



senseable city lab:...

19 Crowdsensing in the Web: Analyzing the Citizen Experience in the Urban Space

Francisco C. Pereira, Andrea Vaccari, Fabien Giardin, Carnaven Chiu, and Carlo Ratti

The mid-sized and large cities of the twenty-first century lead a “double life,” because they exist in both in the physical and the digital worlds. Although these worlds do not physically share the same spatial or temporal dimensions, the anonymous citizen constantly projects the physical world onto the digital world. Websites such as Flickr, Twitter, Facebook, and Wikipedia are repositories of what citizens sense in the city and include reports or announcements of events and descriptions of space.

In this chapter, we analyze the pulse of a city using publicly available user-generated web data. We describe the technical and methodological processes followed and present three case studies that illustrate their potential.

Introduction

The growing popularity of online platforms based on user-generated content is gradually creating a digital world that mirrors the physical world. For almost every city in the world, a parallel digital version exists, spread across different platforms and systems. Such “digital cities” are as rich in diversity and content as their physical counterparts. Furthermore, much of the content is actively generated and updated by residents, tourists, and organizations, in many cases at a quasi-real-time rate. This dynamic creates a historically unprecedented level of intensity to the experience of life in the city. With a common smartphone, a citizen can find and communicate with other individuals, learn of the nearest events at a given time and place, generate content, seek guidance, or report problems, among other capabilities.

For these reasons, the potential impact of an individual’s actions in the digital age is often overrated because of this unprecedented apparent social and technological power. An individual’s behavior, however, rarely deviates significantly from that of the crowd—jointly with peers, his or her behavior *builds* the crowd (Rheingold 2002). The crowd uploads pictures of *popular* events, sends tweets in real time about new happenings, creates and updates pages on Wikipedia about the city, and responds with opinions and hits on the *best* content. These acts of communication generate

different kinds of data that provide unique views on how people experience and view the city.

The crowd therefore becomes a distributed network of sensors that allows us to understand the dynamic patterns of the city and the experiences of its citizens at a quasi-real-time rate, hence the term *crowdsensing*.

In this chapter, we review several types of online data sources related to cities and present techniques available to collect and mine these sources. We then present three case studies that explore what crowdsensing can tell us about how people experience a city. This comprehensive presentation aims to offer the reader insights into the possible applications of crowdsensing in diverse fields and contexts. We focus particularly on collectively built and shared content—content that generates a high level of public interest, or *buzz* (Currid and Williams 2009), during a certain period of time. We describe each case comprehensively to convey the main concepts and results. We invite the interested reader to analyze the suggested literature in more detail.

In the first case study, the *Eyes of the World*, we estimate the attractiveness and popularity of places and events based on the density of user-generated data, in particular the photographs uploaded by Flickr users with tags including information on their location and time in addition to a description. We summarize the results of two experiments in Rome and New York City, employing information visualization to ground and evaluate urban strategies. In the second case study, *My Architect*, we take the idea of exploring the density of user-generated data a step further, using these data to benchmark “iconic architecture.” In this context, the measurement of buzz is correlated to the popularity of an aesthetic, helping to evaluate the success of architects in promoting the image of a city.

In the third and final case study, *Semantics and the City*, we analyze the textual descriptions of points of interest (POIs) and events found on the web, both on individual homepages and on encyclopedic platforms such as Wikipedia. This analysis is carried out with the application of *information extraction* (IE) and *natural language processing* (NLP) techniques that extract the most important concepts found in freely flowing text. From these concepts, we generate an augmented view of the meaning of space, both from a dynamic perspective (based on events), and a static perspective (based on homepages or Wikipedia).

Collecting and Mining Buzz

The Internet is extremely rich in content on the spatiotemporal and semantic dimensions of cities. This wealth of information is generated by the interactions of users with wireless and online services that enable us to develop new methods for collecting and analyzing data related to the social dynamics of the city (O’Neill et al. 2006). The integration of content spread over many websites and services could build new metrics for describing the spatial distribution and temporal evolution of the built environment

(Kostakos et al. 2008). Additionally, the distributed presence of personal devices creates a vast sensor network that could reveal collective behaviors with unprecedented details (Goodchild 2007; Zook et al. 2004; Budhathoki, Bruce, and Nedovic-Budic 2008; Eagle and Pentland 2006). As table 19.1 illustrates, these *digital footprints* present an opportunity in urban and tourism studies to build more efficient ways to collect aggregate information about visitors' activities.

Two of the case studies presented in this chapter focus on the information hidden in every photograph uploaded on the Flickr website. People using the Flickr service to share and organize photographs also have the option to add geographic attributes. When a photograph is anchored to a physical location, Flickr assigns longitude and latitude values together with an accuracy attribute derived from the zoom level of the map in use to position the photographs. Therefore, photographs positioned on a map when the user zooms in to the street level receive a higher accuracy estimate than those positioned when the user zooms out to the map view. The system also adds metadata embedded by the camera into the image using the *Exchangeable Image File Format* (EXIF) information, thus completing the spatiotemporal information.

In the *Eyes of the World* and *My Architect*, we develop an application that retrieves Flickr photographs with specific types of digital information such as georeferences and matching tags. To understand and rank the buzz of user-generated content within the scope of architecture, custom text-filtering software was used to monitor and ensure the quality of the Flickr photos we collected. Data were primarily categorized and filtered textually in English.

These two projects illustrate the wide applicability of user-generated content in the study of urban processes. This dataset allows for the study of mobility and tourism on many different scales, from within the city to between cities or even countries. Connections between different areas of the city become salient (e.g., which tourist hotspots are visited by people from a similar origin), and insights are revealed into how cities are interpreted (e.g., which locations are considered more or less important and what is captured by the eyes of the people who are there).

