CS498 Fall 2016

Social & Information Network

Project Midterm Checkpoint

Bangqi Wang (bwang34)

Kedan Li (kedanli2)

11/01/2016

Introduction

The project focuses on the Learning Social Circles in Network, which helps the user to organize their social networks. The goal of this project is to cluster the ego network of the central user according to the features of the users in that network. With the developing of online social network, organizing the social circles becomes more and more important. People need an effective way to manage their social network, such as, Google+, Facebook, and Twitter. However, managing the social circle is sophisticated and challenging because the data sets is extremely huge and there is no good way to automatically organize the social circles. We use the KMean algorithm to cluster the users and get a submission output.

Related Work

1. McAuly, J., & Leskovec, J. Learning to Discover Social Circles in Ego Network:

This paper defines a novel machine learning task for social circles. This paper considers the social circle problem as the clustering problem with overlapping and hierarchy. What's more, the paper creates the evaluation metrics that could work on unsupervised environment.

The strength of this paper is that the algorithm form the circle based on some common aspects and user can pinpoint the aspects that caused the circle to form. What's more, this paper uses a real, large data sets from Google and Facebook, and the result is credible. The paper test multiple clustering methods only on profiles, only on network, and on the both.

However, the weak of this paper is obvious. The final algorithm used is Multi-Assignment Clustering which focuses on the profile information and ignore the network information.

2. Yang, J., Mcauley, J., & Leskovec, J. (2013). Community Detection in Networks with Node Attributes. 2013 IEEE 13th International Conference on Data Mining. doi:10.1109/icdm.2013.167:

This paper introduces the technical method that clusters the users with the combination of the node attribute and the edge structure, CESNA.

The strength is that the CESNA method is the first overlapping community detection method that model the node and the dependency and the runtime is linear.

The weak is that the method cannot handle more general types of attributes and cannot cluster the attributes into “topics”. The method should rely on other sources except node attributes.

Implementation

Our algorithm is built upon an existing library. The algorithm user k-mean clustering to compute the social circle. It first loads the friend graph data and features data from file, separate the dataset into different groups by original people. Then it computes k-mean clustering twice, according to the features similarity, friendship similarity and both features respectively. The number of clusters is set to 6 if the total number of friends of the original people is below 250, and 12 if above. While doing clustering, the algorithm compute maximum and minimum similarity from centroid for each cluster and drop friends with small similarity from cluster if similarity difference is more than a given threshold.

Interpretation

We compare the result of our cluster with the result in the training set, and found that most clusters are largely consistent with a few elements that are mismatched. We presume that the method didn’t fully exploit the feature data but haven’t fully the graph connectivity.

Interpretation

Build upon the following library: https://github.com/j-a-c/LearningSocialCircles