Proposal 349 Yunfei Gao, Ting Liu, Wanxin Xu

1. What task? Why we are interested in this topic.

   Our topic is to do some research in social media data mining. We are interested in get some knowledge from the social media data. One important area is social-related advertising. For example, when we go through some webpages especially the shopping websites, those websites can always know what I want to buy most. Actually, it is some machine learning way to predict our likes based on our browsing histories. Thus, we decided to figure out how it works.

   Our task is to do the similar people expansion. More specifically, we need to calculate a similar group based on a seed group(the seed package) provided by advertiser. And we then test the those group with our test package to calculate the accuracy of our method.

2. How to get our data?

   We get our data from Tencent(a Chinese internet company). This dataset contains hundreds of seed group people, features of candidate group people, and features of seed crowd corresponding advertising features. All the dataset has been divided into train set and test set. The training set calibrates the users belonging to the seed package in the crowd and users who do not belong to the seed package (i.e. positive and negative samples). The test set will test whether our algorithm can accurately mark whether the user in the test set belongs to the corresponding seed package.

3. Which **features/attributes** will we use for your task?

   We have not decided it yet. From our observation of their dataset, we find that most of the sample data features are sparse. So, we have to find a way to deal with the sparse matrix.

4. What will our initial approach be?

   First we want to convert the dataset into one-hot encode and try to find some good features to represent the whole dataset, because the data is quite large.

   After that, we can use deep learning to calculate those features. A good cost function here is incredible.