The *Semantics and the City* project takes a broad approach and considers a wide range of websites. The initial seed is a *point of interest* (POI) or an event, such as a concert or theater performance. Given its location and title, we conduct a search for the best set of pages associated with the POI or event, and then perform a sequence of natural language analyses: *part-of-speech tagging*, *noun phrase chunking*, *named entity recognition*, and *WordNet concept extraction*, using available NLP tools (Toutanova et al. 2003; Ramshaw and Marcus 1995; Finkel, Grenager, and Manning 2005; Fellbaum 1998). A list of words representing each POI or event is thus extracted. The last step consists of applying *information retrieval* techniques to rank those words in relative importance, more specifically *term frequency times inverse document frequency* (TF IDF). For each document, this measure balances the popularity of certain words with respect to all documents for that POI against the frequency of those words in that specific

Table 19.1

Data capture techniques with their main strength and weakness in the context of tourism and urbanism studies.

Data capture	Strength	Weakness	Example of application
Land-use and census data	Applicable to many scales and over long time period.	Infrastructure and service based, static view of urban dynamics.	Estimate the tourism intensity of an area.
Manual surveys	Capture high-level information such as motivations and reasons for staying in specific areas.	Very costly and applies to limited time periods.	Capture the motivation for visiting and length of stay.
Near-field communication	Precise real-time mobility data.	Costly infrastructure deployment.	Describe the social and spatial characteristics of space (Kostakos et al. 2008).
GPS logs	Precise mobility data.	Does not scale well if deployed for the purpose of a survey alone. Limited in time and participants.	Cluster tourist routes (Asakura and Iryob 2007).
Cellphone (device-based)	Timely mobility data, potentially augmented with in situ survey.	Does not scale well if deployed for the purpose of a survey alone. Limited in time and participants.	Context-aware experience sampling to capture the experience in situ (Froehlich et al. 2006).
Cellphone (aggregated network-based)	Use existing infrastructure to provide real-time density and mobility data, covering multiple geographic scales (neighborhood, city, country).	Reveal large-scale phenomena but do not explain the reasons.	Real-time traffic detection (Yim 2003).
User-generated content	Exploit publicly available data with no need for deployment or preexisting infrastructure.	Credibility of information and no systematic coverage.	Reveal flows of photographers (Giardin et al. 2008).

document. For example, given two terms that appear once in a POI word list (e.g., *menu* and *Indian cuisine*), the one that appears in many documents relating to that POI (e.g., *menu*) in restaurant POIs gets a lower TF IDF value than one that appears only in a few documents (e.g., *Indian cuisine*), because the latter is more informative. Frequency inside the document can influence this balance—for example, if *menu* appears much more often than *Indian cuisine*, it will become more important. The final outcome is a ranked list of words for each POI or event. As expected, this list is very dependent on the initial search; therefore, we divide our approach into a number of perspectives that consist essentially of a preselection of search spaces. In the Wikipedia perspective, we only search on Wikipedia.org; in the events perspective, we search on upcoming.org and specific study-area sites, such as calendar.boston.com, while in the Open Web perspective we use the Yahoo!Search API to perform an unbounded search.

The Eyes of the World: Visualizing Buzz as it Comes Online

Visitors to a city have many ways of leaving voluntary or involuntary electronic trails: prior to their visits, tourists generate server log entries when they consult digital maps (Fischer 2007) or travel websites (Wöber 2007); during their visit, they leave traces on wireless networks (Ahas et al. 2007) whenever they use their mobile phones or credit cards (Houée and Barbier 2008); and after their visit they may contribute reviews (Mummidi and Krumm 2008) and photographs (Crandall et al. 2009) online. In this section we present analyses and visualizations of photographs left by visitors on Flickr.

The explicit character of photograph geotagging and manual disclosure to the world generates many dimensions of interest. Positioning a photo on a map is an act of communication that embodies the locations, times, and experiences that an individual considers to be relevant to himself or herself and of interest to others. Our results clearly show that Flickr users have a tendency to point out the highlights of their visit to a city while skipping over the lowlights.

Previous research has also shown that cellular networks in urban areas are efficient tools to study both individuals (González, Hidalgo, and Barabási 2008) and crowds (Ratti et al. 2006) due to their pervasive coverage and the widespread usage of cellular phones. For example, the analysis of mobile data (see Yim 2003) can generate information about traffic conditions in real time. Cellular network signals can also be correlated (albeit with limited success) to the actual presence of vehicles and pedestrians in the city (Sevtsuk and Ratti 2007). In a case study of tourism dynamics in Estonia, Ahas et al. prove that the sampling and analysis of passive mobile positioning data is a promising resource for tourism research and management.

In the case study of Rome (Girardin et al. 2008), we illustrate the potential of user-generated electronic trails based on the sequences of photographs to reveal the presence and movement of visitors in a city. Our analysis allows us to identify the areas

that attract the greatest attention such as the Coliseum and the main train station next to the Piazza della Repubblica, as well as the temporal signature of these places—that is, the level of congestion with respect to the time of day and day of the week.

The study of digital footprints also reveals *desire lines* embodied in the paths of those traveling throughout the city. We first reveal the most active areas through spatial clustering of the data and then aggregate individual paths to generate lines capturing the sequential preferences of visitors. This process produces multiple directed graphs that allow us to compute the number of sites visited by season, the most visited hotspots, and so forth.

For instance, figure 19.1 illustrates the main paths taken by photographers between points of interest in the city. Significantly, the 753 visiting Italian photographers (top) are active across many areas of the city, while the 675 American visitors stay on a narrow desire line between the Vatican, the Forum, and the Coliseum. One cannot conclude that American visitors only explore these areas of Rome, but rather that they concentrate on these points of interest when relaying their experience of the city.

In a follow-up case study in New York City (Girardin et al. 2009), we exploit the spatiotemporal characteristics of buzz to inform local authorities about the success of the “New York City Waterfalls,” a public art project consisting of four human-made waterfalls rising from the New York Harbor in the East River from June 26 to October 13, 2008. The waterfalls were intended to attract residents and visitors to the city’s waterfront with the goal of stimulating social and economic activity.

