I. PROJECT DEFINITION

Project Overview:

This project builds a mock-up Movie Recommendation System based on MBTI type.

Problem statement:

Due to the high performance of computers today and big data era, many companies such as Netflix, Amazon are making use of these two advantages to improve their webpage effectiveness in product recommendation system by not only reducing time for consumers but also encouraging them to purchase more as well as content with their purchases when their likeness and best items at the top of the screen. In addition, in building a recommendation system, the collaborative filtering approach gains lots of reputations these days since it makes use of existing user data and as many people have the same preferences. Hence, in this project, I propose a recommendation structure for movies using collaborative filtering.

However, there are two major challenges affecting collaborative filtering performances which are data sparsity and system scalability. Data sparsity occurs when the rating matrix is sparse, especially if there are new users or new items enter in the system as well as a new recommender system. As an user has not rated or purchased any items yet, or an item has not been rated, which is called the cold-start problem. As a results, finding a group of similar users or items is a big challenge causing by historical data sparsity.

The scalability issue happens when the number of users or items is increasingly growing. This problem is inevitable as more and more people today are buying products both on the streets and e-commerce platforms. In fact, it is extremely large in real world, which is so-called Big Data. For better recommendation, CF paradigm has to estimate similarities of every users and items. Subsequently, it suffers from serious scalability issue.

Assigning users in the specific groups is a method to support solving scalability, in which the system just needs to estimate similarity within a smaller group rather than a whole existing data. In addition, if a new user has some traits as same as users in a specific group, the systems only needs to suggest items which are beloved by users within that group as well as reflects user's personality

All in all, in this project, I propose to use Myer-Briggs Type Indicator (MBTI) in movie recommendation system to individualize recommendations by user-user collaborative filtering. In personality typology, MBTI is an introspective self-report questionnaire indicating differing psychological preferences in how people perceive the world and make decisions. The test attempts to assign 4 categories: Introverted vs Extroverted, Sensing vs INtuition, Thinking vs Feeling, Judging vs Perceiving. One letter from each category is taken to produce a for letter list such as INFP. There are 16 types in total. I evaluated this approach by doing two experiments. Firstly, predicting existing users' ratings for movies which they haven't seen yet. Secondly, predicting and ranking all ratings of existing items for new users.

Metrics

In this project, I use MAE score and RMSE score to evaluate the recommendation.

II. ANALYSIS

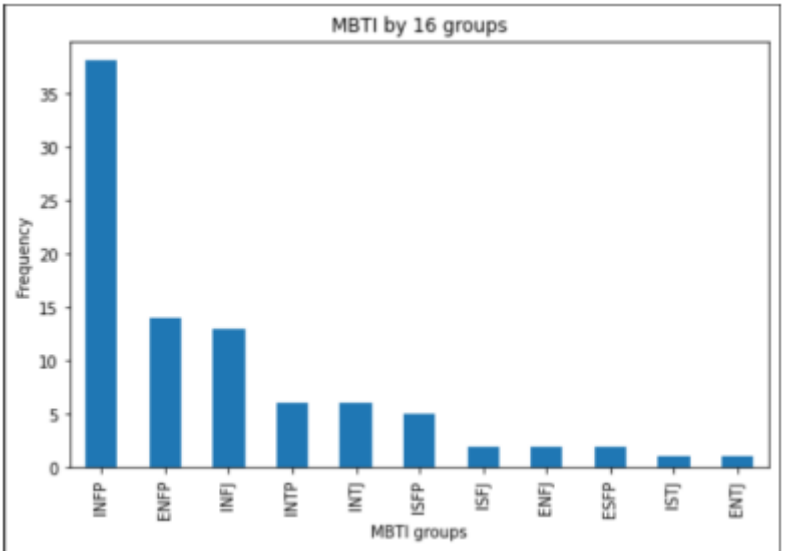
Data Description

1. Collecting user data

There is no specific data dedicated to my experiment. Therefore, I conducted a survey through Facebook in which I asked people for answering: 8 multiple choice questions about their western zodiacs, one of their favorite colors, numbers, seasons, music genres, musical instruments, social media, alignment[1]. Secondly, I proposed a list of 20 Vietnamese and Hollywood movies for rating. Before posting the survey in Facebook, there are some adjustments for the sake of the clear meanings of questions and answers. The number of participants in this experiment is 91, after removing nan values, the final number of users is 90.

