11 Geoinformatics and Social Media

New Big Data Challenge

Arie Croitoru, Andrew Crooks, Jacek Radzikowski, Anthony Stefanidis, Ranga R. Vatsavai, and Nicole Wayant

CONTENTS

11.1 Introduction: Social Media and Ambient Geographic Information ..........205
11.2 Characteristics of Big Geosocial Data......................................................207
11.3 Geosocial Complexity ............................................................................213
11.4 Modeling and Analyzing Geosocial Multimedia: Heterogeneity and Integration...............................................................................................217
11.5 Outlook: Grand Challenges and Opportunities for Big Geosocial Data.....223
References........................................................................................................226

11.1 INTRODUCTION: SOCIAL MEDIA AND AMBIENT GEOGRAPHIC INFORMATION

Fostered by Web 2.0, ubiquitous computing, and corresponding technological advancements, social media have become massively popular during the last decade. The term social media refers to a wide spectrum of digital interaction and information exchange platforms. Broadly, this includes blogs and microblogs (e.g., Blogger, WordPress, Twitter, Tumblr, and Weibo), social networking services (e.g., Facebook, Google+, and LinkedIn), and multimedia content sharing services (e.g., Flickr and YouTube). Regardless of the particularities of each one, these social media services share the common goal of enabling the general public to contribute, disseminate, and exchange information (Kaplan and Haenlein, 2010). And this is exactly what the general public does, making social media content a sizeable and rapidly increasing chunk of the digital universe. Facebook announced in 2012 that its system deals with petabyte scale data* as it processes 2.5 billion content elements and over 500 TB of data daily.†

* http://www.infoq.com/presentations/Data-Facebook
† http://tcrn.ch/NhjAVz
This social media content is often geo-tagged, either in the form of precise coordinates of the location from where these feeds were contributed or as toponyms of these locations. Based on studies by our group using the GeoSocial Gauge system that we developed to harvest and analyze social media content (Croitoru et al., 2012), we have observed that on average, the percentage of precisely geolocated (at the level of exact coordinates) tweets ranges typically between 0.5% and 3%. Depending on the area of study and underlying conditions, this rate may occasionally go higher. For example, a dataset collected from Japan following the Fukushima disaster reflected a data corpus where 16% of the tweets were precisely geolocated (Stefanidis et al., 2013). This spike is attributed to the fact that the dataset from Japan reflected a technologically advanced community that was on the move (following the tsunami and subsequent nuclear accident), in which case users were tweeting using primarily their cell phones. Both of these situations, namely, the proliferation of technology in a society and an increased use of mobile (and other location aware) devices to post tweets, are conditions that tend to produce higher rates of geolocated content. In addition to precisely geolocated tweets, we have observed that approximately 40%–70% of tweets come with a descriptive toponym related to the location of the user. Regarding imagery and video contributed as part of social media, a recent study has indicated that approximately 4.5% of Flickr and 3% of YouTube content is geolocated (Friedland and Sommer, 2010).

The geographic content of social media feeds represents a new type of geographic information. It does not fall under the established geospatial community definitions of crowdsourcing (Fritz et al., 2009) or volunteered geographic information (VGI) (Goodchild, 2007) as it is not the product of a process through which citizens explicitly and purposefully contribute geographic information to update or expand geographic databases. Instead, the type of geographic information that can be harvested from social media feeds can be referred to as ambient geographic information (AGI) (Stefanidis et al., 2013); it is embedded in the content of these feeds, often across the content of numerous entries rather than within a single one, and has to be somehow extracted. Nevertheless, it is of great importance as it communicates instantaneously information about emerging issues. At the same time, it provides an unparalleled view of the complex social networking and cultural dynamics within a society and captures the temporal evolution of the human landscape.

Accordingly, social media feeds are becoming increasingly geosocial in the sense that they often have a substantial geographic content. At the same time, we can observe the underlying social structure of the user community by studying the interactions among users. For example, we can identify the trail of a tweet as it is retweeted within the user community, or we can construct a social network describing the follow connections among numerous users. This allows us for the first time to explore the physical presence of people together with their online activities, enabling us to link the cyber and physical spaces on a massive scale. This information contributes additional content to social media (i.e., space) and provides additional context to analyze these data (i.e., topics and sentiment accord). For example, we can identify geographical hot spots of cyber communities that participate in a specific discussion and their interactions with other communities. Accordingly, geosocial analysis is inherently complex, as it comprises the study of content, connections, and locations.
and their variations over time. As such, it represents an emerging alternate form of geographic information, which, through its volume and richness, opens new avenues and research challenges for the understanding of dynamic events and situations.

The objective of this chapter is to present some particular challenges associated with big geosocial data, in order to provide an understanding of the corresponding analysis and processing needs. Accordingly, the chapter is organized as follows. In Section 11.2, we discuss some particular characteristics of social media as they compare to traditional big spatial data. In Section 11.3, we focus on the complexity aspect of social media content, using representative examples. In Section 11.4, we address the integration of diverse content to enable the integrative geosocial analysis of multiple social media feeds. Finally, in Section 11.5, we conclude with a view of the outlook.