Given the large investment in the temporary installation, its organizers wished to assess the economic impact of the event. Traditional methods such as people counts and surveys, which produce accurate estimates only for confined areas, failed to capture the dynamics of the open spaces around the waterfront. Our analysis of geo-referenced photographs, on the contrary, helped to quantify the influence of the public art exhibition on the attractiveness and popularity of various vantage points in proximity to the event.

We conceptualized the level of attractiveness and popularity as properties of a well-defined place that could vary in intensity. We then used indicators inspired by economic analysis and network theory to measure the attractiveness of the main points of interest around the waterfalls based on their relative strength, and the evolution of the popularity of the waterfront based on its centrality in the network of points of interest.

Table 19.2 shows the variations of the *comparative relative strength* indicator based on the presence of photographers during the summers (June to October) of 2006, 2007, and 2008. It reveals a positive growth in the waterfront’s attractiveness of 8.2 percent in the summer of 2007 and 20.7 percent in the summer of 2008 with respect to other areas of interest in New York City, such as Times Square and Central Park. It should be noted that the maximum growth in attractiveness, 29.9 percent, observed during

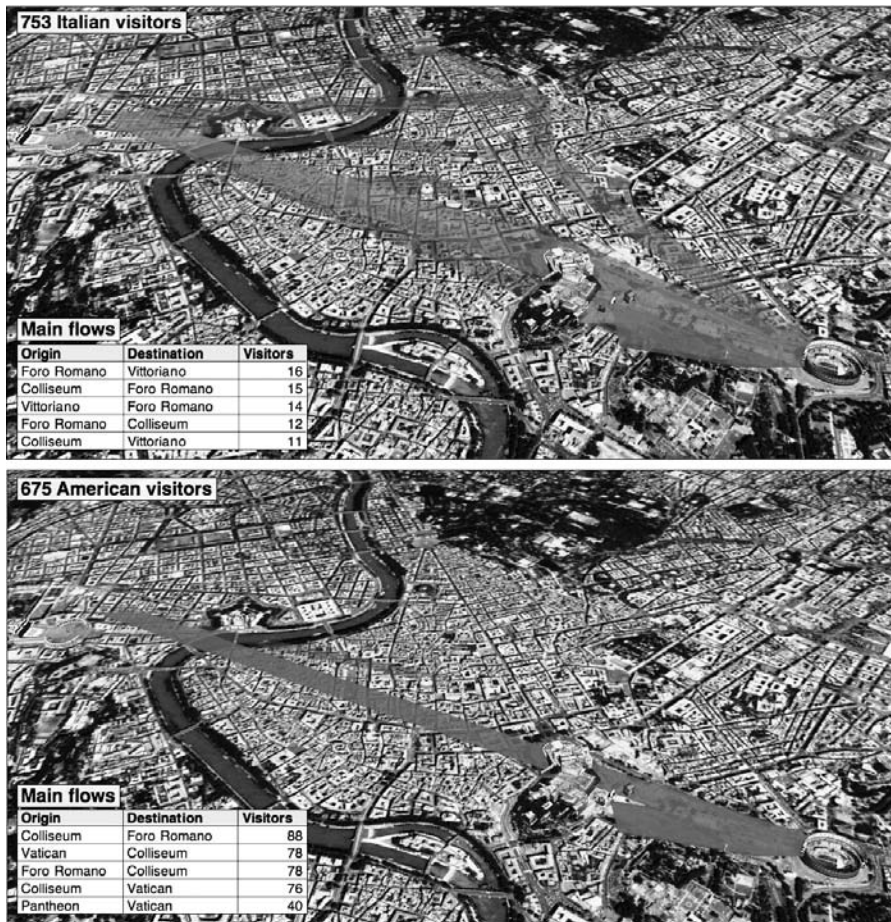


Figure 19.1

Geovisualization of main paths taken by photographers between points of interest in Rome. Significantly, (*top*) the 753 visiting Italian photographers are active across many areas of the city, whereas (*bottom*) the 675 American visitors stay on a narrow path between the Vatican, Forum, and Coliseum. (Different scales apply to each geovisualization.)

Table 19.2
Variation of CRS indicators from 2006 to 2008

	Photographers 2006	Photographers 2007	Photographers 2008	CRS 2006	CRS 2007	CRS 2008	2006 to 2007	2007 to 2008
Central Park	1874	2619	1537	0.111	0.101	0.100	-0.091	-0.008
Chelsea	1146	1790	1125	0.068	0.069	0.073	0.015	0.062
East Village	652	971	606	0.038	0.037	0.039	-0.031	0.054
WTC site	564	775	374	0.032	0.029	0.024	-0.075	-0.184
Time Square	1227	1883	1026	0.073	0.072	0.067	-0.002	-0.079
Vantage Points	538	896	640	0.032	0.034	0.041	0.082	0.207

the summer of 2008, was recorded in DUMBO, Brooklyn: this was likely supported by the increased presence of photographers, elicited by the waterfalls, at the proximate vantage points of Pier 1 and Main Street Park.

We assess the popularity of an area of interest by studying its ties to other areas in the city. The stronger the ties, and as it becomes part of a popular route, the more frequently an area is accessed. This is measured by applying network analysis techniques to study the connectivity of a network in which the nodes represent areas of interest and the edges represent flows of people between them. Figure 19.2 illustrates the flows estimated by analyzing Flickr photos in conjunction with the reported location where and when the photos were taken.

These aggregate, spatiotemporal records lead to a novel perspective on different aspects of mobility and travel. Although the results are still fairly coarse, we clearly show the potential for geographically referenced digital footprints to reveal patterns of mobility and preference among different visitor groups.

We also analyze the flows of visitors between several points of interest in Lower Manhattan to track the evolution of the centrality of the waterfront area in comparison to other points of interest. Mapping this new type of digital footprint analysis shows the capacity of an event to drive people to less explored parts of a city over time, information that can be highly valuable for urban design and tourism studies.

