Using Twitter Data to Infer Users' Demographics

(760 Final Project Proposal)

Xinyi Liu, Mingren Shen, Faust Shi

Research topic:

We are trying to combine geographical data to improve the efficiency of machine learning methods in Twitter data Mining. To be more specific, we plan to mine the gender information and emotion state from Twitter data that has been collected and preprocessed by Professor Qunying Huang of Department of Geography at UW-Madison.

Normally, to mine the gender information, people try to first use the information from the user's reported first name, tweet content and even the color of user's profile. SVM and Bayesian Network were proved to be good models and were able to get a relatively high test-accuracy. Based on previous work, we try to include user's travel pattern features like Twitter frequency at certain type of locations, and use ensemble training methods to optimize the classification results.

Data Descriptions:

We have got 254698 tweets of 8614 users in City of St. Louis, MO from 4:12:06 AM Sep. 11, 2010 to 5:49:50 AM Jul. 6, 2014 (data provided by Prof. Qunying Huang). This dataset is for use of a larger project while this course project is a part of it.) Each tweet consists of following informational fields which are useful for the classification object: content, hashtag, tweet zone type, user account name, user's first name, user's last name and so on. We will divide these data into 2 gender types: male and female.

We also got 20050 testing tweets of 20050 users with gender labels (one tweet for each user) from Kaggle platform.

Existing approaches:

1. Census, normally done by government
2. Machine learning approach:

Supervised learning, unsupervised learning. Supervised learning contains decision tree classifiers, linear classifiers (SVM, Neural Network and etc.), rule-based classifiers and probabilistic classifiers (Naive Bayes, Bayesian Network and Maximum Entropy). Among these methods, SVM and Bayesian Network were used to learn gender distributions based on multiple features and got good results. Effective features which have been discussed include typical word frequencies from several pieces of a user's tweet content, user's first name, user profile color and so on.

Timeline: Nov.16 to Dec. 14

Week1(2017.11.16 -- 2017.11.19) and **Week2**(2017.11.20 -- 2017.11.26)
(1) Project Implementation (2) Discuss with Professors
Week3(2017.11.27 -- 2017.12.03) and **Week4**(2017.12.04 -- 2017.12.10)
(1) Extra time for adjustment
Reporting Week (2017.12.11 -- 2017.12.14)