A CyberGIS Environment for Near-Real-Time Spatial Analysis of Social Media Data

Shaowen Wang
CyberInfrastructure and Geospatial Information Laboratory
Department of Geography and Geographic Information Science
National Center for Supercomputing Applications
University of Illinois at Urbana-Champaign
Email: shaowen@illinois.edu

Social media, such as social networks (e.g. *Facebook*), blogs and micro blogs (e.g. *Twitter*), and photo/audio/video sharing services (e.g., *Youtube* and *Flickr*), can be understood as Internet-based applications that are built on the ideological and technological foundation of participatory Web, and allow for the creation and exchange of user generated content (Kaplan and Haenlein 2010). These online applications and associated data generated have been experiencing a spectacular rise and popularity. Over short periods, hundreds of millions of users have been attracted to these services and generating massive quantities of social media data with unprecedented spatiotemporal scales and extents.

Twitter, for example, has rapidly gained popularity. Although each *tweet* is limited to only 140 characters, the aggregate of millions of *Tweets* may capture dynamic patterns of various topics of interest at the scale of large populations. This new data modality has become increasingly important to the development of human knowledge. For example, the Library of Congress has begun archiving *Twitter* feeds. Extensive studies with significant societal impacts have been conducted by capitalizing on social media data, ranging for example from predicting stock market (Bollen *et al.* 2011), tracking infectious diseases (Signorini and Segre 2011), to measuring public opinion and political sentiment (O'Connor and Balasubramanyan 2010).

Social media have also been recognized as proxies to understand geography (Leetaru *et al.* 2013). Intentionally or unintentionally, people are sharing their whereabouts when using social media services. With widespread of location-aware mobile devices and continuing improvements of location-based services, location-based social media data are becoming increasingly available. Such massive, dynamic, geo-referenced data, despite privacy concerns and quality issues such as noises and possible spams, offer an unprecedented opportunity to understand micro-dynamics of complex social systems across multiple spatiotemporal scales. To gain timely insights and desirable knowledge from social media data, however, poses several fundamental challenges.

Firstly, location-based social media data are often 'big' and "coming" continuously, considering the case of daily new tweets across the globe and even extending the time window to a number of months or years. The magnitude of this data volume is well beyond the capability of any mainstream geographic information systems (GIS). Especially, data access and analytics may not be achievable within a reasonable amount of time without resorting to advanced cyberinfrastructure strategies.

Secondly, social media data are generated dynamically and continuously. Users of social media services are allowed to frequently update or change their status and locations, and for certain emergency events, volunteers can rapidly contribute their information and experiences. These near-real-time crowdsourcing data, complemented with official and authoritative data sources, become especially valuable in such time-critical cases as disaster response and relief (Goodchild and Glennon 2010). Conventional GIS approaches, however, are limited to support timely analysis of such dynamic and massive social media data. While near-real-time spatial analysis of social media data is desirable particularly for time-critical cases, it is computationally intensive and, thus, requires high-performance computing.

Thirdly, in contrast to well-structured geospatial data sources, social media data are often produced in unstructured forms. Extra efforts, such as applying data mining techniques, are often necessary to make such data meaningful and sensible. In addition, social media services usually do not provide direct access to all the data being produced, which causes data access to be a nontrivial

problem. Researchers have to come up with their own *ad hoc* mechanisms to obtain data of particular interests, typically via designated access interfaces provided by social media services. Issues of uncertainty and noises further compound this data access problem, which hinders applications of these data sources to broad use.

CyberGIS, a new modality of GIS based on advanced cyberinfrastructure, is established through the synthesis of advanced cyberinfrastructure, GIS, and spatial analysis and modeling capabilities (Wang 2010). Early research and development of cyberGIS have demonstrated its great potential to address significant challenges of geographic information science and various geo and spatial fields (Wright and Wang 2011). The ongoing National Science Foundation cyberGIS initiative (www.cybergis.org) has made steady progress on advancing the science and applications of cyberGIS, particularly for enabling the analysis of big spatial data, computationally intensive spatial analysis and modeling, and collaborative geospatial problem solving and decision making (Wang *et al.* 2013a).

With an open framework and a concrete implementation, this paper suggests a cyberGIS environment for efficient collection, management, access, analysis and visualization of location-based social media data, *Twitter* feeds in particular, with a focus placed on addressing the aforementioned challenges (Wang *et al.* 2013b). By seamlessly integrating a system of collecting and managing location-based *Twitter* data, a suite of near-real-time spatial analytical services, and an advanced cyberinfrastructure environment with high-performance computational resources, this user-centric cyberGIS environment provides a near-real-time means to explore spatiotemporal patterns hidden in massive *Twitter* data.

Acknowledgements

This material is based in part upon work supported by the U.S. National Science Foundation under grant numbers: 0846655 and 1047916. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Bollen, J., Mao, H., and Zhang, X. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1): 1-8.

Goodchild, M., and Glennon, A. 2010. Crowdsourcing geographic information for disaster response: a research frontier. *International Journal of Digital Earth*, 3(3): 231-241.

Kaplan, A., and Haenlein, M. 2010. Users of the world, unite! The challenges and opportunities of social media. *Business horizons*, 53(1): 59-68.

O'Connor, B., and Balasubramanyan, R. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In: *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pp. 122-129.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., Shook, E. 2013. Mapping the global Twitter heartbeat: The Geography of Twitter. *First Monday*, 18(5), DOI:10.5210/fm.v18i5.4366.

Signorini, A., and Segre, A. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS One*, 6(5): e19467.

Wang, S. 2010. A cyberGIS framework for the synthesis of cyberinfrastructure, GIS, and spatial analysis. *Annals of the Association of American Geographers*, 100 (3): 535-557.

Wang, S., Anselin, L., Bhaduri, B., Crosby, C., Goodchild, M. F., Liu, Y., and Nyerges, T. L. 2013a. CyberGIS software: A synthetic review and integration roadmap. *International Journal of Geographical Information Science*, DOI:10.1080/13658816.2013.776049.

Wang, S., Cao, G., Zhang, Z., Zhao, Y., Padmanabhan, A., and Wu, K. 2013b. A cyberGIS environment for analysis of location-based social media data. In: *Location-Based Computing and Services*, *2nd Edition* edited by A. K. Hassan, CRC Press, pp. 187-205.

Wright, D., and Wang, S. 2011. The emergence of spatial cyberinfrastructure. *Proceedings of the National Academy of Sciences*, 108(14): 5488-5491.