Ranking Buzz: A Case Study in Architecture

The intimate relationship between photography and architecture has been well documented since the early days of photography (Robinson and Herschman 1988).

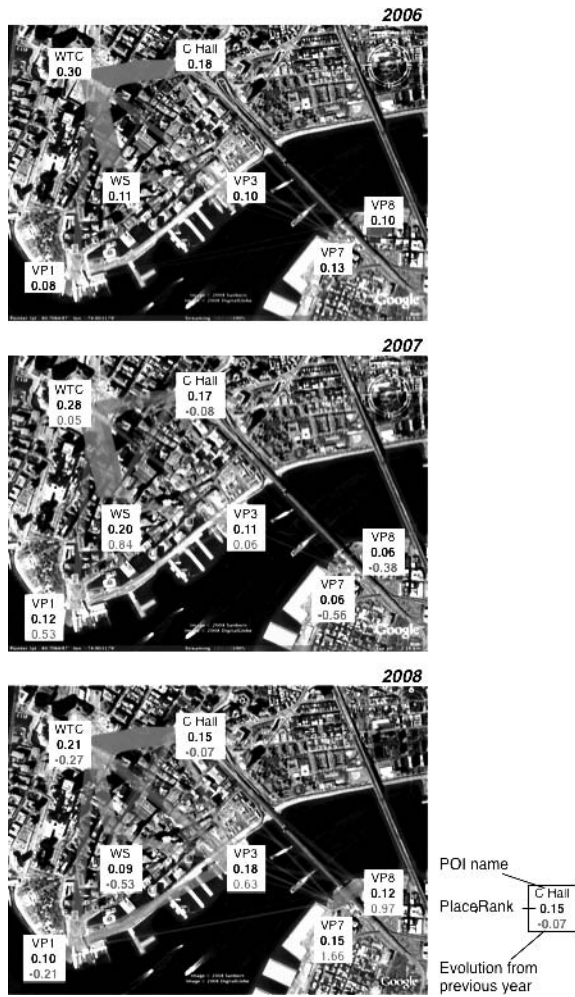


Figure 19.2

Evolution of the flows of photographers in proximity to the exhibit based on the analysis of photos generated in June–October 2006, June–October 2007, and June–August 15 in 2008. In 2008, when Waterfalls opened, VP3, VP6, and VP7 massively increase their PlaceRank.

Photographs reflect on the built world in which we live and generate awareness of its visual qualities, allowing us to make connections and observations that would not have been obvious otherwise. A collection of photographs of a particular place tells a story of the way it is designed to be experienced. Flickr provides vast amounts of this type of data, particularly for places that are of substantial interest, including the most important architectural sites.

My Architect offers a global view of the presence and work of a selection of the world's greatest architects. Designed under the same principles and built from a very similar set of components as the *Eyes of the World*, it serves as an extension of this project but focuses solely on architecture. The public creates narratives based on their activities from which we are able to extract a general opinion of these spaces.

My Architect visualizes the formation of iconic architecture and “starchitecture” culture. Through this visualization, we highlight the clusters of the physical presence of each architect and showcase their most famous works through stunning photographs captured by anonymous photographers around the world. This project also features a ranking of the architects based on the size of their online photograph collection. This ranking deviates from the media's measure in that it is not simply representing which buildings are photographed the most; it is instead an examination of pop architecture culture formalized via peer-to-peer information sharing online. This bottom-up approach to conveying information related to architecture contrasts with the traditional top-down approach in which communication media and other third parties measure the success of architectural practice and professions.

For this case study, we develop several filtering methods to extract desired photographs for evaluation that are from a large collection in the Flickr database. We begin with gathering a list of approximately three hundred living architects from around the world, using information from various online platforms including Wikipedia and the most popular blogs and sites related to architecture. We then establish the preliminary scope for this project by retrieving the total number of available photographs on Flickr for each of the architects included in this list. From this initial search, the top hundred architects were selected for further analysis based on the frequency of photographs of their work. The following paragraphs introduce the procedures we used to further control and refine the quality of the search results, and furthermore analyze the user-generated tags that are associated with the Flickr photographs.

Using simply the names of the architects creates unreliable search queries, because they are not unique to those specific individuals (e.g., Smith). Therefore, to ensure that the photographs returned from the search are relevant to architecture, we compile a list of the most popular words with which Flickr users tag photographs related to architecture, including, among others, *exterior*, *interior*, *architecture*, *building*, *windows*, *doors*, *facade*, *design*, *architect*, *designer*, *skyscraper*, *museum*, *outdoor*, *tower*, *house*, *contemporary*, and *modern*. All available tags retrieved for each photograph are matched

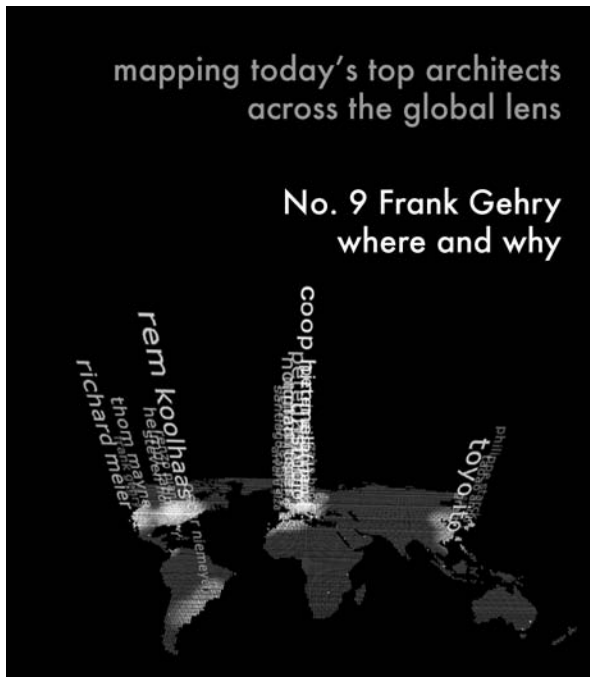


Figure 19.3
Scanning for top architects.

against these words, and the selection is narrowed to those photographs overlapping most strongly with the list. Using a filter script to check for matches between the lists, we extract photographs that contain a minimum of one match. The top fifty architects in the world are selected based on the size of the photograph collection filtered through this process.

