Clique Finder

By Ryan Lange, Thomas Dvornik, Wesley Hamilton, and Bill Hess

Abstract

Facebook has a very large user base with a lot of data, and we thought that it might be interesting to see if we can use data mining to find different cliques that might exist within a person's friend list. This could be useful when you want to find relationships between certain groups of people within your own friends. This proposal will identify our problem, our data sets, and our proposed solution.
1. Introduction

Modern Social Networking sites like Facebook allow users to create a large network of friends. These sites allow users to coordinate events and share interesting information about what is happening in their lives. Users can become online friends with hundreds or even thousands of people. But when friend networks get this large, it can be hard to tell who is hanging out with whom, or who in your friends list actually knows each other.

We are going to develop a Facebook application that solves this problem using data mining. Using a clustering algorithm, we will find out how your friends naturally group together. You will be able to find out information like which of your old high school friends are still hanging out together without having to tediously look through albums or wall posts. We believe this app will be a valuable addition to Facebook and will give people the opportunity to find out new information about their friends.

2. Problem Statement

Cluster friends based on their profile information and the relationships they share with other users on your friends list, thus forming cliques or groups of people within your friend’s list that may not have been considered before.

3. Proposed Solution

We are going to develop a Facebook application that helps people discover the cliques within their friends list. This application will use the clustering technique (unsupervised learning) to group friends together. The distance metric for doing this will be left up to the user based on what kind of information they are interested in or seems most relevant. Some of these options include hometown, religion, images, networks, shared friends, or a combination of these.

We are interested in trying both bottom-up and top-down approaches in clustering to see which method gives the best results. The reason for using a bottom-up approach over a top-down is because then we do not need to know how many cliques exist in a user's friends, and this will be different for each user or for each distance metric the user chooses. However, if we could determine a suitable number of cliques for the user then we could try the top-down approach.

For the bottom-up method called hierarchical clustering that was mentioned in class to group the user's friends into cliques. The result of this algorithm is a tree known as a dendogram. To achieve grouping we will split the dendogram at a depth that gives a reasonable number of cliques for the user.

For the top-down method we would use the max-flow technique for community discovery. This would be useful in discovering cliques because it is more likely to find split groups that are weakly connected, unlike the bottom-up approach. This would give us better odds of finding sub-cliques within larger cliques.

To test our clustering algorithms we are going to create a fake dataset and compare the results for varying number of friends.
4. Data / Data Sets

Facebook’s API allows applications to access all data of users who accept the application. The data can be cached for 24 hours before it is erased from the application. A user’s data includes any entered data on their profile; such as gender, age, networks, relationship status, hometown, religious, status feed, etc. Any profile’s visible data can also be accessed; such as wall post, comments, messages, notes, pages, notifications, events, groups, links, streams, friends, pictures, videos, tags. Friend’s user data can be retrieved from within the application. Pictures and videos include all of the user’s pictures and videos as well as any pictures and videos that the user is tagged in.

Our training set will be Thomas Dvornik’s data. This set will include about 7000 pieces of personal data, over 40000 wall post, 350 photos, about 1600 tags, and tons of other miscellaneous data; such as status, events, links, notes, etc.

Our data set includes at least 10 heavy users. A heavy user being people who have more data then our training set. So about 70000 pieces of personal data, 400000 wall post, 3500 photos, 16000 tags, and much more miscellaneous data. Of course, some of this data will be repeat information, which will not be considered when analyzing users individually.

5. References