Project Proposal

Project Choice and Rationale

We decided to pick the second dataset (from grouplens) because we are both interested in understanding how recommender systems work.
Additionally, the dataset seemed the simplest to understand and preprocess. Therefore it will allow us to dedicate more time to exploring and comparing different data mining algorithms for our research.

Team Composition

The team is composed by Giovanni Ilacqua and Sharfuz Shifat.

Project Goals

Our goal is to build and compare the performances of different movie recommender systems. In particular, we would like to compare the results of both item-based and user-based collaborative filtering against a more sophisticated CF strategy, using clustering.

Planned Approach

For the simpler approaches we intend to measure item-item and user-user similarity, using Cosine Distance. To determine similarity between movies we will be considering the movie tags and other metadata, while to determine the similarity between users we will consider the users' previous movie ratings.
Once we have identified the similarity between data objects we will compute movie predictions using K-nearest Neighbors.

For the more sophisticated approach we plan to partition users with a clustering algorithm, such as DBSCAN. Then define a voting algorithm which we would apply to each partition separately, so that a movie would be recommended to a user *U* only by analysing the ratings given by the users belonging to *U*'s cluster.
One thing we want to consider is performing probabilistic clustering and use a cluster based Collaborative Filtering algorithm.

Role of Team Members

Since the team is made of only two members, there will not be a precise definition of roles. We both want to practice and learn Data Mining techniques, so we will both be contributing to the software implementation as well as to the report writing and research.