Final Project Requirements

Max Wagner

July 7, 2016

System

The MovieLens data was put into a smaller, more compact version that includes only user and movies with suffient information by the University of Minnesota. One of the problems with the data set is that time is not taken into consideration when looking at user's recommendations. The set provides a timestamp for each recommendation, abit in a strange format. The timestamp is in number of seconds from midnight on January 1, 1970. Using this information, it will be possible to weight ratings that are closer to present day higher than older ratings.

The process will be completely done on Apache Spark, which includes the use of `SparkR` and likely a minimal SQL database. Ideally, this will allow for better management of data and queries than pure CSV. Additionally, multiple methods of recommendation will be used to find the most accurate prediction strategy.

Business Application

The most important aspect of weighting rating dates chronologically is that users' preferences change over time. A movie they liked 5 years ago, may not be a movie they like currently. Being able to more reliably recommend movies is critical to the success of sites like Youtube and Netflix, and likewise for this small project.

Apache Spark introduces the ability to use the project on larger and more computationally expensive data sets.

Audience

The theoretical audience this specific enhancement would benefit is the long-time users of a site. Users with years of ratings would benefit from weighting dates as they fit the profile in the business section above.

Data Source

The numerous verions of the MovieLens set can be found here. http://grouplens.org/datasets/movielens/latest/

For this project, the `ml-latest-small` version will be used.

Similar to other systems, `NA` values will be filled in with 0's. The alternative to this would be filling in with user averages, however most information seems to agree that 0's provide a better end result.