Movie Recommendation System

Tianyu Gao, Xu Wang
The Wang Yanan Institute for Studies in Economics, Xiamen University

Abstract : This report mainly develop a recommender system to recommend 5 movies from 9000 movies to a specific user. Firstly, this system uses UB-CF, IB-CF and NLP to select 30 movies among 9000 movies, and then use Wilson interval and regular term to recommend final 5 movies. The recommended movies consider both user's preference and public aesthetics.

Keywords : Recommender System; UB-CF; IB-CF; NLP; Wilson Interval

1 Introduction

Nowadays, it's popular to make recommendation in internet companies, and it's also one of the key core competitiveness of TikTok. We try to use existing movies data and data mining methods to build movie recommender system.

In this report, three recall methods are used. Firstly, user-based collaborative filtering (UB-CF), means to recommend movies to you which people similar to you had admired. Secondly, item-based collaborative filtering (IB-CF), means to recommend movies with genre or main actor that user have always admired. Thirdly, natural language processing (NLP) on movies' text comments, means to recommend movies that have similar comments with users' movies comments. They are used in multiple recall which means the recalls are done simultaneously. Re-rank comes after recall. Wilson interval and regular term are used to make recommendation based on user's preference and public aesthetics. The final goal of this report is to recommend 5 movies to a specific user.

The structure of this report is shown as follows. Section 2 presents data source, data visualization and descriptive statistical analysis.

2 Data

2.1 Data Source

The dataset used in the report comes from ”Movielens”, which is provided by GroupLens, a research lab at the University of Minnesota. Specifically, we use ”ml-20m”, which is slightly old but comprehensive dataset in ”Movielens” released in 2015, and it is recommended for education.

The dataset describes 5-star rating and free-text tagging activity from [MovieLens](http://movielens.org), a movie recommendation service. It contains 20,000,263 ratings and 465,564 tag applications across 27278 movies. These data were created by 138,493 users between January 09, 1995 and March 31, 2015.

Links.csv: part of URL of the movies on IMDB and Rotten Tomato, which explains where the data are from.

Genome-tags.csv: connects the tags and the numbers which are used in the tag genome.
Genome-scores.csv: show the relevance between the tags and the movies.
Movies.csv: provides movies' genre and other relevant details
Ratings.csv: contains 20,000,263 movie ratings given by users
Tags.csv: contains 465,564 movie text comments given by users

2.2 Data Visualization

The movie genres contained in the dataset is given as figure 1.From the figure we can see that the most popular movie genres are Drama and Comedy, followed by Thriller, Action and Romance. However, Musical and War movies are less shown in the data.

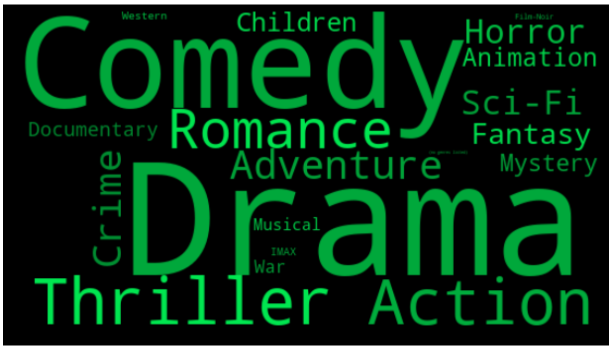


Figure 1: Word Cloud of Genre

Figure 2 shows the relationship between movies rating and rating time, and it's clear that there are more ratings and larger variance ratings for modern movies.

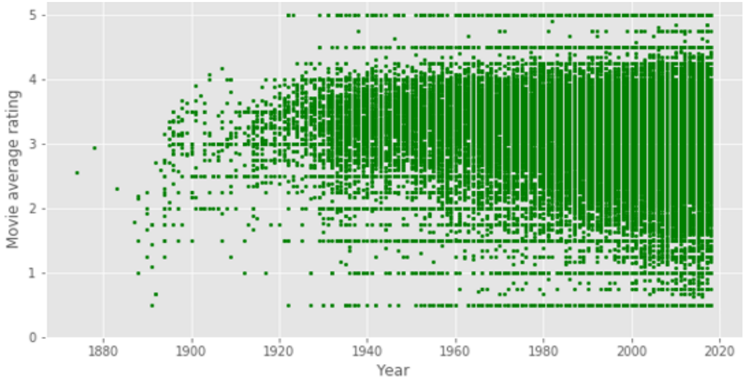


Figure 2: Ratings Change with Time

Figure 3 shows the empirical distribution of average movie ratings of each user. It's roughly normal but not that good, because there are many users tends to give 5 points all the time. In later proprecessing, we deleted the users whose ratings are all 5 because we could not distinguish the movies using their ratings.

Figure 4 and 5 show the relationship between movie releasing time and numbers of ratings and the ratings. Basically, users tend to mark movies released 20 to 30 years before, however, movies released in this time period tend to have lower ratings than previous movies. At the same time, users' rating got lower after 2000, which may because users criteria got higher or movies' overall quality got lower.