We design two distinct visualizations to exhibit the data collected. The first one (figure 19.3) introduces the overall concept of the project to the audience. A “crowd detector” scans across the global map, revealing the names of architects as it scans over the geographic area where each architect contributes or is photographed the most in relation to the rest of his or her work. The size and color of each architect’s name correspond to the frequency with which his or her projects are photographed. The second visualization (figure 19.4) focuses on displaying the data for each individual architect. A line scans across the screen and generates a particular heat map for the selected architect while presenting the Flickr photographs to viewers.

The analysis and mapping of these data reveal the global points of architectural interests and how they contribute to the level of tourist activity. They also expose how



Figure 19.4

A sequence of four snapshots. Notice the sweep line crossing from west to east, leaving a trace of photos.

architectural landmarks and the role of architects contribute to shaping urban activity and aesthetics. The image of a place or city reveals the dynamic aspect of how people utilize the photographic medium to explore architecture and the architectural experience.

Understanding Buzz: Natural Language Text Analysis of User-Generated Content

Our main goal in *Semantics and the City* is to understand the meaning of space in static and dynamic terms—that is, what exists and what happens in a particular space. Coupled with other types of data from the city such as mobile-phone usage and GPS traces, we can extract relationships between space and mobility in the city. For example, understanding that a specific place has recurrent congestion due to its point of interest (POI) profile and events—for example, a sports stadium and important games or rock concerts—would facilitate the development of a traffic prediction model. Other examples concern the sociological relationship between space and mobility: people who go to family concerts also go to interactive museums; people from specific neighborhoods prefer sports events or musical concerts, and so on.

As described in the above section, given a POI or event, we apply a sequence of steps that use natural language processing, information retrieval, and information extraction techniques to obtain a list of words that correspond to the “meaning” of that place. The first decisive step is the choice of the document search space. We call each choice a perspective, because it reflects the type of knowledge embedded:

Wikipedia This perspective relies on encyclopedic information. Although the accuracy of its content may vary, Wikipedia contains over three million pages created and maintained on a crowdsourcing basis. Given a POI, we consider two different analyses: documents corresponding to the type(s) of the POI (red wiki perspective) and docu-

ments with the same or a similar title to the name of the POI (yellow wiki perspective).

Events This perspective is based on analyzing events occurring at a given place. For a specific study area, we manually search for the website(s) that cover the largest number of public events. The minimum information on each event normally contains details on the time and location in addition to a small description. For a given location or venue, we analyze the descriptions of the events, and furthermore, for each of the most relevant words, search on Wikipedia for related documents. For example, the description of an event containing the words *Bach*, *Partita*, *Handel*, and *Vivaldi* could lead to terms in Wikipedia such as *baroque* or *concerto grosso*, which were not referred to in the original text.

Open Web This is the unconstrained perspective. It simply uses the Yahoo!Search API to find pages on a POI. This perspective possesses the highest potential because it can obtain the “official” homepages; it is also the most challenging because there is no a priori control to exclude noise or secondary information.

We initially considered the Flickr perspective. For the POIs and events in our database, however, the top words obtained from Flickr typically focus on geographic information, which becomes redundant (e.g., photos of the Brooklyn Bridge are mostly tagged as “Brooklyn,” “Bridge,” “Manhattan,” “New York,” etc.). There are also many nongeographic tags, but these are extremely varied according to the individual and therefore become submerged under the geographic tags because they appear with a lower frequency. Only in POIs with a strong aesthetic component such as architectural monuments, visual arts museums, and so on, will such data be more relevant, because descriptive words become more common (e.g., *baroque*, *cubist*, etc.), as we can see in *My Architect*.

In any of the perspectives, the choice of the “ideal” set of words concerning an event or place ordered by relevance is ultimately a subjective task. What we obtain is either the result of a consensus (Wikipedia) or the union of individual contributions (events, Open Web), and the process we apply is ultimately based on statistical analysis. We therefore claim to obtain a valid set of words that accurately represents the way places and events are referred to by the crowd in all of these perspectives, as opposed to an ideal single set of words. As an example, while the words *fat*, *unhealthy*, and *capitalism* are associated with many fast food chains, they are rarely used on Wikipedia and the homepages of the chains. This justifies the exploration of other perspectives including crowd opinion and sentiment analysis.

From our collection of a large set of POIs from Boston, New York, and San Francisco, we begin the computationally time-consuming extraction of words that results in what we term “enrichment.” The average time for the analysis of a POI from the Open Web perspective is approximately 108 seconds; analysis from the red and yellow wiki

Table 19.3
Summary of statistics

	New York	Boston	San Francisco	Overall
Yahoo!	183144	64133	94466	251839
YellowPages	7694	12878	—	21333
Boston Calendar	13999	2867	9497	26364
OpenWeb	757	2020	—	2896
Red Wiki	69011	20309	—	90210
Yellow Wiki	4400	1928	—	6330
Events	—	7591	—	—
Enriched Events	—	3827	—	—

requires 57 and 31 seconds, respectively; and analysis from the events perspective takes 30 seconds on average. The Open Web is the most time consuming, because it searches the entire web. In the events perspective, the system already contains the initial text descriptions for the events, and its subsequent steps are similar to those involved in the yellow wiki perspective, hence the similarity in time requirements. In table 19.3, we present the overall statistics.

