Detection and Prediction Anomalies in Twitter Stream using Geographic Information

Liron Marcus

Advisor: Dr. Ron Bekkerman

Social media services such as Twitter are generating an ever increasing amount of information on daily events. The ability of monitoring their large-scale data stream can provide nearly real-time detection of events – from local festivals to nation-wide natural disasters. This turns Twitter into a social sensing tool (Sakaki et al., 2010) which allows researchers to answer a variety of questions, such as how information propagates between users given its local context and cultural aspects.

Twitter itself has developed techniques to geotag tweets based on the user's coordinates at the time of tweeting, however the major challenge is to extract location information from tweets that were not geotagged.

In this work, we present a novel graph-based semi-supervised approach to detect locations in the tweet content. Unlike existing approaches, which either focus on geographical keyword extraction (at the granularity level of a country, state/region, city) or attempt to predict where the tweets was issued, we aim to detect locations that tweets refer to at lower level (down to level of a street or a point of interest).

We construct a graph of Twitter users and geographic mentions, and overcome the ambiguity of geographic names by exploiting the local property of the graph: given a set of unambiguous geographic names, we bound the ambiguous names based on their proximity to unambiguous names in the graph. For this purpose, we propose a new probabilistic model for label propagation in a large-scale graph and derive a parallelized algorithm for optimal labeling of nodes in a large-scale graph. We call the resulting methodology *Splash Label Propagation* (SLP).