11.2 CHARACTERISTICS OF BIG GEOSOCIAL DATA

A recent definition of big data (TechAmerica, 2012) is moving beyond sheer data volume to identify it through two additional properties, namely, velocity and variety as shown Figure 11.1. In this context, velocity refers not only to the rate at which the data is produced but also to the currency of its content and the corresponding need for timely analysis. The need to process data and extract information from them at streaming rates is imposing substantially higher computational demands than the periodic (e.g., daily or weekly) processing of comparable information (e.g., as may be the case when addressing remote sensing data). Variety on the other hand refers to the diversity of the data sources and types that are processed and on the degree to which the information to be extracted is distributed among such diverse sources. It is not uncommon for an event to be communicated by the general public in fragments, as individuals may only have a partial view of the event they are reporting. Accordingly, the event may be communicated across numerous social media channels by multiple users using various modalities (e.g., text in Twitter, images in

![Diagram showing the three dimensions of big data: volume, velocity, and variety.](image-url)
Flickr and Instagram, and videos in YouTube). One of the earliest manifestations of this was during the Mumbai terrorist attacks in 2008, where Flickr imagery, Twitter streams, Google maps mashups, and Wikipedia articles were set up immediately, to provide real-time coverage of the unfolding events (Arthur, 2008). While each piece of information separately adds to our understanding of the event, the aggregate view of all these pieces offers a far better understanding of the full complexity of the actual event. This introduces the need to develop a better capacity to process geosocial multimedia in order to extract knowledge from diverse social media feeds. This is analogous to information aggregation in a geosensor network (Stefanidis and Nittel, 2004), where each sensor contributes a piece of information, but it is through aggregation across that the observed event is revealed in all its complexity. When it comes to social media, people act as sensors too, reporting their observations in the form of multimedia feeds, and the challenge is to compose these fragmented contributions into a bigger picture, overcoming the limitations of individual perception. Handling this variety imposes constraints on both IT architecture and algorithmic needs and will be discussed in more detail in Section 11.4. At the same time, it is important to note that the ability to monitor human observations at a massive scale and to cross-reference such data across a variety of sources and modalities (e.g., text, imagery, video, and audio) presents a unique opportunity to validate information regarding events as they unfold in space and time. We can therefore postulate that Linus’s law (Raymond, 1999), as has been discussed in the context of VGI (Haklay, 2010; Sui and Goodchild, 2011) also has a central role in AGI and geosocial analysis.

Since their inception, geospatial datasets have always been large volume datasets, at the edge of the computational capabilities of each era. This was true at the time of early seminal computer mapping software environments in the late 1960s and early 1970s, such as SYMAP and its use in Waldo Tobler’s movie of the urban expansion of Detroit (Tobler, 1970), and it is certainly true today. As a matter of fact, one could argue that geospatial data collection advances have outpaced computational improvements over the past 40+ years, leading, for example, to Digital Globe currently having an archive of imagery exceeding 4 billion km² in coverage and generating 1–2 PB of data annually to add to its estimated up to 30 PB of archived data. At the same time, the EROS Center is estimated to have over 4 PB of satellite imagery available, the NASA Earth Observing System Data and Information System (EOSDIS) is estimated to house over 7.5 PB of archived imagery (Fares et al., 2012), and NOAA’s climatic data center houses a digital archive of over 6 PB. NASA is currently generating approximately 5 TB of data daily (Vatsavai and Bhaduri, 2013). Furthermore, the proliferation of Google Earth has led to Google maintaining an archive of over 20 PB of imagery, from satellite to aerial and ground-level street view imagery.*

Technological advances moved the geospatial community further into big data territory, by broadening the variety of geospatial datasets that are collected and analyzed and by improving the rates at which such data are generated. A Google Street View car driving down a street is equipped with multiple cameras and LIDAR sensors, capturing millions of image pixels per second, while it also simultaneously scans thousands of point locations to generate thick point clouds (Golovinskiy et al., 2009).

At the same time, video surveillance is generating massive datasets using ground- or airborne-based platforms. DARPA recently unveiled ARGUS-IS,* a 1.8-gigapixel video surveillance platform that can monitor a 25 km² wide area at a pixel resolution of 15 cm and temporal resolution of 12 fps from an altitude of 6 km. At the same time, ground-based video surveillance has been growing extremely rapidly. As a telling reference, in the United Kingdom alone, it is estimated that between 2 and 4 million CCTVs are deployed, with over 500,000 of them operating in London (Norris et al., 2004; Gerrard and Thompson, 2011). By adding velocity and variety to big volumes, these advances have further solidified the big data nature of geospatial datasets. While these challenges are indeed substantial, they reflected an evolution rather than a breakpoint for the geoinformatics community. The objective was only partially altered: applying established analysis techniques onto larger volumes of data.