We obtain a total of 77,558 different words, of which 9,746 (12.6 percent) are also identified in WordNet, a lexical database for the English language. An analysis of these concepts was performed on the average information content (IC) (Resnik 1995), which reflects the balance of *s* of a concept on a scale of 0 to 17. The average we obtain is 16.313395 (st. dev. = 1.7263386), indicating that the concepts the words illustrate are very specific. This level of specificity can, however, pose risks. If the concepts prove to be generic as opposed to very specific, the probability of being accurate with respect to the place is much higher. Table 19.4 includes excerpts from results illustrating both the strengths and weaknesses of the approach.

Taking into account the top five words in each case, the experiments show that the median of correct words¹ is four in both the Wikipedia and events perspectives and two in the Open Web perspective. More specifically, the red wiki perspective most reliably ensures correctness, because it sacrifices the specificity of each POI for the analysis of its category and then typically yields correct results. The yellow wiki and Open Web perspectives prove to be the best for extracting exact information on a specific place. The former is preferable when the POI exists on Wikipedia, while the latter is the only alternative option. Finally, the events perspective is unique in that it brings information about what happens in the place rather than how the place is defined on the web.

This work is currently being applied in two different analyses: (1) the correlation of cellphone usage with event types and semantics, and (2) the identification of

Table 19.4

Excerpt of results

Name	Categories	Terms
Red Wiki		
Gas stations	pumps, gasoline, fuel dispenser, filling station, gasoline stand	Bowdoin Square Exxon
Grocery stores	groceries, retailing, food, vegetables, products	Harvard Market
Interior design	office space, architects, private residence, code, decoration	Kim Depole Design Incorporated
Yellow Wiki		
Clothing, women's clothing	Victoria, wear, limited brands, top model, fashion models	Victoria's Secret
Law enforcement	Massachusetts, law enforcement agency, correction, investigation, responsibility	Boston Police Department
Entertainment venues	Boston Celtics, arena, Boston Blazers, naming rights, National Lacrosse League	TD Garden
Events		
Nature	pond, falls, streams, currents, winter	Nature Trail and Cranberry Bog
Theater	orpheum, journey, spirit, tale, surprises	The Haunted House
Farmers' markets	cultures, consumption, carbohydrate, food safety, gastronomy	Salem Farmers' Market
Open Web		
Waste and environmental consulting	Industrial services, asbestos management, mildew removal, asbestos removal, residential services	Envirotech Incorporated
Telecommunications	Boston Telecommunications, Gary, communication services, Boston Business Directory, telephone communications	Grasshopper
Banks	Houston, reading room, Allston, Senior Commercial Loan and Business Development, federally chartered	Cambridge Savings Bank

attendees' origins according to event types and semantics. In the first case, we focus again on the waterfalls area during the same period described above and search for causality relations between the semantics of events and cellphone activity. The hypothesis is that, for example, an event that involves the concept "films" or "cinema" would have less activity than one that involves "family" or "sports." We partition areas using cellphone coverage maps and cluster event venues according to those areas. We then choose events isolated in time with respect to their area to avoid ambiguities with concurrent events, extract their words and categories, and compare cellphone activity with similar periods for the same area. When "activity," the number of calls received or sent, is higher or lower than typical cellphone activity by more than one or two standard deviations, we classify the event as having high/very high or low/very low activity, respectively. We next run associative machine learning algorithms (the a priori algorithm) to understand nominal correlations between the two datasets. In table 19.5, we show the top twelve rules obtained. In general, the results confirm our hypothesis, but further experimentation is needed to reach more confident conclusions.

Table 19.5

Top 20 association rules considering all attributes (S/C=support/confidence)

Precondition	Activity	S/C
category=Music time=afternoon	normal	35/35
category=Performing/Visual Arts day=weekday	normal	25/25
category=Other day=weekend	normal	25/25
category=Commercial day=weekday	high	15/15
category=Commercial time=morning	high	15/15
category=Festivals	high	10/10
category=Education time=morning	normal	10/10
category=Performing/Visual Arts day=weekend	high	10/10
category=Performing/Visual Arts time=night	normal	10/10
category=Education day=weekend time=afternoon	normal	10/10
word=request	normal	6/6
category=Sports	normal	5/5
time=dawn	very high	5/5
category=Education time=night	high	5/5
category=Media day=weekday	very low	5/5
category=Media time=morning	normal	5/5
category=Commercial day=weekend	normal	5/5
category=Commercial time=afternoon	normal	5/5
category=Social time=night	normal	5/5
day=weekend time=morning	normal	5/5