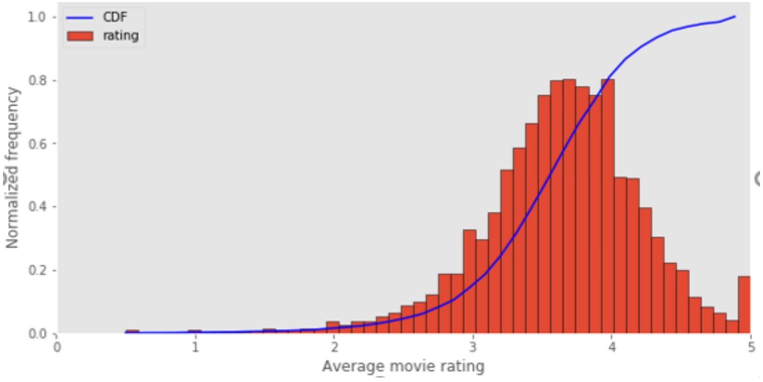


Figure 3: Users' Average Rating

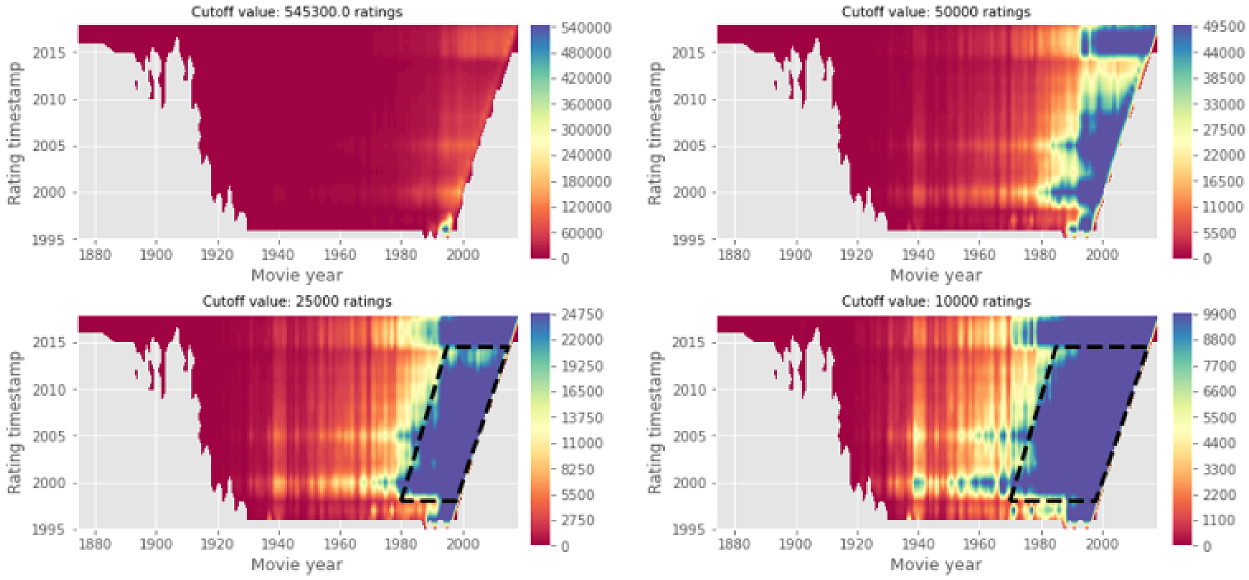


Figure 4: Number of Ratings per Movie

2.3 Descriptive Statistical Analysis

Figure 6 shows normality test of each genre's rating, it's roughly follow normal distribution for every genre's rating.

There is a outlier, meaning this user mark for about 9000 movies, as shown in figure 7. However, variance inflation factor suggests that ratings that this user gave to 9000 movies are in the rational interval, leaving no influence on regression results. Thus, it's highly possible that it is a official account, and we keep in that later studies.

To analyze the relationship between different genres, we try to discovery association rule between genre in figure 8. It is not surprising that different genres are connected, for example, Animation and Children positively connected with each other, Drama and Comedy negatively connected with each other.