The emergence of social media however is posing a different type of big data challenge to the geoinformatics community, due to the particularities of the analysis that these data support, a hybrid mix of spatial and social analysis, as we will see in Sections 11.3 and 11.4 of this chapter. At the same time, these social media datasets have some particular characteristics that differentiate them from traditional geospatial datasets. We can identify three such particular characteristics as follows:

1. Social media datasets are streaming big data that are best-suited for real-time analysis. As public participation in social media is increasing very rapidly, the information published through such sites is clearly exceeding big data levels. For example, in 2012, Twitter users were posting nearly 400 million tweets daily, or over 275k tweets per minute (Forbes, 2012), doubling the corresponding rates of 2011 (Twitter, 2011). At the same time, 100 million active users are uploading daily an estimated 40 millions of images in Instagram.† Furthermore, every minute, Flickr users upload in excess of 3000 images (Sapiro, 2011), and YouTube users upload approximately 72 h of video‡ (YouTube, 2013). Accordingly, one can argue that, compared to traditional geospatial big data, social media impact more the velocity component of the diagram in Figure 11.1. Furthermore, because of their nature, social media are primarily suited to communicate information about rapidly evolving situations, ranging from civil unrest in the streets of Cairo during the Arab Spring events (Christensen, 2011) or New York during Occupy Wall Street (Wayant et al., 2012) to reporting natural disasters like a wildfire (De Longueville et al., 2009) or an earthquake (Crooks et al., 2013). Accordingly, not only are social media data available at streaming rates, but also their analysis must often be done at comparable rates, in order to best take advantage of the unique opportunities that they introduce.

2. Social media data are non-curated and therefore their reliability varies substantially. Traditional geospatial data collection campaigns are based on

---

† http://instagram.com/press/#
‡ http://www.youtube.com/yt/press/statistics.html
meeting strict accuracy standards. This tends to produce datasets that are authoritative and fully reliable. In contrast, social media content is uncontrolled and as such displays substantial variations in accuracy. As is the case with other online media outlets, social media is not immune to the dissemination of misinformation. In Figure 11.2, we show, for example, how after the Boston Marathon bomb attacks the wrong names of suspects were circulated in Twitter traffic for few hours before the real suspects were identified. Nevertheless, studies indicate that information dissemination patterns can be analyzed to assess the likelihood that a trending story may be deemed likely to be true or not (Castillo et al., 2011). Furthermore, the massive membership of social media is offering another mechanism to verify breaking stories. As we see in Figure 11.3, real events reported from the ground during a demonstration tend to be clustered in space and time, making it easier to verify content reliability (Croitoru et al., 2011). These two examples demonstrate quite vividly both the unique suitability of social media to report breaking stories, as well as the variability in accuracy of their content, which are both distinguishing them from traditional data sources.

3. The spatial distribution of social media contributions is nonuniform and heavily skewed. Traditional geospatial data offer broad and consistent coverage: satellites provide global remote sensing imagery coverage with high revisit frequency (from hours to few days), and numerous local campaigns complement that foundation by providing additional data. Social media contributions do not follow this pattern though. Their content reflects the participation and contributions of a particular demographic. Internet users under the age of 50 are mostly involved with them (with those in the 18–29 age group slightly leading the 30–49 group in terms of participation), college graduates participating more than nongraduates and women also slightly leading men (Duggan and Brenner, 2013). Each particular service has its own profile, but they all tend to fall under these general participation patterns. The users are also living primarily in urban settings, and this is affecting the spatial coverage of these feeds.

Social media contributions display a spatial distribution that reflects the aggregation of the previous trends. They tend to initiate more from urban, technologically advanced areas, rather than rural remote locations. Figure 11.4 is quite representative of this pattern of predominantly urban contributions, with tweets in the broader New York area reporting the Virginia earthquake in the summer of 2011. We can see that the reports are most heavily contributed from the Manhattan region, and their density is dropping substantially as we move away toward suburbs and the less urban regions in New Jersey. Nevertheless, these contributions also often originate from locations of interest to the general public. Figure 11.5 shows a sample of the variety of imagery that has been contributed to Flickr from the Chernobyl area (site of the 1986 nuclear disaster).
FIGURE 11.2 A graph showing how Twitter users reported the names of the wrong suspects following the Boston Marathon bombing of April 15, 2013. The horizontal axis is time of the day, extending from the evening of the 18th (leftmost point) to the morning of the 20th (rightmost point). The vertical axis represents the log of tweet mentions for each suspect name. The orange and purple lines correspond to the names of two wrongly accused individuals, while the green, blue, and red lines are the last (Tsarnaev—green) and first (Dzokhar, blue; Tamerlan, red) names of the actual suspects.
Figure 11.3  Clustered geolocated tweets (red dots) and Flickr imagery (green dots) reporting events as part of Occupy Wall Street’s Day of Action on November 17, 2011. (From Croitoru A. et al., Towards a collaborative GeoSocial analysis workbench, COM-Geo, Washington, DC, 2012.)