2. User MBTI

As MBTI has not been known by my friends in Facebook, so I using 8 multiple choice questions implying that each user's answers would reflecting their personality. Data for all of the answers for each question is collected from Personality Database [2], where MBTI is contributed by many people from all over the world. Each choice corresponds to a list of percentages of E, I, S, N, T, F, J ,P (E + I = S + N = T + F = J + P = 100). After answering 8 questions, for each user ,I averaged out 8 answers to get a one list of percentages and compared to get the higher percentage between the two features within each of 4 categories: E vs I, S vs N, T vs F and J vs P. And the final MBTI contains 4 letter such as INFP. The following picture is the distribution of user MBTI in my data.



Picture 1. Participants MBTI distribution

3. Rating matrix

Likert scale is ranging from 0 to 10, 0 – indifference, 1- deadly awful ~ 10- fantastic. The survey data contains 1800 ratings. As this project only took rating different from 0 into predicting rating. Thus, the final rating matrix decreased to 1453 ratings.

User_ID	Em chưa 18 - 2017	Người nhện,	Tháng năm rực rỡ	Aquaman 2018	Cua lại vợ bầu -	Trạng Quỳnh
1	3	8	2	8	6	1
2	10	4	7	2	10	10
3	7	5	8	6	8	9
4	8	7	5	5	7	5
5	8	9	9	8	8	8
6	8	8	8	9	9	7
7	6	0	1	0	1	0
8	1	10	0	0	4	5
9	8	9	9	9	8	8
10	9	9	9	9	10	10
11	3	8	3	7	7	3
12	8	10	0	9	0	0
13	0	0	0	0	0	0
14	3	7	9	8	2	1
15	0	8	0	8	0	0

Picture 2. Movie ratings

4. Movie dataset

Movie dataset contains 20 movies from different genres: Romantic, Romantic Comedy, Action, Criminal, Horror, Sci-fi. The majority is Vietnamese movies as our movie industry has gradually developing in recent years as well as many Vietnamese people or foreigners are going to love our country movies. In addition, I also added some worldwide famous movies such as Avengers: Endgame to measure if there is not any influence of personality in famous movies.

Movie_ID	Title	Title_Eng
1	Em chưa 18 - 2017	Jailbait(2017)
2	Người nhện, trở về nhà 2017	Spider-man Homecoming
3	Tháng năm rực rỡ 2018	Go-Go Sisters 2018
4	Aquaman 2018	Aquaman
5	Cua lại vợ bầu - 2019	Win My Baby Back 2019
6	Trạng Quỳnh 2019	Mandarin Quynh
7	Mắt biếc 2019	Blue Eyes
8	Ròm 2019	Rom
9	Hai phượng 2019	Furie
10	Lật mặt: nhà có khách 2019	Face Off: The Walking Guests
11	Ký sinh trùng 2019	Parasite 2019
12	Gã hề ma quái 2 2019	IT Chapter 2
13	Annabelle: ác quỷ trở về 2019	Annabelle Comes Home 2019
14	Chúa tể Godzilla: Đế vương bất tử 2019	Godzilla: King of the Monsters (2019)
15	Avengers: Hồi kết 2019	Avengers: Endgame
16	Nắng 3: Lời hứa của cha 2020	Nang 3: Loi Hua Cua Cha
17	Chị Mười Ba: 3 ngày sinh tử 2020	13rd Sister: Three Deadly Days 2020
18	Bố già 2021	Dad, I'm sorry 2021 (Old Father)
19	Lật mặt: 48H 2021	Face Off: 48 Hours 2021
20	Trò chơi con mực 2021	Squid Game

Picture 3. Movies in mocked archive.

