Towards Web-Scale Geo-Semantic Crowd Discovery

JAMES CAVERLEE
Assistant Professor
Computer Science and Engineering
Texas A & M University
Tel: (979) 845-0537; **Email**: caverlee@cse.tamu.edu

Introduction and Opportunity

There is a growing need to fundamentally advance research for enabling a new generation of applications for monitoring, analyzing, and distilling information from the prospective web of real-time content that reflects the current activity of the web's participants. Highly-dynamic real-time social systems like Twitter, Facebook, and Google Buzz have already published terabytes of real-time human "sensor" data in the form of status updates. Coupled with growing location-based social media services like Gowalla, Foursquare, and Google Latitude, we can see unprecedented access to the activities, actions, and trails of millions of people, with the promise of deeper and more insightful understanding of the emergent collective knowledge ("wisdom of the crowds") embedded in these activities and actions. Toward the goal of web-scale social media mining and inference, our lab (http://infolab.tamu.edu) is pursuing a set of related research activities, two of which are briefly described here: (i) Identifying and tracking the evolution of semantic crowds; and (i) Social media location estimation.

Identifying and Tracking the Evolution of Semantic Crowds

First, we believe that "crowd-based" information holds the key to effective modeling and understanding of the real-time web. A single user action—for example, posting a picture of a smoke plume to Flickr—though perhaps interesting itself, does not convey a strong community or social-based importance to the user action. In contrast, a flurry of activity associated with a "crowd" is a strong indicator of an emergent online phenomenon that may be worth identifying and directing to interested users. We refer to these ad-hoc collections of users that reflect the real-time interests and affiliations of users as semantic crowds. Unlike the more static and perhaps staler group-based membership offered on many social networks, semantic crowds are naturally organic and reflect highly-temporal group affiliation.

Identifying coherent crowds in real-time from the massive scale of the real-time web across a collection of non-obviously connected user actions is a major challenge. Considering Twitter alone, there are potentially 100s of millions of active users inserting new messages into the system at a high-rate. Concretely, we consider three overlapping crowd perspectives: (i) communication-based, reflecting groups of users who are actively messaging each other, e.g., users coordinating a meeting; (ii) location-based, reflecting groups of users who are geographically bounded, e.g., users posting messages from

Houston, Texas; and (iii) interest-based, reflecting groups of users who share a common interest, e.g., users posting messages about a presidential debate. While each of these crowd perspectives can be studied separately, in many cases it will be important to study cross-cutting crowds, e.g., users in Houston (location-based), messaging each other (messaging-based) about a local fire (interest-based). In addition to identifying a crowd at a point-in-time, we must additionally track the crowd over time as users join, crowds merge, and disband (as in Figure 1).

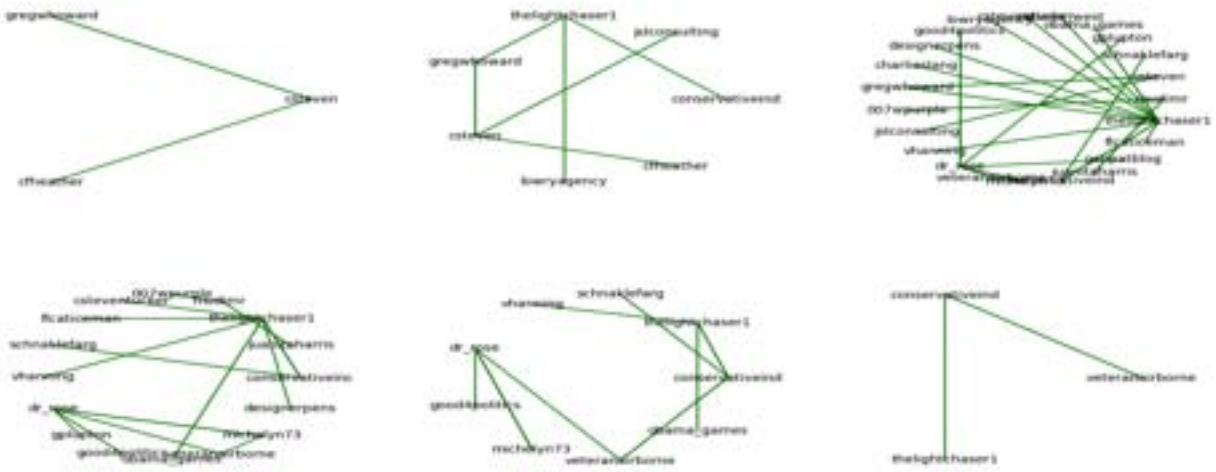

Figure 1: Example crowd growth and dispersal over 6 hours on July 23, 2010

Social Media Location Estimation

Second, the potential of social media as a medium for geospatial and temporal research is potentially limited by the aversion of participants to engage in and use location-revealing technologies. The increasing popularity of location-based social media (including Facebook Places, Google Latitude, Foursquare, etc.) belies the slowness of the vast majority of social media users to adopt geospatial features. To illustrate, in a random sample of over 1 million Twitter users, we find that only 26% of users have listed a location as granular as a city name and that fewer than 0.42% of all tweets actually use geotags. To overcome this location sparsity problem and to support our overall objectives of web-scale social media mining, we believe it is necessary to develop novel algorithms for automatically estimating a user's location through an analysis of the publicly-available data in a user's profile and the social media community itself. By relying only on publicly-available data, these algorithms can be generalized across social media sites and future human-powered sensing systems for providing accurate and reliable location estimation without requiring expensive or proprietary data from system operators or privacy-sensitive data from users.

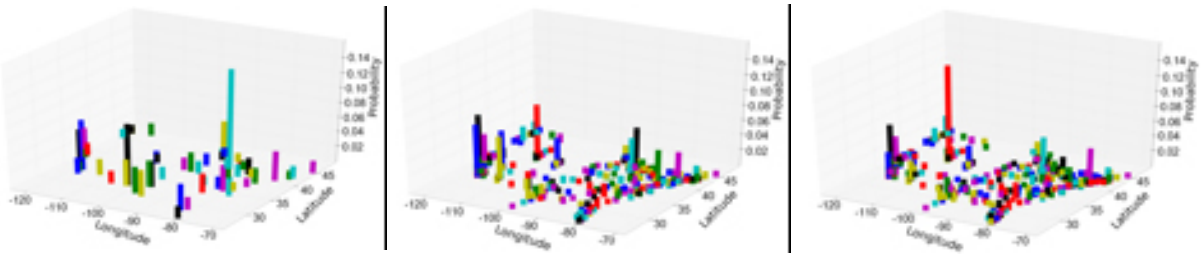

Figure 2: *Example: Location Estimation Convergence as Number of Tweets Increases*

As an illustration, we have built a simple probabilistic model that estimates the likelihood of a user in a particular location "emitting" a word. By aggregating over all of the words that a particular user posts (e.g., on Twitter), we can infer the user's most likely location. In essence, the hope is that as a user continues to tweet, more location-sensitive information is "leaked" which can be used to refine the user's location estimation. We find that the location estimates converge quickly (needing just 100s of tweets), placing 51% of Twitter users within 100 miles of their actual location. Figure 2 illustrates the increasingly refined geo-location estimate for a test user with an actual location in Salt Lake City, converging after only 500 tweets.

By augmenting the enormous human-powered sensing capabilities with algorithmically-derived location information, this framework can overcome the sparsity of geo-enabled features in these services and bring enhanced scope and breadth to emerging location-based personalized information services. This in turn could lead to even broader applications of social media in time-critical situations such as emergency management and tracking the diffusion of infectious diseases.