Figure 11.4  Spatial distribution of tweets reporting the Virginia earthquake in August 23, 2011. (From Crooks A.T. et al., Trans. GIS, 17(1), 124, 2013.)
11.3 GEOSOCIAL COMPLEXITY

In addition to the particular characteristics that we identified previously in Section 11.2, geosocial data are also differentiated from traditional geospatial datasets due to their complexity. More specifically, they are predominantly linked information; links are established among users to establish a social network and among words to define a storyline that is communicated through pieces of information. Regarding user connections, they can be established through specific actions. For example, user A following, replying, referring to, or retweeting user B can establish a connection among them. Aggregating all these connections provides us with a view of the users as a structured, networked community that can be represented as a graph.

In Figure 11.6, we show a sample network, generated by considering retweet activities. The network was constructed using Twitter data discussing the 2012 US presidential elections. Nodes (dots) correspond to users, and lines connecting them indicate instances where a user retweeted another. This graph representation offers detailed insights into the structure of a social network. We can identify, for example, a bigger cluster of nodes on the left, formed around nodes 687, 685, 667, and 642, and their connected groups. These groups are interconnected among themselves to form a larger cluster, in which these four nodes play a leading role (as they lead the corresponding smaller groups). We have used different colors for each group to better communicate individual group structure and the points where two groups connect to form a larger community. On the right side of the same figure, we identify a smaller cluster, formed around nodes 1046, 1044, and 805 (and their corresponding groups), that remains disjoint from the left cluster. We also observe smaller disconnected
group formed around node 635 (blue cluster at the upper middle) and some other further disjoint small clusters at the top and bottom of that figure.

Similarly to users, words are also linked, as they appear in the same messages, to form word networks that define the discussion themes. In Figure 11.7, we show a snapshot of such a word network. The size of the node corresponds to the frequency of this word, and we can see how *obama* is a keyword that is emerging from the discussion, followed by *romnei, bain, barack, patriot*, etc. Connections among words are established every time two words appear together in a single record (e.g., a tweet or the caption of an image), and the thickness of the edge connecting two words in this network is proportional to the frequency of the two words appeared together in the same tweet. It should be noted that in this word network, we use the stemmed
form of words, a process that reduces (normalizes) words to a common root form to reduce the effect of form variations on the word frequency count. For example, the word *Romney* would be stemmed to the form *romney*, and the word *people* would be stemmed to the form *peopl*. A detailed review of stemming algorithms can be found in (Hall, 1996).

While we tried to use a small data corpus to visualize these networks, it is clear to see that they can become very complicated very fast. A collection of tweets captured during Hurricane Sandy (October 2012) resulted in a data corpus of nearly 10 million tweets, including over 4.7 million retweets (47.3% of the data corpus) and over 4.1 million (41.7% of the data corpus) links to external sites (e.g., Flickr photos, news websites). We can understand that analyzing such highly complex datasets is a computational challenge, which is typically addressed through graph analysis. Enumerating interesting patterns, for example, triangles in a large network, is a computationally expensive task. This task requires a listing algorithm, which requires at least one operation per triangle, leading to a worst-case lower computational bound of $\Omega(n^3)$.
A variety of solutions exist to support feature extraction from such network graphs by constructing (or extracting) variables from them. These features (variables) typically encode the position and importance of each node with respect to the other nodes in the graph. Centrality, topological, and distance measures are fundamental metrics to support such analysis:

- **Centrality metrics**: Centrality is one of the most important structural attributes of a social network (Freeman, 1979). Often, it is defined in terms of reachability of a given node in a network. Some of the centrality measures can also be defined in terms of flow. In recent years, a great many measures have been proposed, including degree, closeness, betweenness, and flow betweenness (Scott, 2012). Degree centrality refers to the network activity for a node, often measured as the number of connections a node has. The betweenness measure reflects the influence of a node in the network, often measured as the number of times a node acts as a bridge along the shortest path between two other nodes (Wasserman et al., 1994). The closeness measure refers to the visibility of a node. Closeness can be extracted in terms of geodesic distances or eigenvector measure. PageRank (Page et al., 1999) and the Katz centrality (Katz, 1953; Borgatti, 2005) are two other measures that are closely related to the eigenvector centrality measure. Computing these features on large graphs is superlinearly expensive, and scalability of these algorithms in terms of three graph partitioning methods and GPU implementation were studied by Sharma (2011). Some of the centrality measures like betweenness (shortest path or random walk) are computationally expensive –O(n^3). It is noted in Kang (2011) that parallelizing (graph partitioning) these measures is not straightforward. Therefore, implementing these measures in main memory systems like uRiKA is an important research task.