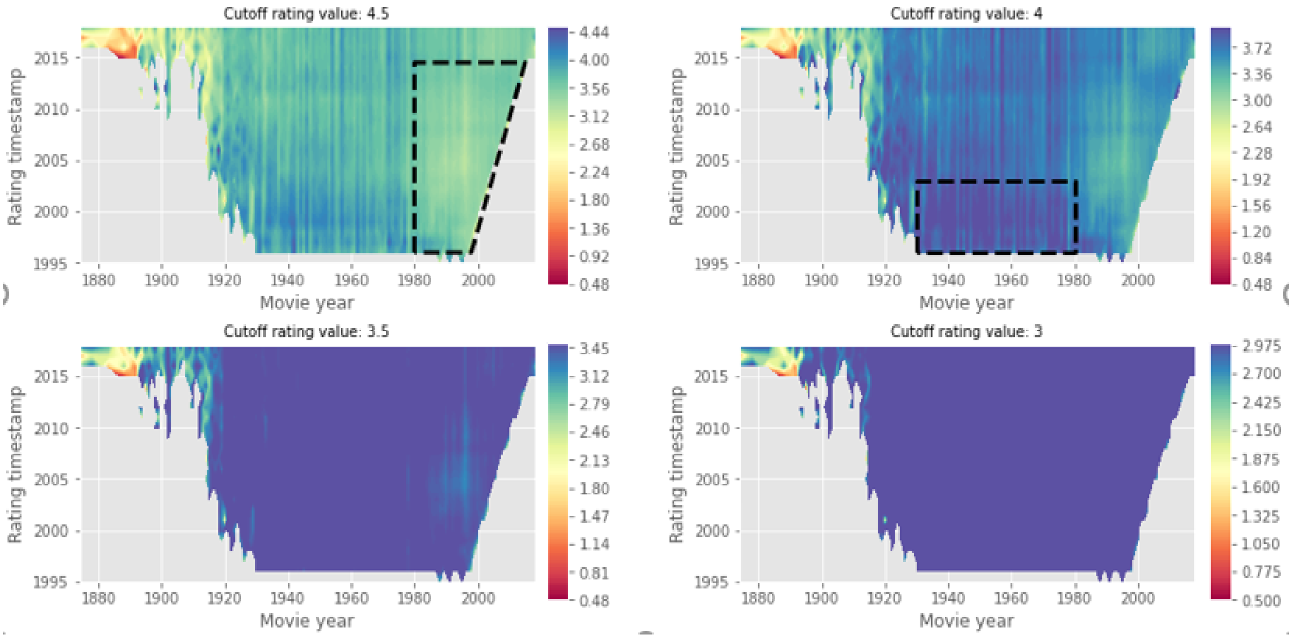


Figure 5: Movies' Average Rating

Figure 9 shows association rule between genres, it shows clearly that Thrill and Sci-Fi are highly correlated.

