École Polytechnique Fédérale de Lausanne

EE558 - Network Tour of Data Science

Project : Build a network based recommendation system

Author:
Lukas De Loose
Niklas Glaser
Maxime Lamborelle
Nils Ter-Borch

Professor:
Frossard Pascal
Vandergheynst Pierre

Assistants:
Michaël Defferrard

December 4, 2019

1 Introduction

What movie should I watch to tonight? A very common, yet personal question, that movie recommender systems try to answer. In our project, we want to get insights in the way users rate movies and make a prediction on the most likely rating for a movie, as well as possible influences on the rating behaviour.

In comparison to most currently applied recommendation algorithms, which are based on single value decomposition, we want to use a network-based approach. We want to improve the user rating based recommendation system by taking into account movie features.

2 Exploring the data

Our main data source is the `MovieLens` database, which we will combine with data directly provided from IMDb, resulting in 3537 movies. Based on the `MovieLens` data, we can create a user network based on their ratings. This network is created as a epsilon-similarity-graph, which uses the difference in the commonly rated movies as a measure of distance. To get better insights into the graph and create clusters within the Network the nodes will be combined by an user network which looks especially at their features, which are:

- 20 types of occupations,
- the zipcodes divide into 10 groups of 1000,
- ages in the with a range of 10 years,
- and gender.

The movie features will be used for training our dataset which will be used as an graph signal to predict the rating, using convolution.The following key features are available from tMDb:

- Genre
- Descriptive Keywords
- Popularity
- Year

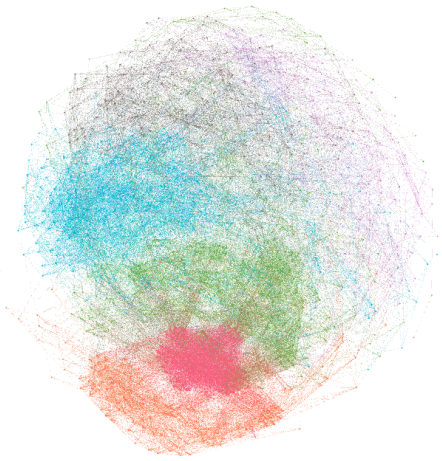
More features may be gathered if necessary. Each movie is linked to the movie lens database by a pair of unique ID's

3 Visualising the graphs

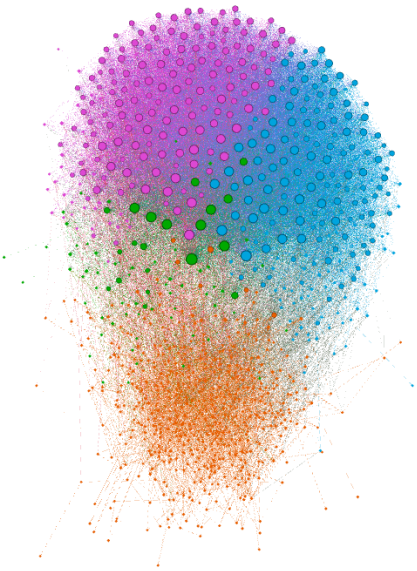
When creating these graphs, we expect to find clusters of people with similar interests. By visualizing our graphs with the Gephi tool we can create distinguishable areas in the feature graph, which were held in the graph by design 1a. On the other hand, we see that the ratings based graphs are distributed pretty homogeneously 1b. Which might lead to the hypothesis that we do not have a fanbase of a certain genre in this dataset, although this has to be further investigated. Still we can find separation by sparsifying and calculating the modulation, which is heavily connected to the other sections.

4 Exploitation

There are various techniques and approaches possible to reach the prediction. We will use semi-supervised learning, to predict how users would rate an unseen movie, so we can recommend them the one that they would rate the best. For this, we have to split our data in a train and test set. For the recommendation a graph convolution approach will be used. We will assign features from the movies to every node in the user graph. With the help of graph convolution and back propagation, we want to achieve a graph signal which contains a predicted rating. By doing this we can consider a loss function that evaluates the predictions we have, in a way that we can fill the gaps in our missing graph.



(a) User graph based on user features only.



(b) User graph based on user ratings of movies only.

Figure 1