III. METHODOLOGY

User-collaborative filtering recommendation system

The basic idea of collaborative filtering method is that unspecified ratings can be estimated as the observed ratings are often highly correlated across various users and items. For example, consider two users named Alice and Bob, who have very similar tastes. If the ratings, which both have specified, are very similar, then their similarity can be identified by the underlying algorithm. In such cases, it is very likely that the ratings in which only one of them has specified a value, are also likely to be similar to the other. This similarity can be used to make inferences about unspecified values. User-based collaborative filtering is a memory-based method of collaborative filtering. In this case, the ratings provided by like-minded users of a target user A are used in order to make the recommendations for A. Hence, the basic idea is to determine users, who are similar to the target user A, and estimate ratings for the unobserved ratings of A by computing weighted averages of the ratings of this peer group. In this project, instead of the conventional approach in building a recommendation system by using only ratings matrix, I assigned users in the specific MBTI group in which my system only needs to estimate ratings for a user in a specific group. This method is called formally neighborhood-based collaborative filtering. In this report, I 5 only used weighted average rating prediction that is conventionally used in rating prediction by using observed rating, similarity score.

Experiment

In this project, I created 4 types of data from the original data for various evaluation on the pros and cons of methodology as well as my data. Firstly, predicting ratings for historical watching users, denoted as (*). I took out 10% ~ 145 ratings from rating matrix for test dataset. In this experiment, I tested in 3 kinds of data:

- Using only rating matrix which is called as Ratings only.
- 8 answers from users, assigning them in MBTI group and naming it as Info only. Data is described clearly in the following paragraph.
- Aggregating two types of prior data as name Info+ Ratings.

Secondly, for cold-start resolving method, denoted as (**), I picked 9 out of 90 users and treated them as new users. The information I took from them were 8 answers they gave and predicting their ratings by finding similarity within an MBTI group which I found out by 8 answers they gave. And encoding 8 answers into one-hot encoder for estimating similarities to targeted user.

I used cosine similarity for estimating similarity score and chose top 10 users have closest correlation with targeted user. Using weighted average rating prediction for predicting user ratings with raw ratings.