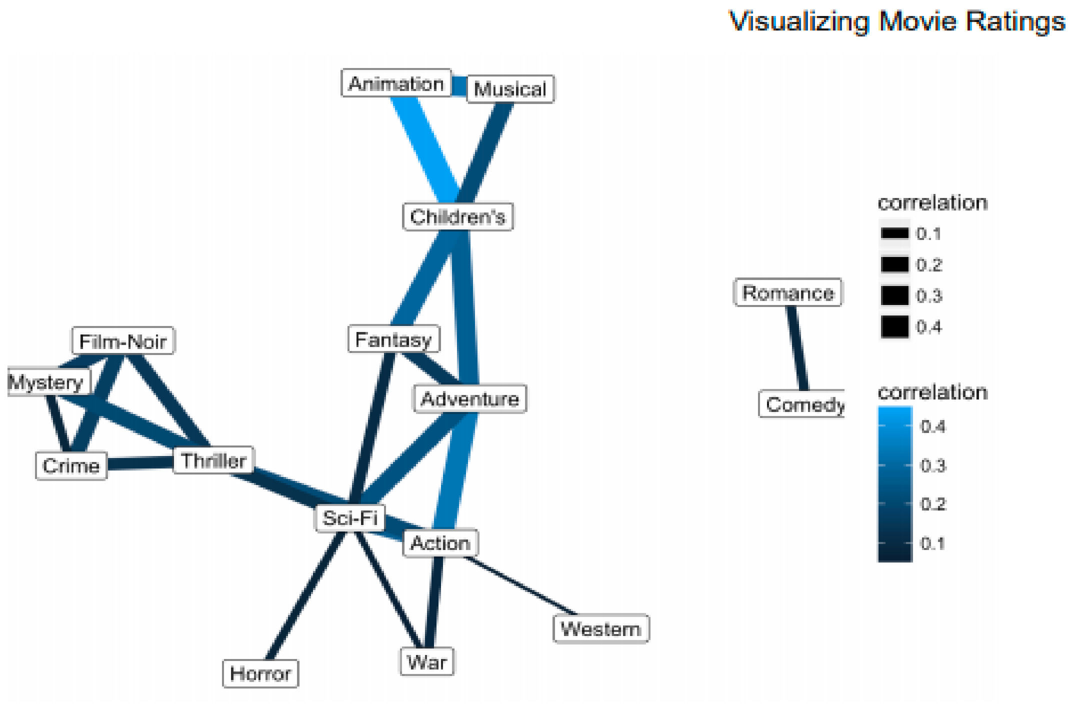


Figure 9: Association Rule

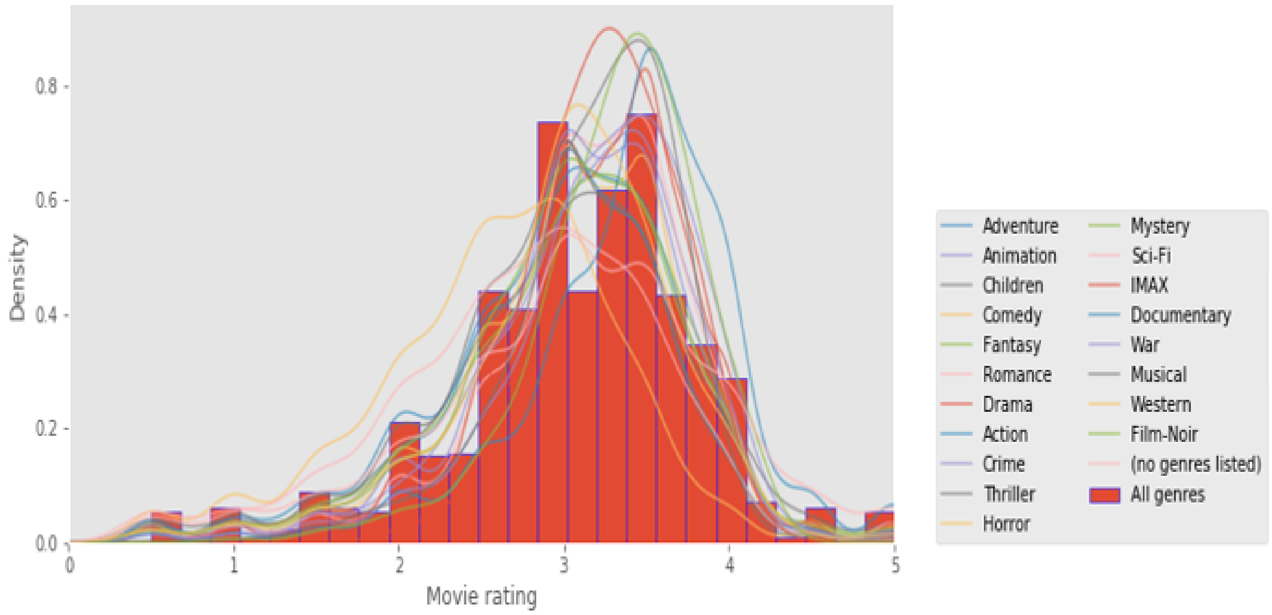


Figure 6: Normality of Genre's Rating

3 Algorithm

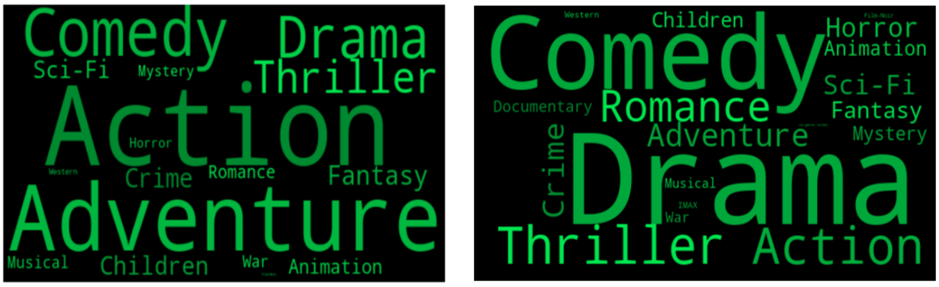
This report uses both recall and re-rank to make recommendation. Recall step uses both UB-CF and IB-CF methods to make better selection from large amount of movies. Re-rank step contains Wilson interval and regular term.

3.1 Recall

As research experience suggests, UB-CF is more widely used in recalling socialized goods and IB-CF is more widely used in recalling individualized goods. Thus, to make our recommendation both socialized and individualized, we decide to recall 10 movies using both method.

3.1.1 UB-CF

The following recommendation are all made for user 1. Figure 10 shows what user 1 had watched and all genres in the sample.



(user 1)

(all users)

Figure 10: Movie Genres Watched

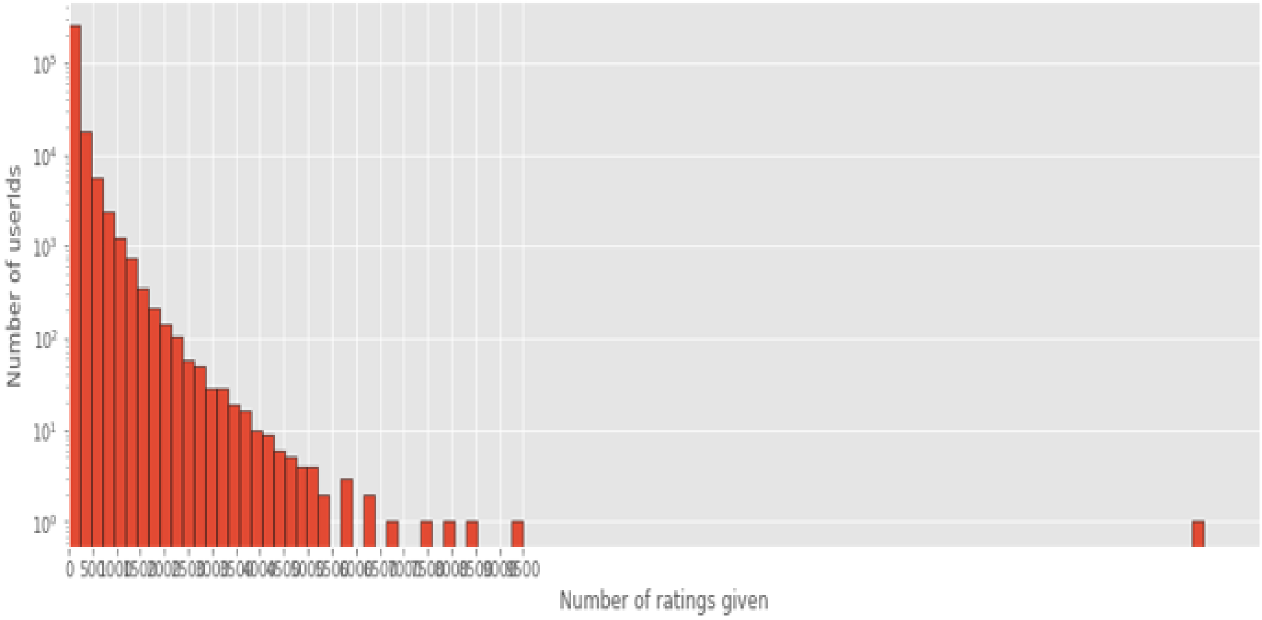
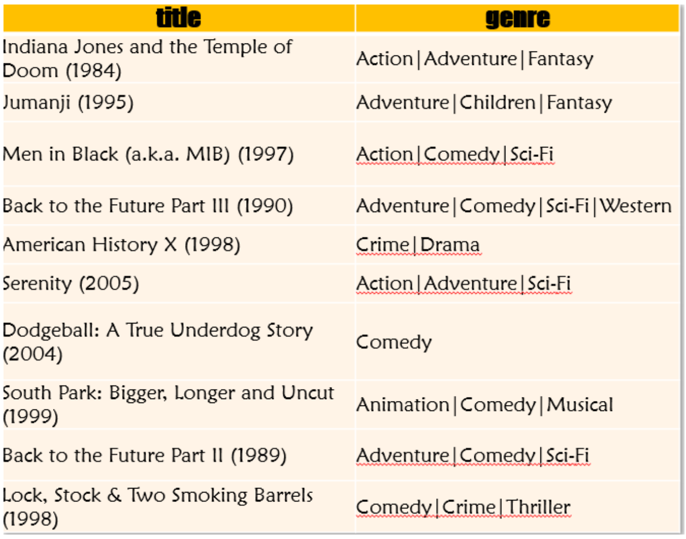


