CS579 Project Proposal

Jiada Tu A20306906
Lv Zhang A20305221

Problem Statement

People may want to know how many “retweets” their next tweet will get. Base on it, they may find a way to increase their retweets. Company may need it for advertising. Organizations may need it for notification spread out. It is about message propagation.

Hypothesis

- It is possible to predict the number of retweets

- Some attributes like location info, amount of followers, images/videos, posted URL or number of days passed by, etc. will affect the number of retweets

Overview of Method

—Use some machine-learning algorithm to training and predicting the # of retweets and Favorite

—Use the attributes like:

1) # of followers/followings/ratio of them
2) If has hashtag/mention
3) # of hashtag/mention
4) If has images/videos/http link/location info
5) log(time passed by)

Overview of Method—continue

6) what the tweet said? pos./neg., length, etc.

7) User profile

8) more……

—Use Bootstrap Aggregating or Random Forest to deal with the bias-variance problem

Data Usage

—Use both Streaming & REST API to collect data

—Start with several user, and then some of they're following and followers. Get the time_line of them. (also, the related attributes)

—The training data we want is at least 200MB with purge text. As large as possible.

—Collect data from this/next week until … as long as possible/ or until we get the data size we need

Timeline Evaluate

—Collect the relevant data — as long as possible

—Compare multiple algorithm and find one to support our app. And implement our app. We may also find some way to refine the idea. Also add some attributes to the vector presents tweet

---------- (Three weeks)

Analysis the result and check the accuracy of predicted value and refine

---------- (Three weeks)

Related work

http://ciir-publications.cs.umass.edu/getpdf.php?id=1071 //also use image pixel as attributes

http://research-srv.microsoft.com/pubs/141866/NIPS10_Twitter_final.pdf

https://gigaom.com/2013/04/26/predicting-twitter-popularity-is-all-about-probability/

http://homepages.inf.ed.ac.uk/miles/papers/icwsm11.pdf // use passive-aggressive algorithm

Questions?