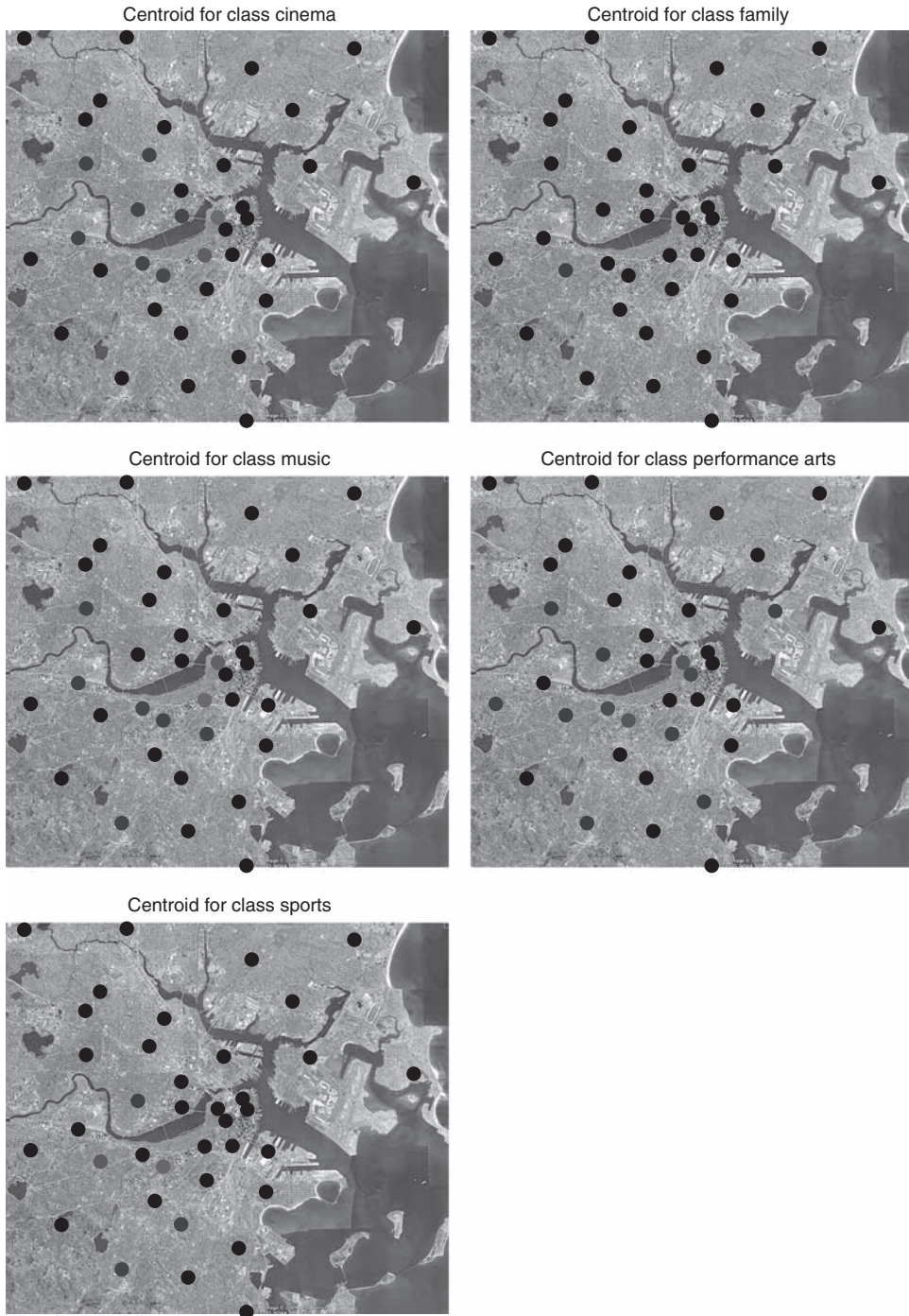


Figure 19.5
 Spatial visualization of clusters centroids. The circles correspond to zip-code areas with values greater than zero. The shade from light (low) to dark (high) is proportional to the values.

To explore the nature of events, we analyze the correlations between large events and attendees' origins in Boston. Using over one million anonymized cellphone user traces triangulated to obtain general location (zip code), we obtain the origins for a sample of attendees at each of the selected events. For example, for "Shakespeare on the Common," a large outdoor theater event on Boston Common, we identify that 11 percent of attendees came from the immediate neighborhood, followed by 16 percent in the bordering areas, and so on. From the categories of events not concurrent in space and time, we built a neural network classifier that predicts event types given origin distributions with accuracy ranging from 60 to 90 percent (Calabrese et al. 2010). In figure 19.5 we summarize the overall clustering results of origin distributions for each event type.

Conclusions

As with any dataset, user-generated data yields incomplete conclusions. The subjective nature of this type of data facilitates an understanding of the experience of a city based on the preferences and values of residents and visitors. Instead of being the result of intentional contributions, user-generated data is a form of *implicit engagement*; digital content is a contribution to the city itself, and social networks are now part of the experience of the city. The collection and analysis of these data enable the development of new urban indicators useful to evaluate urban strategies.

In all of the works mentioned, we face the challenge of validating this pervasive, user-generated content. Indeed, our data processing techniques try to account for the inconsistent quality of user-generated data, which can substantially impede our ability to generate accurate information. For instance, the timestamps extracted from the camera-generated EXIF metadata do not necessarily match the real time when a photo was taken because the user may not have set the clock on their camera to the local time, or even to the time from their area of origin. User-generated data points can also be apparently idiosyncratic and, for instance, indicate not the point where the photo was taken but the location of the photographed object.

Strong concerns also arise in the validation of semantics obtained from online texts. The ambiguous and subjective nature of information hinders the identification of a ground truth; the decision on whether a word is related to a context on a qualitative scale, however, should be attainable. In our case, care was taken to rank words correctly by applying TF IDF, then using extended stop word lists² to filter out format words such as *http*, and finally analyzing self-consistency. For example, a clustering analysis shows that each category of POI attracts a different set of words. A field survey analyzing the results beyond the laboratory is a necessary next step to this project.

Finally, another important limitation of these approaches is the bias toward "loud" crowds and events—that is, those that due to the nature and character of their crowd lead to more Flickr photos or textual contributions on the web, because their attendees

tend to be more technologically savvy. Any application of these analyses should heed this bias.

These shortcomings currently limit the usability of digital footprints for city management and decision making. Visualizations produced from these datasets, however, are already offering new tools to communicate different features of activity in urban areas. Along with interactive software, visualization is a useful tool for many actors in the city such as researchers, practitioners, service providers, and local authorities to discuss how to interpret data and put information in context. For example, an area with strong calling activity and weak photographic activity could be primarily commercial; one with weak calling activity and weak photographic activity could be residential; one with strong calling activity and strong photographic activity could be friendly toward tourist- and leisure-related activities. Similarly, an outdoor area with strong attractiveness indicators even during adverse weather conditions suggests that it is critical for visitors; an indoor area with weak attractiveness indicators during adverse weather conditions indicates that it may not be easily accessible.

Virtual cities are growing. Everyday more people become new residents who increasingly use more websites and devices. While the virtual city may never merge into the physical one, it will gradually reflect reality more accurately and become an even more powerful resource to understand the cities in which we live and the cities we would like to visit.

Acknowledgments

We are extremely grateful to our colleague Caitlin Zacharias for her valuable comments and her expert editorial guidance. We also acknowledge generous support from the members of the SENSEable City Lab Consortium, Volkswagen of America, the AT&T Foundation, Airsage, and the MIT-Portugal Program.