Figure 7: Number of Rating

The similarity between movies is measured by cosine similarity, and movies recalled based on UB-CF. 10 selected movies and its genres are shown in figure 11.



title	genre
Indiana Jones and the Temple of Doom (1984)	Action|Adventure|Fantasy
Jumanji (1995)	Adventure|Children|Fantasy
Men in Black (a.k.a. MIB) (1997)	Action|Comedy|Sci-Fi
Back to the Future Part III (1990)	Adventure|Comedy|Sci-Fi|Western
American History X (1998)	Crime|Drama
Serenity (2005)	Action|Adventure|Sci-Fi
Dodgeball: A True Underdog Story (2004)	Comedy
South Park: Bigger, Longer and Uncut (1999)	Animation|Comedy|Musical
Back to the Future Part II (1989)	Adventure|Comedy|Sci-Fi
Lock, Stock & Two Smoking Barrels (1998)	Comedy|Crime|Thriller

Figure 11: Movies Selected Based on UB-CF

Comparing it with figure 10, we can see that recommendation method based on UB-CF consider both user 1's preference, such as Adventure and Comedy, and public aesthetics, such as Comedy and Action.

3.1.2 IB-CF

IB-CF recommend similar users' high-rating movies to users, and the similarity between users are also measured by cosine similarity. The recommendation results and their genres are shown in figure 12:

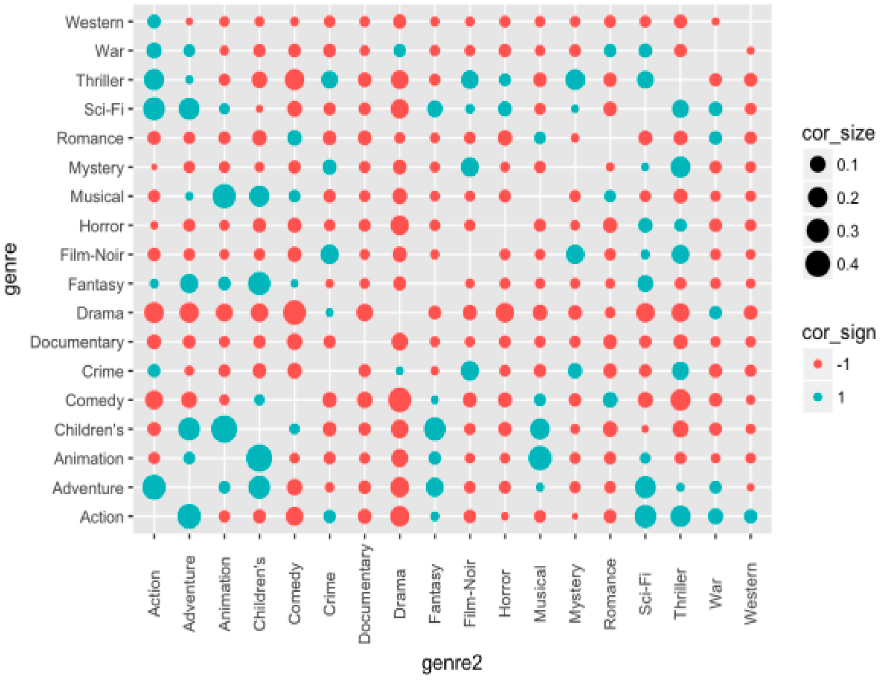
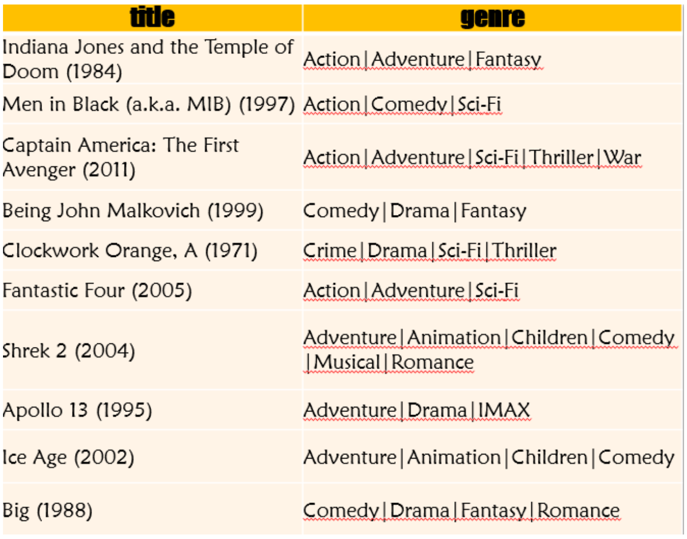


Figure 8: Correlation Matrix between Genres



title	genre
Indiana Jones and the Temple of Doom (1984)	Action|Adventure|Fantasy
Men in Black (a.k.a. MIB) (1997)	Action|Comedy|Sci-Fi
Captain America: The First Avenger (2011)	Action|Adventure|Sci-Fi|Thriller|War
Being John Malkovich (1999)	Comedy|Drama|Fantasy
Clockwork Orange, A (1971)	Crime|Drama|Sci-Fi|Thriller
Fantastic Four (2005)	Action|Adventure|Sci-Fi
Shrek 2 (2004)	Adventure|Animation|Children|Comedy|Musical|Romance
Apollo 13 (1995)	Adventure|Drama|IMAX
Ice Age (2002)	Adventure|Animation|Children|Comedy
Big (1988)	Comedy|Drama|Fantasy|Romance

Figure 12: Movies Selected Based on IB-CF

It's clearly that recommendation based on IB-CF is more individualized than UB-CF, since movies with genre Action and Adventure, which is user 1 admired, are frequently recommended here.