- **Topological features**: Topological features are dependent on abstract structure of the graph and not on the representation. Commonly used features include degree distribution, clustering coefficient, and connectivity. The degree distribution is probability distribution of the degree (of nodes) over the entire network. It is an important measure, as random graphs often have binomial or Poisson distributions, whereas real-world networks are highly skewed (Bollobas et al., 2001), that is, the majority of their nodes have low degree but a small number of them (also known as hubs) have high degree. In scale-free networks, the degree distribution follows a power law. The clustering coefficient determines the degree to which the nodes tend to cluster together (Huang, 2006). Another reliable measure is the effective diameter, which is measured as the minimum number of hops in which some fraction of all connected pairs of nodes can reach each other.

- **Similarity and distance measures**: Searching for similar patterns in heterogeneous information networks is a challenging task. Heterogeneous networks are directed graphs, which contain multiple types of objects or links. There exist several similarity measures for networks with same type of nodes or links. However, the same is not true for heterogeneous networks.
Recently, several new similarity measures have been defined for heterogeneous networks (Sun, 2012). A particular similarity measure, called PathSim (Sun, 2011), was shown to find many meaningful objects as compared to the random-walk-based similarity measures. Authors have also provided prediction and clustering algorithms using this similarity measure.

11.4 MODELING AND ANALYZING GEOSOCIAL MULTIMEDIA: HETEROGENEITY AND INTEGRATION

Currently, social media services offer a wide range of platforms using various technologies and platforms. As a result, their content tends to be very diverse both in content—ranging from text to photos and videos—and in form, ranging from structured content to semi- or non-structured content. In addition, the form of raw social media data tends to be unstructured or ill-defined, making valuable knowledge hidden and limiting the capability to process it through automation (Sahito et al., 2011). For example, both Twitter and Flickr provide application programming interfaces (APIs) to query their content, but their responses are often structurally incompatible, not only between services but also within a single service. For example, the Twitter API can return the same content in different formats depending on the particular API calls made. Managing and integrating such diverse social media data requires the development of a unified conceptual data model that will support the various data structures under a single scheme. Generally, this task can be viewed as a data-cleaning problem, that is, the removal of errors and inconsistencies in databases (Rahm and Do, 2000), from either a single source or multiple sources of data. For multiple data sources, data-cleaning problems can arise at the schema level (e.g., structural inconsistencies) or at the instance level (e.g., uniqueness violations).

A step toward a more general solution for integrating social data was recently presented by Lyons and Lessard (2012), who introduced a social feature integration technique for existing information systems that are not socially oriented. However, there seem to be no widely established universal conceptual models that could be directly applied to multiple social media sources for geographical analysis. For example, Sahito et al. (2011) presented a framework for enriching and deriving linkages in Twitter data by using semantic web resources, such as DBpedia* and GeoNames†. However, this work considers only a single social media source, that is, Twitter. A more generalized conceptual model has been recently introduced by Reinhardt et al. (2010), in which two networks are considered: a set of artifact networks that describes the relationships between data elements (e.g., chats, blogs, or wiki articles) and a social network. This model, Artefact-Actor-Network (AAN), is created by linking the two networks through semantic relationships. Another closely related work is the data model presented by Shimojo et al. (2010), which focuses on lifelogs: digital records of the experiences and events a person encounters during a period of time, which are generated by individuals

* http://dbpedia.org/
† http://www.geonames.org/
(Kalnikaitė et al., 2010). Their work presents a model that is geared toward the integration of multiple heterogeneous social media sources through the introduction of a common model for lifelog data.

Building on this prior work, we present here a data model for converting social media data into structured geosocial information, from which knowledge can be extracted through further analysis. Focusing on Twitter (text) and Flickr (imagery), Figure 11.8 depicts an entity–relation (ER) diagram of our data model. It is comprised

![ER Diagram](image)

**FIGURE 11.8** An ER diagram of the geosocial media data model.
of several primary building blocks, namely, entry, geolocation, time, keywords, and authors. As some of these components may have many-to-many relationships between them, additional tables were added to represent such relationships. In the following discussion, we describe these components and the relations between them:

- **Entry**: An information entry serves as an abstract record of social media data, a high-level entity that binds the various components that constitute a data entry. An entry instance is uniquely identified and is linked to one author instance of the author component. In addition, each entry is associated with a time stamp indicating when the entry instance was published.

- **Source API**: The source API data are comprised of the data elements that are unique to each social media source. For example, the Twitter API returns a tweet attribute that contains the content of a tweet, while the Flickr API returns a photo URL and has no equivalent to the tweet attribute. Such source-specific attributes, which are driven by the characteristics of each social media source, are regarded as source dependent and are therefore stored separately in dedicated tables.

- **Author**: Author instances represent social media service users that contributed content. As user identification across sources is rather limited, each social media service creates its own author namespace, that is, a unique set of user names. As an author is identified by a tuple of a user name and a social media service identifier, different users can have the same identifier value in different services. It should be noted that authors can be referenced to in the content of social media feeds, through which the underlying social network can be recovered. In our model, this is accomplished by linking entries to users through the *mentions* table.