Note

1. For determining *correctness* for each sample we examined whether the words were specifically related to a POI or event or if they were unrelated—for example, whether “laundry service” refers to a place for laundry as opposed to Shakira’s pop music album—and if the information the words refer to was too general.
2. Stop word lists are a specific type of list that contains words that have little semantic meaning, such as articles, “as” “and” “or,” HTTP codes, tags, and so on.

References

Ahas, R., A. Aasa, S. Silm, and M. Tiru. 2007. Mobile positioning data in tourism studies and monitoring: Case study in Tartu, Estonia. In Marianna Sigala, L. Mich and J. Murphy, eds.,

Information and Communication Technologies in Tourism 2007, 119–128. Ljubljana, Slovenia: Springer Vienna.

Asakura, Y., and T. Iryob. 2007. Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. *Transportation Research, Part A: Policy and Practice* 41 (7): 684–690.

Budhathoki, N., B. Bruce, and Z. Nedovic-Budic. 2008. Reconceptualizing the role of the user of spatial data infrastructure. *GeoJournal* 72 (3): 149–160.

Calabrese, F., C. Pereira, L. Liu, G. Lorenzo, and C. Ratti. 2010. The geography of taste: Analyzing cell-phone mobility and social events. In Patrik Floréen, Antonio Krüger, and Mirjana Spasojevic, eds. *Proceedings of the 8th International Conference on Pervasive Computing*. 22–37. Heidelberg: Springer.

Crandall, D. J., L. Backstrom, D. Huttenlocher, and J. Kleinberg. 2009. Mapping the world's photos. In *Proceedings of the 18th international conference on the World Wide Web*, 761–770. New York: ACM.

Currid, E., and S. Williams. 2009. The geography of buzz: Art, culture and the social milieu in Los Angeles and New York. *Journal of Economic Geography*.

Eagle, N., and A. S. Pentland. 2006. Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing* 10 (4): 255–268.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Finkel, J. R., T. Grenager, and C. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*, 363–370. Stroudsburg, PA: ACL.

Fisher, D. 2007. Hotmap: Looking at geographic attention. *IEEE Transactions on Visualization and Computer Graphics* 13 (6): 1184–1191.

Froehlich, J., M. Y. Chen, I. E. Smith, and F. Potter. 2006. Voting with your feet: An investigative study of the relationship between place visit behavior and preference. In P. Dourish and A. Friday, eds., *Ubicomp: Lecture Notes in Computer Science*, 333–350. Berlin: Springer.

Giardin, F., F. Calabrese, F. Dal Fiore, C. Ratti, and J. Blat. 2008. Digital footprinting: Uncovering tourists with user-generated content. *IEEE Pervasive Computing* 7 (4): 36–43.

Giardin, F., A. Vaccari, A. Gerber, A. Biderman, and C. Ratti. 2009. Quantifying urban attractiveness from the distribution and density of digital footprints. *International Journal of Spatial Data Infrastructure Research* 4:175–200.

González, M. C., C. A. Hidalgo, and A.-L. Barabási. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196): 779–782.

Goodchild, M. F. 2007. Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research* 2:24–32.

Houée, M., and C. Barbier. 2008. Estimating foreign visitors flows from motorways toll management system. Paper presented at the 8th International Conference on Survey Methods in Transport: Harmonisation and Data Comparability, Annecy, France, May 25–31.

Kostakos, V., T. Nicolai, E. Yoneki, E. O'Neill, H. Kenn, and J. Crowcroft. 2008. Understanding and measuring the urban pervasive infrastructure. *Personal and Ubiquitous Computing* 13 (5): 355–364.

Mummidi, L. N., and J. Krumm. 2008. Discovering points of interest from users' map annotations. *GeoJournal* 72(3): 215–227.

O'Neill, E., V. Kostakos, T. Kindberg., A. F. gen. Schieck, A. Penn, D. S. Fraser, and T. Jones. 2006. Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In P. Dourish and A. Friday, eds. *Ubicomp: Lecture Notes in Computer Science*, 315–332. Berlin: Springer..

Ramshaw, L., and M. Marcus. 1995. Text chunking using transformation-based learning. Paper presented at *The 3rd Workshop on Very Large Corpora: WVLC-1995*.

Ratti, C., R. M. Pulselli, S. Williams, and D. Frenchman. 2006. Mobile landscapes: Using location data from cell-phones for urban analysis. *Environment and Planning, B: Planning & Design* 33 (5): 727–748.

Resnik, P. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'95)*, 448–453. San Francisco, CA: Morgan Kaufmann.

Rheingold, H. 2002. *Smart Mobs: The Next Social Revolution*. New York: Basic Books.

Robinson, C., and J. Herschman. 1988. *Architecture Transformed: A History of the Photography of Buildings from 1839 to the Present*. Cambridge, MA: MIT Press.

Sevtsuk, A., and C. Ratti. 2007. Mobile surveys. In J. Van Schaick and S. C. Van Der Spek, eds. *Urbanism of Track*, 103–119. Delft: Delft University Press.

Toutanova, K., D. Klein, and C. Manning C. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology—Volume 1*, 173–180. Stroudsburg, PA: ACL.

Wöber, K. 2007. Similarities in information search of city break travelers - a web usage mining exercise. In Marianna Sigala, L. Mich, and J. Murphy, eds., *Information and Communication Technologies in Tourism 2007*, 119–128, Ljubljana, Slovenia: Springer Vienna.

Yim, Y. 2003. *The State of Cellular Probes*. Technical report. Institute of Transportation Studies, UC Berkeley.

Zook, M., M. Dodge, Y. Aoyama, and A. Townsend. 2004. New digital geographies: Information, communication, and place. In S. Brunn, S. Cutter, and J. W. Harrington, eds. *Geography and Technology* 155–176. Dordrecht, Netherlands: Kluwer Academic Publishers



G