3.1.3 NLP

NLP is an unsupervised document quantification method, based on natural network. NLP method recommend movies that have similar comments with users' existing movies comments, and the similarity between users are also measured by cosine similarity.

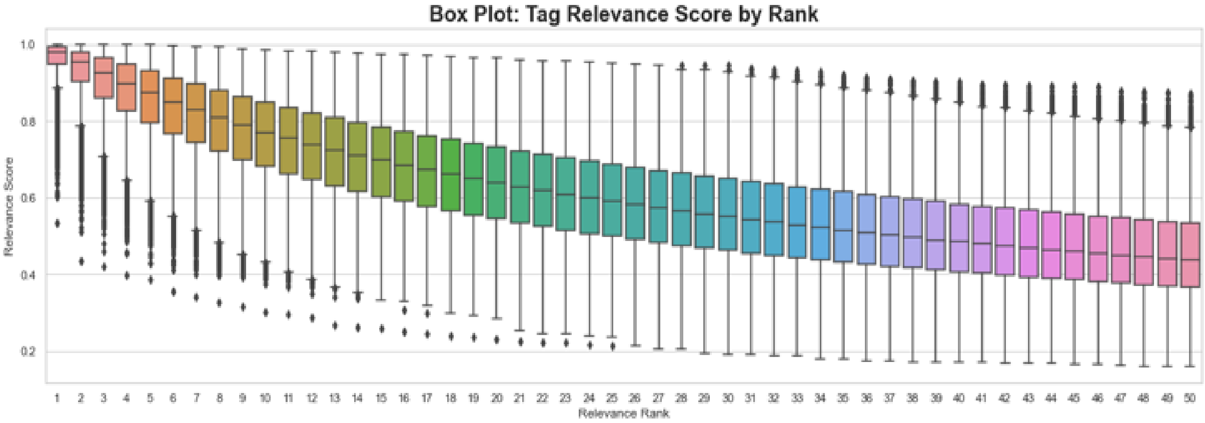
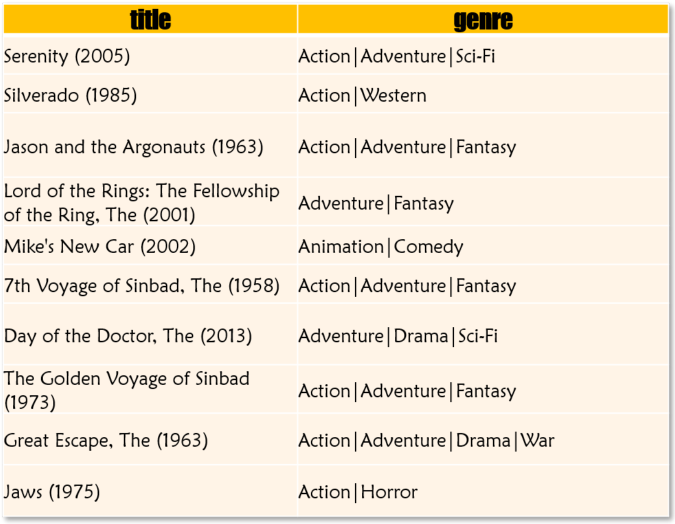


Figure 13: Correlation between Movies' Comments

Figure 13 shows the correlation between movies' comments, and it's clear that the most relevant 50 movies have pretty high correlation. The recommendation results and their genres are shown in figure 14:



title	genre
Serenity (2005)	Action|Adventure|Sci-Fi
Silverado (1985)	Action|Western
Jason and the Argonauts (1963)	Action|Adventure|Fantasy
Lord of the Rings: The Fellowship of the Ring, The (2001)	Adventure|Fantasy
Mike's New Car (2002)	Animation|Comedy
7th Voyage of Sinbad, The (1958)	Action|Adventure|Fantasy
Day of the Doctor, The (2013)	Adventure|Drama|Sci-Fi
The Golden Voyage of Sinbad (1973)	Action|Adventure|Fantasy
Great Escape, The (1963)	Action|Adventure|Drama|War
Jaws (1975)	Action|Horror

Figure 14: Movies Selected Based on NLP

3.2 Re-rank

The re-rank process involves two attributes: rating of each user as the identification of the overall goodness of the movie, and the genre to measure similarity, which is used as a penalty for the movies the user might feel used to, even feel bored with. In this process, we use the result of the multi-channel recall as the input, and we will finally return a subset series of the input, as a final result to recommand to the users.

