Individual report for COMP7630 project

Movie Recommendation base on IMBD

Student No.19403844

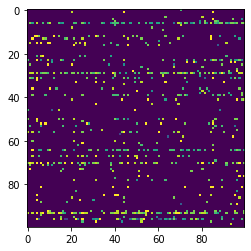
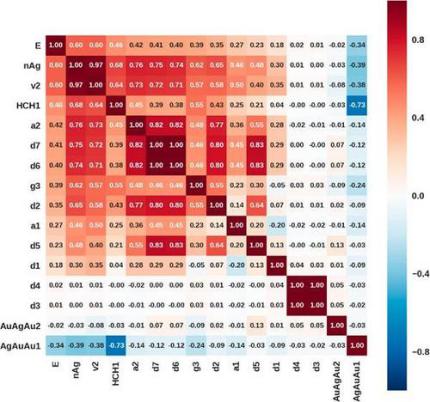
Student Name:Deng Guoquan

To some degree, I have written the most of work into project report. In this report, I will involve some experiment I fail and share some experience I gain in this project.

1, First and the most important hint, starting from the easiest assumption and model. In fact, at beginning of this project, I try to use a big model to train the whole data and it waste a lot of time and fail at the end. The reason why it fail is the data I achieve is not complete, there are too much data online and I can only crawl small part of them, so the data incomplete will become the noise which hurt the model. So after failure, I give up that purpose and make a data clean, data clean almost take up 50% time of the whole project. So starting from the easiest model and data set, after finding some useful pattern try to use other model powerful model to explore and enhance the performance. Base on that, trying to solve more challenge and generalize data set.

2, I also try some other model but they do not provide good enough result and I still find the reason. (1) I try association rule and try to analysis it base on graph model, the algorithm I use is fp-max which is used to find the maximum item set based on fp-growth, and visualized by spring layout. But I find that is not really realistic because the graph is very high density and I also find that the size of most item set is 2 which means that connection between movie is too sparse.

3, A very interesting idea is using hierarchical feature to improve Collaborative filtering. Although I fail in this project, I believe it is reasonable. As image, each movie often contain different topics, people who love that movie is because he he love a specific topic of that movie but all topics of that movie, this is why error happen in collaborative. So if the similarity computation of collaborative filtering is base on small set of movies which high relate to the target movie we want to estimate instead of the the whole set of movie, the performance of collaborative filtering can be improve. And how to do that, I suppose we can use association rule to find the most related movie set of target movie we want to estimate, and calculate the similarity based on this movie set.



If each movie just contain one topic the correlation movie matrix will similar to graph on left, only deep colored square at the diagonal line, but what we get is chart at right. The item set of association rule algorithm as feature can transform the feature map from right to left. More experiments need to be finished in the feature,

4, The initial purpose of me in this project is create some auto-tag model. This is the reason why the model I use usually based on the users’ history in my part in project report. The insight of all these models are using a subset of movie to identify the each user profile, and using a subset of user to identify the each movie’s feature ( the latter part I do not realized but you can image transposing the user-item matrix and train the model ).

5, I do not write a lot about network spider in the report but you can still find a lot of trick from the code.