CS341: Project in Mining Massive Datasets

Predicting Information Virality on WeChat

Project Proposal

Submitted by:

Ruomiao Guo

Yuchen Li

Sijun He

1 Problem Description

WeChat started as a mobile text and voice instant messaging service by Tencent in 2011 and now provides a variety of features including social network, public accounts, payment, and location service. WeChat's social network component, Moment, allows users to share pictures, statuses and external post with the users' choice of friends and has become the dominant social networking service in China with more than 650 million monthly active users.

External posts in Moment may get **reshared** by friends: a user shares a post with his/her circle of friends and some of these friends share it with their friends and a **cascade** is developed along the way. The goal of the project is to develop a framework to quickly predict the final number of reshares of a post, given the number of reshare up to a time T.

2 Data

While we haven't had the opportunity to examine the data ourselves, we believe that we have access to all WeChat data. In additional to the data of reshare histories of all the posts, there are three aspects of data that we are currently considering.

- User profile data: name, gender, age, location and etc.
- Social Network data: underlying friendship networks between users, chats, likes, public account follows and etc.
- Payment data: WeChat Wallet activities like sending money between users and paying for services.

3 Approaches

We have two potential approaches to the problem, a featured based approach and a point process based approach. We may do both approaches or may pick only one and will make the decision later when we get our hands on the data.

Featured Based Approach
This approach is inspired by the work[1] of Cheng et al., where they developed a feature based framework to predict whether a cascade will continue to grow in the future and implemented it on a large sample of photo sharing cascades on Facebook. Instead of predicting the final size of the cascades, which some

argued to be unpredicable, the framework turns the problem into a classification problem: given a cascade that currently has a size k, predict whether it will grow beyond $f(k)$, which is the mean of the eventual sizes of all cascades that reach size k. Cheng et al. made use of a range of features, which could be grouped into four major categories: content, original poster/resharer features, structural, and temporal features.

We consider our project similar to the work of Cheng et al. and feel most of the features mentioned above could also be retrieved from our WeChat dataset. Furthermore, since the resharing of posts in WeChat contains considerably richer text content compared with the Facebook photo reshares, we would have much more content features from text mining. While Cheng et al. found content features to be weak predicators, it would be interesting to see how content features performs in our case.

Point Process Based Approach

This approach is inspired by the work[2] of Zhao et al., where they developed a statistical model SEISMIC based on the theory of self-exciting point processes and evaluated their model on one month of complete Twitter data. The two key components in SEISMIC are the human reaction time, which is the time it takes for a person to reshare a post, and the infectiousness, which defines the probability that a given user will reshare a given post. By combining the two, the speed at which post will spread through the network can be accurately modeled. The human reaction time is modeled through the memory kernel $\phi(s)$ and can be estimated using the distribution of retweet times of a sample set of the tweets. The post infectiousness $p_t(w)$ is estimated through Maximum Likelihood Estimate (MLE) with points close to time t. SEISMIC predicts the final popularity combining the memory kernel, the post infectiousness and the underlying post infectiousness.

The main advantage of using the point process based approach is that it imposes no parametric assumptions and requires no expensive feature engineering, compared with the feature based approach. SEISMIC also is very scalable with a computational time linear in the number of observed reshares. Our doubt about the SEISMIC is that Twitter's unique limited short messages and time-sensitive characteristics may not directly extend to Moment WeChat since Moment is only one feature out of many in WeChat that people spent time in. Furthermore, the underlying network structure is different between WeChat and Twitter, as twitter features a directed graph of ”followers” while WeChat is a undirected graph of ”friends” similar to Facebook.

4 Performance Indicator

The performance indicator for the featured based approach would be the classification accuracy and the area under the ROC curve (AUC) of the 10-fold cross validations. The performance indicator for the point process based approach would be the absolute percentage error (APE) as a function of time.

5 Plan

We will implement one or two approaches proposed above based on WeChat datasets and features that can predict the final reshare count for a post.

6 Group Member Bios

Ruomiao Guo
Ruomiao Guo is a second-year master student in EE department, working on the Computer Software and Hardware track. She did her internship in Altera last summer to reconstruct the internal testing flow for scalability. She enjoys the materials covered in CS246 and wants to implement the advanced algorithms to real datasets to get deeper understanding in this field.

Yuchen Li
Yuchen Li is a second-year CS Master student focusing on data management and analytics. He'll be graduating this summer and work at Google afterwards. He had several internships. The most recent one was at Yahoo to work in the backend dev team of a large Advertising Bidding system. He took CS246 in 2015 and really wants to do a course project on data mining.

Sijun He
Sijun He is a first-year master student in Civil & Environmental Engineering(CEE) department who has a passion for data science. Apart from his experience in CEE, he interned for Baidu last summer on a user profiling project and will be joining for Autodesk this upcoming summer as a data science intern. Sijun is currently taking CS246 and would love to gain hands-on experience on massive data mining through a project on the social network app he frequently uses everyday.

References

[1] J. Cheng, L. Adamic, P. Dow, J.Kleinberg, and J.Leskovec. Can cascades be predicted? *Proceedings of the 23rd international conference on World wide Web*, 2014.

[2] Q. Zhao, M. Erdogdu, H. He, A. Rajaraman, and J.Leskovec. Seismic: A self-exciting point process model for predicting tweet popularity. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015.