3.2.1 Wilson Interval

The rating of a movie make up with rating and rating volume. So here comes a question: which movie should be recommended if one movie has high rating but low rating quantity and another movie has low rating but high rating quantity? When making recommendation, we actually face a trade-off between rating and rating volume. Luckily, we can use Wilson interval, which is a widely used arithmetic, to balance these two variables, and the definition of Wilson interval is given by:

$$(\frac{p+\frac{1}{2n}z^2-z\sqrt{\frac{p(1-p)}{n}+\frac{z^2}{4n^2}}}{1+\frac{1}{n}z^2},\frac{p+\frac{1}{2n}z^2+z\sqrt{\frac{p(1-p)}{n}+\frac{z^2}{4n^2}}}{1+\frac{1}{n}z^2})$$

where p stands for proportion of good rating, z is the quantile of normal distribution and n stands for quantity of rating. Lower boundary means popular movies and upper boundary means niche excellent movies. To derive p, we need to transform rating, which ranges from 0 to 5, to good and bad. The empirical distribution of rating is shown in figure 15.

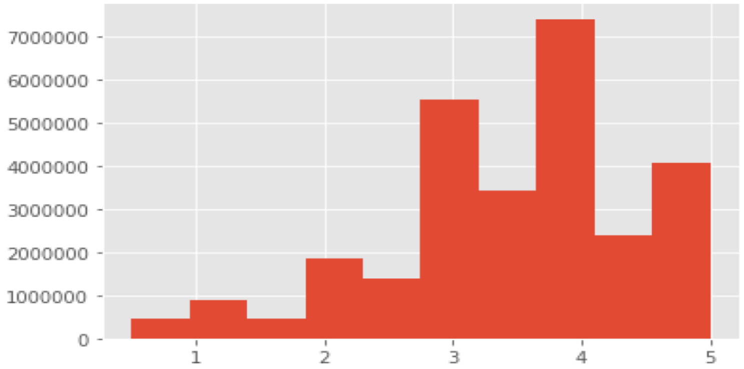


Figure 15: Histogram of Rating

We find that the distribution looks like a bimodal distribution instead of normal distribution. Here We assume both of the peaks are normally distributed. To test whether the distribution is as we assumed, we can use the em-test to conduct the hypothesis test that:

H_0 : The ratings are normally distributed

H_1 : The distribution is a combination of two normal distributions.

Here we use the expectation maximization test (EM-test), which is based on the MLE of the observations $\boldsymbol{X}=\{X_1,\cdots,X_n\}$, the hidden variable $\boldsymbol{Z}$, and the parameter $\boldsymbol{\theta}$. The likelihood of the model can be written as:

$$L(\boldsymbol{X}|\boldsymbol{\theta})=\prod_{c=1}^{k}p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}),\ \boldsymbol{Z}=\{Z_1,\cdots,Z_k\}$$

where X is discrete in our case. After taking log like solving for any single distribution, we have:

$$logL(\boldsymbol{X}|\boldsymbol{\theta})=\sum_{i=1}^{N}log\left[\ \sum_{c=1}^{k}p(X_i,Z_c|\boldsymbol{\theta})\ \right]$$

According to Jensen's inequality, we have:

$$logL(\boldsymbol{X}|\boldsymbol{\theta})=\sum_{i=1}^{N}log\left[\ \sum_{c=1}^{k}p(X_i,Z_c|\boldsymbol{\theta})\ \right]\geq\sum_{i=1}^{N}\sum_{c=1}^{k}\left[\ q(Z_c)logp\frac{X_i,Z_c|\boldsymbol{\theta}}{q(Z_c)}\ \right]=L(\boldsymbol{\theta},q)$$

When the RHS is globally maximized, the LHS is locally maximized, so the algorithm is to find out:

$$\hat{\boldsymbol{\theta}}=\underset{\boldsymbol{\theta}}{argmax}L(\boldsymbol{\theta},q)$$

Considering the inquality listed above, we have such expansion:

$$[logL(\boldsymbol{X}|\boldsymbol{\theta}) - L(\boldsymbol{\theta}, q)$$

$$= \sum_{i=1}^{N} log \left[\sum_{c=1}^{k} p(X_i, Z_c|\boldsymbol{\theta}) \right] - \sum_{i=1}^{N} \sum_{c=1}^{k} \left[q(Z_c) log \frac{p(X_i, Z_c|\boldsymbol{\theta})}{q(Z_c)} \right]$$

$$= \sum_{i=1}^{N} \left[logp(X_i|\boldsymbol{\theta}) \sum_{c=1}^{k} q(Z_c) - \sum_{c=1}^{k} q(Z_c) log \frac{p(X_i, Z_c|\boldsymbol{\theta})}{q(Z_c)} \right]$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{k} q(Z_c) \left[logp(X_i|\boldsymbol{\theta}) - log \frac{p(X_i, Z_c|\boldsymbol{\theta})}{q(Z_c)} \right]$$

$$= \sum_{i=1}^{N} \sum_{c=1}^{k} q(Z_c) \left[log \frac{p(X_i, Z_c|\boldsymbol{\theta}) q(Z_c)}{p(X, Z_c|\boldsymbol{\theta})} \right]$$