IV. RESULTS & CONCLUSION

Results

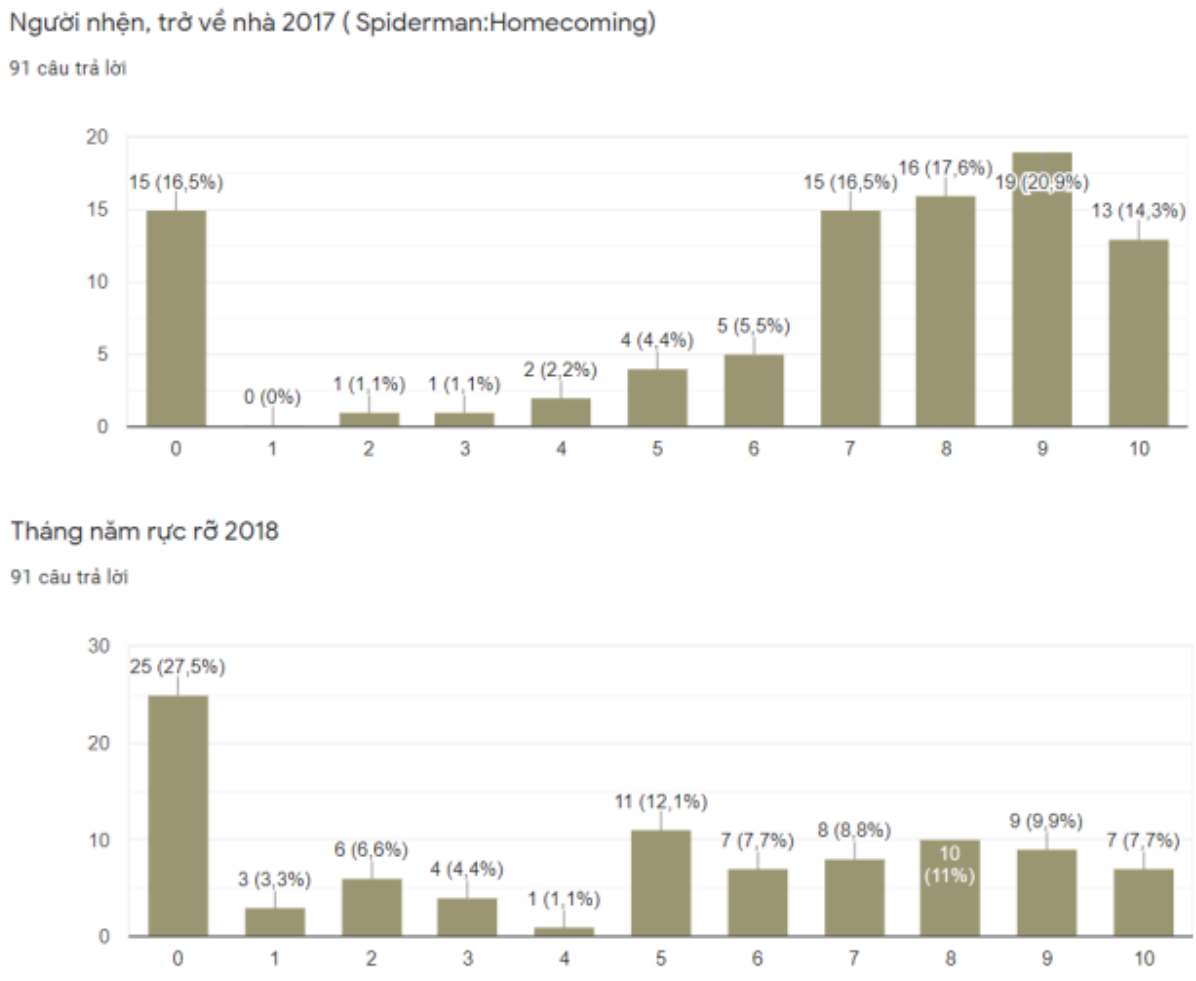
As we can see from the table below, for historical watching users, there is an outstanding result from Info+ Rating input, the reason is having more data the models would predict better, this input together with Info Only result is far more better than Rating Only. However, the most significant result is Cold-start.

For more insights, when looking into predicting rating distribution, most of 4 types of data approach are good at predicting ratings at around 4 ~ 8. This is because when using weighted average ratings prediction

with raw ratings, it ignores biases while in fact, they usually occur in rating movies. In addition, this method are prone to outliers such as low true ratings, scores 1~ 2 are predicted at high ratings at around 6~ 8 and sometimes true rating scores are high but models predict as around 5 ~ 7. The imbalance distribution between MBTI groups and small dataset in this experiment also contribute to difficulties in predicting ratings within the small scale MBTI groups and sometimes models predict 'nan' values as lacking of data, particularly group ISFJ, ENFJ, ESFP, ISTJ, ENTJ. In addition, this is especially a case with famous movies such as Dad, I'm sorry, Avengers: Endgame or Spiderman have predicting ratings quite high at around 6 7~ 8. This is because these movies are famous and people are high rating them. While movies such as Go-go Sisters has average and low predicted rating as not many people have watched this movie already as well as the scores are scattered from 1 to 10.

Inputs	(*) Ratings only	(*) Info+ Ratings	(*) Info only	(**) Cold-start
MAE	2.40	2.14	2.17	**1.90**
RMSE	2.85	2.68	2.77	**2.39**

Table 1. MAE and RMSE of predicted ratings.



Picture 4. Spiderman (top) and Go-go Sisters (bellow) ratings distribution

Obstacles:

Lacking of data, imbalance distribution between MBTI groups cause difficulty in finding historical ratings within a small groups MBTI such as ISFJ. Sometimes models even predicted ‘nan’ values.