Managing Data in Large-Scale Participatory Sensing Applications

Farnoush Banaei-Kashani
Researcher at Computer Science Department
Associate Director of NSF's IMSC[1]
University of Southern California
Los Angeles, CA 90089
Email: banaeika@usc.edu
Web: persis.usc.edu

Cyrus Shahabi
Professor at Computer Science Department
Director of NSF's IMSC
University of Southern California
Los Angeles, CA 90089
Email: shahabi@usc.edu
Web: infolab.usc.edu

Our background and experience in the field

Both authors come from a database research background, with less focus on conventional/core databases and more interest in addressing data management research issues in the context of atypical data systems, such as peer-to-peer databases, geospatial databases, sensor databases, and multimedia databases. Most relevantly, authors have led a participatory sensing technology development project, termed *GeoSIM* (short for *GeoSocial Image Mapping*) [1], as part of the on-going participatory sensing project [2] at NSF's Center of Embedded Networked Sensing (CENS) [3]. GeoSIM is an urban sensing system with which a group of individuals with camera-equipped mobile phones participate in collaborative/social mapping of the urban image (i.e., the texture of the urban environment) at some target geolocation. The participating group, which may either consist of dedicated individuals or the general public, manually capture geotagged images of the urban environment. The collected images are progressively used for documentation of the dynamic urban image in multiple spatial resolutions at different times. We have published on GeoSIM's enabling technologies [4,5] and have developed a real-world GeoSIM prototype [6].

More recently, with our *iCampus* project [7], we are developing a geospatial social networking web-portal for the entire USC community (including students, staff, and faculty) to enable participatory data sharing in the context of various applications including ambience monitoring in learning environments (e.g., lighting condition in classrooms), health analysis (e.g., pollution, noise and UV exposure assessment), and evacuation planning.

Our vision

Many studies suggest significant future growth in the number of mobile phone users, the phone's hardware and software features, and the broadband bandwidth. Therefore, a transformative area of research is to fully utilize this new platform for various tasks, among which the most promising is *participatory sensing*. Participatory sensing (PS), aka people-centric sensing, engages individuals, groups, and communities in the act of collecting, analyzing, and disseminating urban, social, and other spatiotemporal information. This new paradigm of data collection has shown to be useful when traditional means either fail (e.g., due to disaster), censored or do not scale in time and space.

We argue that large-scale implementation of PS applications imposes new unique data management challenges that, unless addressed efficiently, will hinder applicability of PS in real-world scenarios. These data management challenges are spread across the entire lifecycle of the participatory data as exemplified below:

1. *Data Collection*: Regardless of whether PS is deployed in directed or opportunistic mode, existing approaches for "optimal planning/scheduling" of data collection are unscalable, particularly

[1] The Integrated Media Systems Center (IMSC) is a National Science Foundation (NSF) Engineering Research Center (ERC) at USC, established in 1996.

considering the dynamic, autonomous and error-prone nature of participatory data collection (see [4,5] for more details). New approaches are required that allow for on-the-fly planning, while scaling as the number of participants grows.

2. *Data Communication*: Some data modalities (e.g., video data or high frequency GPS samples) require high bandwidth and are costly to communicate for the participants. Without efficient "data representation" solutions that mitigate the data communication cost, participants are less likely to participate, rendering PS practically inoperative.
3. *Data Verification*: Once the participatory data is collected and contributed, an important issue is how to verify that the collected data are valid. Without effective "data verification" techniques that are applicable with dynamic and autonomous PS, we will be in a garbage-in-garbage-out situation, rendering PS useless.
4. *Data Storage* and *Querying*: As PS applications scale in number of participants, the size of the collected participatory data grows rapidly. OLAP (online analytical processing) techniques have been introduced to store and query data at such scales; however, these techniques fail to apply to the participatory data which are intrinsically uncertain and sparse.
5. *Data Visualization*: Visualization of the participatory data is often required for enabling data exploration/analysis or to incentivize participation (e.g., visualizing traffic congestion conditions to incentivize users to share their GPS locations). New data visualization techniques are required that can handle the dynamicity and sparsity of the participatory data.

Evidence that pursuing our vision will lead to major advances in the field

We are observing upward trends in both the phone's data transmission capabilities and in the number of people using phones; the multiplication of the two suggests the exponential growth for PS usage. On the one hand, the mobile phone's bandwidth is constantly increasing: from 2.5G (up to 384Kbps) to 3G (up to 14.7Mbps) and recently 4G (up to 100 Mbps). On the other hand, we observe the unbelievable growth in mobile phone subscription that has surpassed 4.35 billion worldwide by the end of 2009, a global penetration rate of 71.3 percent (a population of 6.8 billion by year 2009 worldwide). In fact, by 2013, Gartner predicts mobile phones will overtake PCs as the most common Web access device worldwide. However, even though participatory sensing has proven valuable and the technological trends only point to its growth, without scalable solutions to manage PS data, PS will be limited to academic and toy applications without realizing its full benefit in real world.

References

[1] GeoSIM project website at CENS: http://urban.cens.ucla.edu/technologies/geosim/, and GeoSIM project website at USC's Information Laboratory (InfoLab): http://infolab.usc.edu/projects/GeoSIM/

[2] CENS participatory sensing project website: http://urban.cens.ucla.edu/

[3] CENS website: http://cens.ucla.edu/

[4] H. Shirani-Mehr, F. Banaei-Kashani and C. Shahabi, "*Efficient Viewpoint Selection for Urban Texture Documentation* ", GSN 2009.

[5] H. Shirani-Mehr, F. Banaei-Kashani and C. Shahabi, "*Efficient Viewpoint Assignment for Urban Texture Documentation* ", ACMGIS 2009.

[6] F. Banaei-Kashani, H. Shirani-Mehr, B. Pan, N. Bopp, L. Nocera, C. Shahabi, "*GeoSIM: A GeoSpatial Data Collection System for Participatory Urban Texture Documentation* ", Special Issue of IEEE Data Engineering Bulletin on Spatial and Spatiotemporal Databases, June 2010.

[7] iCampus project website at USC/NSF's Integrated Media Systems Center (IMSC): http://imscwww.usc.edu/project.php