Active discovery of location-based twitter data

Rambabu Thakkalapelli
Advisor: Dr. Christopher Homan
B. Thomas Golisano College of Computing and Information, Rochester Institute of Technology

Introduction

- Social media is comprised of platforms that collect data from individuals share their interests, activities, professional careers etc.
- Many social networking sites and communities help to bring people together based on their common interests.
- Our study focuses on the data collected from the social networking sites and tries to locate an individual geographically using relevant data.

Objective

- This research analyzes different methods and approaches for determining where an individual is geographically located based on his/her own social connections.

Background

- Davis Jr et al. (2011) - used voting algorithm, in view of the majority of each user’s friends’ locations from social connections of significant users in a particular area.
- Chen et al. (2013) - used interest of the users, in three phases: interest detection, mapping from the location function to interest, and location estimation.
- Sadilek et al. (2012) - friendship prediction on the sim-ilariites between two users, and identifying the amount of overlapping data(time, location and vocabulary) between the two users.
- Cheng et al. (2010) estimates user’s city location using a probabilistic framework based on the user’s tweets.

Results

- Precision, Recall and F1 Scores (Averages from 6 experiments) of Logistic Regression, Decision Tree, Gradient Boosting and Random Forest for all 2704 labelled in an interval of 500 users.

Conclusions and Future Work

- The best technique showed to be gradient boosted trees with a prediction accuracy of 96% but with high false negatives.
- In future this work can be done using Reddit, Neural Networks.

References


Design and Implementation

Design and Research Flow

- Data Collection – Twitter
- Building Meteor Web Application
- Label the user – Manually
- Cohen Kappa Annotation
- Generate the at-graph

Methods

- Baseline classification using network features
- Graph features
  - Average path length
  - Cluster coefficient
  - Largest connected component
  - Number of connected components
  - Number triangles on graph
  - Graph Degree
  - Coreness of Graph
  - Diameter of Graph
  - Graph Closeness Centrality Score
- Logistic Regression, Decision Tree, Gradient boosting, Random Forest
- Using Active learning to boost performance

Contact:
Rambabu Thakkalapelli - rt4469@cs.rit.edu
Dr. Christopher Homan - cmh@cs.rit.edu