According to Bayes' theorem, we have:

$$[logL(\boldsymbol{X}|\boldsymbol{\theta}) - L(\boldsymbol{\theta}, q) = \sum_{i=1}^{N} \sum_{c=1}^{k} q(Z_c) \left[log \frac{q(Z_c)}{p(Z_c|\boldsymbol{\theta}, X_i)} \right] = \sum_{i=1}^{N} KL[q(Z)||p(Z|X_i, \boldsymbol{\theta})]$$

$$\Rightarrow L(\boldsymbol{\theta}, q) = logp(X|\boldsymbol{\theta}) - \sum_{i=1}^{N} KL[q(Z)||p(Z|X_i, \boldsymbol{\theta})] = F(\boldsymbol{\theta}, q)] \quad \text{F is the Gibbs free energy}$$

$$\underset{\boldsymbol{\theta}}{max} L \Rightarrow \frac{\partial}{\partial \boldsymbol{\theta}} [L(\boldsymbol{\theta}, q)] = 0 \Rightarrow \frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{i=1}^{N} \sum_{c=1}^{k} q(Z_c) logp(X_i, Z_c|\boldsymbol{\theta}) \right] = 0 \Rightarrow \frac{\partial}{\partial \boldsymbol{\theta}} E \left[logp(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right] = 0$$

The em-test can solve the problem whether the θ is a scalar or a vector combined of several values. After the em-test, we find that the p-value of H_0 is 0, thus we can confidently reject H_0. According to the em-test, the two distributions are:

$$N_1(4.326, 1.96)$$

$$N_2(3.203, 4.01)$$

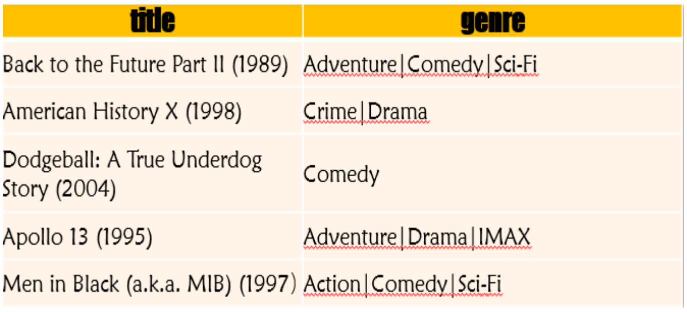
And we should find a criteria that minimizes the probability of classifying a bad movie into good ones while control controlling the probability of classifying a good movie into bad ones. Using Bayes rule for classification, we know that when two pdfs are equal, the total probability is minimized, so we solve the equation:

$$\frac{1}{1.4} e^{\frac{(x-4.326)^2}{2 \times 1.96}} - \frac{1}{2.01} e^{\frac{(x-3.203)^2}{2 \times 4.01}} = ln \frac{4.01}{1.96}$$

The solution is 3.13, which means when rating >3.13, the movie is thought to be a good movie. So we can calculate p in the Wilson interval.

3.2.2 Regular Term

Regular term related with genre is also considered, using cosine similarity between movies had been watched and recommended movies, our algorithm award movies similar to what user had watched and penalize movies different from what user had watched. The final recommended 5 movies are:



title	genre
Back to the Future Part II (1989)	Adventure|Comedy|Sci-Fi
American History X (1998)	Crime|Drama
Dodgeball: A True Underdog Story (2004)	Comedy
Apollo 13 (1995)	Adventure|Drama|IMAX
Men in Black (a.k.a. MIB) (1997)	Action|Comedy|Sci-Fi

Figure 16: Final Recommended Movies

4 Algorithm Evaluation

Our group think our algorithm has following advantages and disadvantages.

4.1 Advantages

- Multi-recall is used, which guarantee the diversity of recommended movies.
- Re-rank using statistical methods instead of machine learning algorithm, which makes the recommend results easy to interpret and make individualized recommendation.
- Our model is built by basic data mining algorithm, so it easy to interpret and time complexity is not high.

4.2 Disadvantages

- When there comes a totally new user, we can't make recommendation, since there is no information at all.
- The model can hardly renew when there is new movies included.
- Recommendation results are not as good as deep neural network.

5 References

[1]Bellogin, et al. (2013). A comparative study of heterogeneous item recommendations in social systems. Information Sciences, 221, 142–169.

[2]Bobadilla, et al. (2012). A collaborative filtering similarity measure based on singularities. Information Processing and Amp; Management, 48(2), 204–217.

[3]Da u, et al. (2019). Recommendation system based on deep learning methods: a systematic review and new directions. Artificial Intelligence Review, 53(4), 2709–2748.

[4]Devi, et al. (2018). A Hybrid Approach for Movie Recommendation System Using Feature Engineering. 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT).

[5]Gold, et al (2001). Knowledge Management: An Organizational Capabilities Perspective. Journal of Management Information Systems, 18(1), 185–214.

[6]Hasija, H. (2017). An effective approach of feature selection for recommender systems using fuzzy C means clustering along with ant colony optimization and neural networks. 2017

8th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

[7]Levandoski, et al. (2012). LARS: A Location-Aware Recommender System. 2012 IEEE 28th International Conference on Data Engineering.

[8]Nie, et al. (2013). Movie Recommendation Using Unrated Data. 2013 12th International Conference on Machine Learning and Applications.

[9]Pujahari, et al. (2014). An Approach to Content Based Recommender Systems Using Decision List Based Classification with k-DNF Rule Set. 2014 International Conference on Information Technology.

[10]Rossetti, et al. (2013). Towards Explaining Latent Factors with Topic Models in Collaborative Recommender Systems. 2013 24th International Workshop on Database and Expert Systems Applications.

[11]Simas, et al. (2012). Semi-metric Networks for Recommender Systems. 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology.

[12]Singh, et al. (2015). Relative similarity based approach for improving aggregate recommendation diversity. 2015 Annual IEEE India Conference (INDICON).

[13]Yildirim, et al. (2008). A random walk method for alleviating the sparsity problem in collaborative filtering. Proceedings of the 2008 ACM Conference on Recommender Systems - RecSys '08.