Amanreet Bajwa

KDD Cup

Introduction

The KDD Cup is a data mining challenge hosted by Yahoo! Labs. In this project we were given the task of creating a working solution for Track 1 of the KDD Cup 2011 Challenge. The dataset for this challenge consists of Yahoo! Music data. It is made up of tracks, albums, artists, and genres which all tie together. The different items (albums, users, tracks, etc.) are given meaningless anonymous numbers so that no identifying information is released.

Preprocessing

The first step in any Data Mining project is to take the time to explore and familiarize yourself with the data. This particular data set is made up of tracks, albums, artists and genres. Each track has a unique id, associated album id, associated artist id, and optional genres. Each album has a unique id, an associated artist id, and optional genres. The genre and artist files each hold a list of unique ids. Looking at this non-standardized text format the first step was to put it into a database. The database includes four main tables: albums, artists, genres, tracks, and two join tables: albums\_genres and genres\_tracks. First, I created a SQLITE database that met these specifications. Then I took the time to write a program using Active Record that handles reading in the data files and loading them into the correct places in the database. The loading process ended up taking a lot longer than I was expecting. It took around 2 hours to load all of the data in to my SQLITE database. At this point I realized I forgot to load the users and their ratings into the database so I went back and added a users and ratings table. The loading of the ratings file didn’t take nearly as long since there are only 11696 lines in the files.

Statistics

After all the data was loaded into the database my first thought was to see what exactly the item id rated in the ratings file was referring to. Is it just tracks, albums, and artists? Or does it also include ratings for genres? First I checked the number of ratings there were for albums which turned out to be only 1391. Then I checked tracks which was 7295.The tracks ended up being the most rated item which seems natural. Artists had 2487 and surprisingly there were 479 ratings for genres. I did not expect there to be any ratings for genres because it doesn’t seem like a common thing to rate to me.