- **Geolocation**: As discussed earlier, geolocation information for social media feeds can be inferred indirectly from content analysis, or it can be extracted directly from the data itself. The contributors themselves may directly provide geolocation information, either in the form of exact coordinates or as a toponym (e.g., listing a city name) that in turn can be geolocated using any gazetteer service (e.g., Lieberman et al., 2010). A more detailed discussion on the various forms of this geolocation content can be found in Marcus et al. (2011), Croitoru et al. (2012), and Crooks et al. (2012). In our data model, geolocation information is stored in the form of coordinates, alongside information about the source through which they were acquired.

- **Time**: Temporal information can typically be found in all social media platforms. In the case of Twitter, temporal information is provided in the form of a time stamp of submission time (instance when it was submitted to Twitter by the user). In Flickr, the time tag can represent three different types of time attributes: posted time (the actual upload time), taken time (the time stamp embedded in the image metadata), and last update time (the last moment that the description of the picture in Flickr has been updated). In the data model, time information is embedded with each entry instance along with a time stamp-type identifier.
Keywords: As part of a social media entry, users contribute keywords or tags, like hashtags (#) in Twitter (Huang et al., 2010) or user tags in Flickr (Ames and Naaman, 2007) to emphasize views and ideas and engage other users (Romero et al., 2011). Hashtag usage, for example, has been shown to accelerate data retrieval from Twitter (Zhao et al., 2011) and Flickr (Sun et al., 2011). Hashtags also support the building of semantic networks by allowing individual tweets to be linked thematically based on their content. Unlike explicit tagging, implicit keyword may emerge from user conversations, when certain words become widely adopted as a result of a noteworthy event. In the data model, unique keyword instances are stored in a separate tags table, which is linked to the entries table.

Source: The social media source component (and its corresponding sources table) serves as a directory of the different source types that can provide data to the analysis framework. Each source can be associated with one or more feeds, which are instances of a given source type. As there may be situations where a single entry is present in multiple feeds (e.g., an identical tweet that is retrieved from two different feeds using different queries), only a single copy of the tweet is stored and a reference to the feed is made to avoid data redundancy along with a unique identifier.

Based on this integrated geosocial multimedia data model, an analysis framework can be designed. The analysis of geosocial multimedia data requires a novel analysis framework that expands the capabilities of traditional spatial analysis to account for the unique components of social data, for example, social links and non-geographic content (Wayant et al., 2012). Toward this goal, we present a framework for collaborative geosocial analysis, which is designed around data harvesting from social media feeds and a collaborative analysis workbench geared toward producing actionable knowledge. This framework is designed to enable and support distributed synthesis, knowledge production and information dissemination by communities of subject matter experts (SMEs), and other end users. To support such activities, the framework is designed around four primary themes: data harvesting and gathering; analysis tools, including spatial, temporal, and social analyses; support for SME communities interaction; and support of end-user communities that process data and interact and provide feedback. Each of these components should be modular and extensible, to allow the framework to adapt to a wide range of problem domains, for example, a new social media feed can be easily added without a need to redesign existing data handling and functionality capabilities.

An overview of this framework is shown in Figure 11.9. The process starts by gathering data from social multimedia, using, for example, specific regions of interest (ROIs), time intervals, and user-provided keywords. The harvested social media data are then analyzed and disseminated to the user community. Analysis results can then be used to refine the data gathering process by refining, for example, the ROIs, time intervals, or the keywords used for the analysis. Social multimedia is therefore not regarded simply as an additional data source in this process, but rather as a primary driver for the entire analysis and knowledge generation workflow.
The continuous refinement process is dependent upon the ability to effectively collect and store social multimedia (as was outlined in the geosocial multimedia model). A general architecture of a system to collect data and extract geosocial information from multiple social media feeds comprises three components (Figure 11.10): data collection from the corresponding social media data providers via APIs; data parsing, integrating, and storing in a resident database; and data analysis to extract information of interest from them.

Once data are collected and processed through the social media ingester, analysis of the data can take place through a collaborative workbench (Figure 11.11). An underlying design goal of this workbench framework is to foster knowledge discovery and support the collaboration of SME communities through a collaborative web-based
platform. Using such platform, the collection, aggregation, analysis, and synthesis of geosocial data can be carried out and disseminated to geographically distributed user communities. At the same time, end-user communities are also empowered to interact with SME communities, to respond and contribute insights and knowledge by providing feedback and inputs to the analysis process. To accomplish this design goal, the workbench is designed around several interrelated enablement principles (Croitoru and Arazy, 2007):

- **Spatial enablement**—the capacity to represent and reference events in a real-world spatial setting. Representation includes the ability to model data for different spatial entities, for example, points, lines, and polygons, as well as other high-level representations of events, especially spatiotemporal (see, e.g., Laube et al., 2007). Such enablement also includes the ability to store, query, and retrieve spatial data and support auxiliary thematic layers (e.g., census data).
- **Temporal enablement**—the capacity to associate data entities with a time stamp. Time is essential not only for understanding when things happen but also for evaluating data relevancy. In addition, temporal enablement allows users to make predictions through modeling, thus enabling users to combine historical observations and future predictions in their analysis.
- **Social media enablement**—the capacity to harness social media feeds for geosocial analysis (e.g., Twitter or Flickr) and issue data collection requests. The collection mechanism provides the capability to harvest social media...
based on location and time, as well as based on thematic keywords (or tags), therefore supporting a theme-based query mechanism for social media.

