Project Progress Report

Sensing Semantic Information from Mobile Social Networks

Ting Zhang and Wen Yi
School of Industiral Engineering, School of Electrical and Computer Engineering
zhan1013@purdue.edu, yi35@purdue.edu

November 5, 2015

1 Problem Definition

As explained in the project proposal, our project focuses on the study of social network information in two parts: Friendship Inference and Community Detection. In the following sections, we will explain these two parts separately.

1.1 Friendship Inference

Topology structures have been analyzed to study the friendship relations between two individuals. Mostly, researchers are building topologies based on personal information, such as hobbits, favorite movies and frequently visited places. And then, "similarity" between individuals are computed based on topological structures. However, in this project, we are not interested in topological approaches. Instead, we want to apply the machine learning algorithms we learnt from this class. Here, we will try three algorithms, including Decision Tree, Perceptron and Adaboost. Bag of features or attributes are extracted and explained in section 2. We denote the input feature set as X, and a feature i as $x_i \in X$. The output would be a boolean label indicating the relation of friend or not.

1.2 Community Detection

The problem of community detection is defined as removing the edge that is evaluated as most "crossing the communities" until the graph is divided into several components with high purity and modularity. For each iteration of the edge removing, we evaluate the "betweenness" of all edges by computing the shortest path between each pair of vertices inside the graph and add 1 weight to each edge in that shortest path. Then we remove the edge with the highest "betweeness" value inside the graph, and recompute the shortest graph of the vertices pair which is influenced by the edge removing, then change the "betweeness" of the edges accordingly.

By continuously removing the edge, we get the cluster treeview of the original friendship connection graph.

2 Data

The dataset originally consisted completed data from 94 subjects with 106 participants in total. A mapping is provided to get the complete data of 94 subjects from the whole dataset. However, in the mapping, we found an error index of participants 107. Therefore, only 93 participants' data are used in this project.

2.1 Friendship Inference

As explained in the project proposal, three types of information from the dataset are extracted and used as input feature for pari of individuals.

* Bluetooth information: The most commom begin time and duration of bluetooth meeting.

* Phone call record: The most common begin time and duration of phone call.

* Location record: The most common shared location and its corresponding begin time.

Here, we catgorized the begin time into seven groups: early morning (5:00am - 7:00am), morning (7:00am - 11:00am), noon (11:00am - 1:00pm), afternoon (1:00pm - 5:00pm), early evening (5:00pm - 8:00pm), late night (8:00pm - 1:00am) and midnight (1:00am - 5:00am).

2.2 Community Detection

The classic community detection problem is based on unweighted and undirected graph, while the edges in the graph should not sparse, which means at least the vertices from the same community should be connected with each other. After the initial study of the dataset, we notice that the data from the questionnaire indicating people's close friend circle and the time people usually spend with others every week are quite sparse. Thus, we need to use other type of data with larger data size for the community detection input.

The Bluetooth information in the dataset are collected by scanning the cellphones inside user's Bluetooth connecting range every 5 minutes within the 9-months experiment. This constantly scanning results in about 4 million proximity events in the dataset. By creating edge between the cellphone users or adding weight to the edge for each appearance of a user's cellphone on others' devices, we generated a weighted graph indicating the frequency of meeting between people. As the devices would continuously listing the devices while people were in long-time meeting, this generating function would automatically adding more weight for the long-time meeting.

After getting the weighted graph between people, we would extract the meeting frequency between each pair of people and combine it with the ground truth of whether the pair of people are in the same affiliation, then computes the best threshold for dropping the low frequency connection between people by evaluating the entropy in original class and maximizing the information after classifying the connections by the threshold. Then, we leave only high frequency connections between people as the potential friendship, which are the unweighted and undirected edges in the graph.

3 Baseline Results

3.1 Friendship Inference

As stated in the project proposal, baseline approach of frienship inference is using the community information of each individual. If two individuals are in the same community, then they are predicted as friends. The results of this baseline approach is with precision of 6.17% and recall of 75.3%.

3.2 Community Detection

As stated in the project proposal, the result of the community detection is measured by the combining purity of the whole clustering result. For the baseline of the detection, we treat the whole original graph as one community, then compute the purity of the community. The result is 35.5%.