- **Collaboration enablement**—the capacity to empower groups of SMEs and user communities to collaboratively forge a data corpus for addressing a common goal and then analyze it toward the production of actionable knowledge. Such enablement includes the ability to collaboratively isolate and gather relevant data and process and analyze it toward the production of knowledge.

- **Geosocial analytics enablement**—the capacity to embed social media in a spatiotemporal domain for aggregation, analysis, and derivation of analysis products, both qualitative and quantitative. Such tools are provided to users in the form of thematic toolboxes, for example, a social network analysis (SNA) toolbox, a spatial analysis toolbox, and simulation and a prediction analysis toolbox. Based on these toolboxes, users can create and share subject matter workflows for processing social media data, thus allowing experts to collaboratively perform analysis. It is worth noting that the workbench also enables the creation of subject matter toolboxes in order to encapsulate domain-specific expert knowledge into analysis tools.

### 11.5 OUTLOOK: GRAND CHALLENGES AND OPPORTUNITIES FOR BIG GEOSOCIAL DATA

The motivation for this chapter comes from the unprecedented developments in social media and the resulting effects on actively and passively contributed geographic information as we discussed in Section 11.1. Big data, especially big geosocial data, are only expected to become more bigger as time progresses. We have already seen this occur with the proliferation of smartphones and other location-aware devices, and we do not expect it to stop in the near future. We would argue that big geosocial data provide us with unique opportunities to collect real-time data in epic scale and geolocate this information for analysis. Through geosocial analysis, we can gain a greater understanding of various parameters that shape the human landscape. For example, people’s tweets can act as sensor with respect to earthquakes (e.g., Crooks et al., 2013) or highlight potential hot spot emergence of political events (e.g., Stefanidis et al., 2013). Notwithstanding its application to various other avenues of human activity including languages (e.g., Mocanu et al., 2013), elections (e.g., Tumasjan et al., 2011), riots (e.g., Tonkin et al., 2012), emotions (e.g., Quercia et al., 2012), happiness (e.g., Bliss et al., 2012), college aspirations (e.g., Wohn et al., 2013), disease spreads (e.g., Achrekar et al., 2011), and flooding (e.g., Graham et al., 2012), one could consider these as altering the notions of how we explore geographical and social systems. For example, we can observe the collapse of a physical infrastructure as it is affecting its people (e.g., Sakaki et al., 2010) or the collapse of a social system while leaving the physical infrastructure intact such as in some cases of the Arab Spring (e.g., Howard et al., 2011). In a sense, such data streams harvested from humans acting as sensors have similarities to how one uses rain and stream gauges to monitor flooding in early warning systems.
Big geosocial data therefore offer us a new lens to study human systems. Traditional geographical analysis of human systems often focused on large-scale studies of population aggregates, remotely sensed imagery, or on small-scale studies of small groups of individuals. This was dictated by data availability limitations, as census, satellites, and individual interviews were the only practical means available to collect information on societal aspects and their distribution in space. These limitations kept us from getting a fine resolution and extensive coverage view of societal issues, problems, trends, etc. Similar to the progress that the invention of the microscope brought to biology or the telescope to astronomy, we are now witnessing a comparable paradigm shift in our ability to observe and analyze sociocultural expressions in space. Our ability to harvest geolocated social media feeds and crowdsourced data is offering unprecedented opportunities to monitor people’s actions, reactions, and interactions at a unique combination of fine spatiotemporal resolution and massive coverage.

But just like the invention of the microscope that introduced new types and volumes of data in biology and fostered the emergence of novel analysis approaches and microbiology, so too the emergence of social media and crowdsourced data presents the need for a new geosocial analysis paradigm as was discussed in Sections 11.2 through 11.4. For the first time, we can bridge the gap between the aforementioned large-scale and small-scale analyses, and we can begin to understand intrinsic connections between individuals, groups, and their environments based on sociocultural similarities. The discovery through novel social network and spatial analysis techniques of up-to-now invisible sociocultural patterns in space will advance our understanding of how our increasingly networked society is shaped, operates, and evolves in space. This is particularly applicable to urban areas, as they are home to the vast majority of the world’s population; they are also home to the majority of the geosocial media crowd, and furthermore, their urban form and infrastructure define to a great extent activities and behaviors of the individuals. This urban setting will serve as the stage for our analysis.

However, there is also a growing discussion that big data need a big theory. For example, West (2013) argues we need a deeper understanding of complexity science, in the sense that we are looking at a system of many parts that can interact in many different ways that adapts and evolves through space and time. Big data offer us a way to observe these interactions, but we do not have a unified framework for addressing questions of complexity. To fully address this issue, we also need to link cross-disciplinary expertise that will foster the further development of transdisciplinary knowledge. For example, by linking SNA with geographical analysis, using information harvested from geosocial media feeds and crowdsourced data, we gain a new understanding of space as it is defined by and in turn defines the interaction of people within it. For the first time, we can test notions such as Tobler’s (1970) first law of geography that “everything is related to everything else, but near things are more related than distant things.” This has already been shown to be valid with respect to Wikipedia spatial articles (Hecht and Moxley, 2009), but is it true for other aspects of geosocial media? We can also explore on a massive scale time (time–space) geography that has seen of much research since the 1960s (Hagerstrand, 1967).

However, even with the growth of geosocial media, one has to note that there is an issue of only getting a sample of the population when collecting information from
geosocial media feeds, namely, individuals who are active in this arena. Nevertheless, this sample is rapidly growing, as relevant technology adoption is becoming more ubiquitous (see Smith, 2011; Nielsen, 2012). That said, however, researchers need to consider new measures of engagement (e.g., contrasting rates with say Internet usage) along with other normalization methods (e.g., Stefanidis et al., 2013a) when looking at results from geosocial data. Coinciding with participation rates are issues relating to false information that can range from state-sponsored propaganda such as the 50-cent party that is supported by the Chinese government and pays for blog or message-board postings that promote the Chinese Communists Party line in an attempt to promote public opinion their way (see Bremmer, 2010) to that of spam and falsification of events (e.g., Gupta et al., 2013) that are active areas of research in the data mining community.

This rise in geosocial media and the ability for analysis also raises several concerns with respect to the suitability of traditional mapping and GIS solutions to handle this type of information. We no longer map just buildings and infrastructure, but we can now map abstract concepts like the flow of information in a society and contextual information to place and link both quantitative and qualitative analyses in human geography. In a sense, one could consider AGI to be addressing the fact that the human social system is a constantly evolving complex organism where people’s roles and activities are adapting to changing conditions and affect events in space and time. By moving beyond simple mashups of social media feeds to actual analysis of their content, we gain valuable insight into this complex system. What is to come next is difficult to predict. For example, consider that only 10 years ago the idea of location-based services and GPS embedded into mobile devices was still in its infancy. Advances in tools and software made geographic information gathering easier, resulting into growing trends in crowdsourcing geographical data rather than using authoritative sources (such as national mapping agencies). More recently, the popularity of geographically tagged social media is facilitating the emergence of location as a commodity that can be used in organizing content, planning activities, and delivering services. We expect this trend to increase as mobile devices become more location aware. One could relate this to the growing usage of online activities and services (such as real-time information on geosocial media sites like Foursquare, Facebook places, Google Latitude, Twitter, and Gowalla and a host of new sites and services emerging with time). But also more static sites (in the sense, one can upload when wants), such as Flickr and YouTube, provide means of viewing and in a sense forming an opinion of a place without actually visiting.

There is also the issue of privacy with respect to geosocial media. Nobody wants to live in Orwell’s version of Big Brother; however, harvesting ambient information brings forward novel challenges to the issue of privacy, as analysis can reveal information that the contributor did not explicitly communicate (see Friedland and Sommer, 2010). But this is not a new trend; it has actually been happening for a while now. Google itself is basically a marketing tool using the information it collects to improve its customer service. Similarly, Twitter makes money by licensing their tweet fire hose to search engines, while companies can pay for promoted tweets (see Financial Times, 2010). And this trend has already spread to locational information. For example, TomTom (2011) has been using passively sensed data for
helping police with placing speed cameras. This comes alongside Apple’s iPhones storing locational data while the user is unaware (BBC 2011). However, people are making progress in highlighting the issue of privacy relinquishing when sharing locational information. Sites and apps such as pleaserobme.com or the creepy (http://ilektrojohn.github.io/creepy/), a geolocation aggregator, have demonstrated the potential for aggregating social media to pinpoint user locations. Trying to protect people’s identities in times of unrest is also a well-recognized concern, for example, the Stand by Task Force (2011) suggests ways of limiting expose and delay information for the recent unrest in North Africa. But the power of harvesting AGI stems from gaining a deeper understanding of groups rather than looking at specific individuals. As the popularity of social media is growing exponentially, we are presented with unique opportunities to identify and understand information dissemination mechanisms and patterns of activity in both the geographical and social dimensions, allowing us to optimize responses to specific events, while the identification of hot spot emergence helps us allocate resources to meet forthcoming needs.

REFERENCES


TomTom (2011) This is what we really do with your data, Available at http://www.tomtom.com/page/facts (accessed on June 3, 2013).