# 535 Semester Long Project

Xianzhi Luo B00814692

## 1 Brief Solution

In this final project, we are requested to build a recommender system. Generally, there are two types of recommender systems: Context filtering based recommender system and collaborative-filtering based recommender system. Because our dataset is a dataset include user id, item id and rating value, I choose memory based collaborative filtering to build this recommender system.

The first thing to do is data preprocess. First step is split the training and test data set and do the summarization. I use the Pandas library to get the mean, total number of users' rating, total value of users' rating. Then I calculate the items that have been rated and fill them with default value. This step is to solve cold start issue.

To predict the rating R on item I given by user U, there are three main tasks to do. First, find all similar users that have rated on item I by cosine similarity (or other similarity algorithm). Second, I choose top 100 of these similar users and use the sum of (weights times mean rating) to get the predict rating Rp. Third, build and train the dataset by KNNwithmean and calculate the RMSE. After training and model selection, I get the optimal model for this dataset and use it to predict remain 0 values in the dataset. I also compare KNNwithmean with other algorithms or models such as Co-Clustering, SVD and so on.

## 2 Cold Start Issue

In collaborative filtering, there is a big issue which is called cold start problem. And in cold start, there are user cold start problem and item cold start problem which means a new user or new item comes in, there is no relation you can use to predict this newcomer. In this dataset, among 943 users, the user with least rating number (the number of items rated by one user) is 3 while the mean of rating number is 63. Among 1628 items, there are 65 items have not been rated. What I do is using representative based Matrix Factorization to solve this cold start. The RBMF means an additional constraint that m items should be represented by a linear combination of k items. So, I choose 100 most popular users by their rating number, with user 517 who has rated 517 items and so on. For those 65 items have not been rated, I randomly choose rating number by quantile 0.25 which is 18, mean which is 37, quantile 0.75 which is 84 as their rating number. And I choose this number from 100 most popular users, and the rating on the 65 items will be the mean of these most popular users.

## 3 Two Real-world Examples

Document-term matrix: A document-term matrix or term-document matrix is mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a document-term matrix, rows correspond to documents in the collection and columns correspond to terms. Each entry corresponds to the number of times the associated term appears in the indicated document.

Geographic Triangulation: Suppose we are given partial information about the distances between objects and would like to reconstruct the low-dimensional geometry describing their locations. For example, we may have a network of low-power wirelessly networked sensors scattered randomly across a region. We can only get a partially observed distance matrix, the row can be the place of departure and the column can be the destination. Each entry corresponds to the distance between departure and destination.

## 4 Mini Survey

*Abstract*— Today's Recommender system is a relatively new but important area of research in data mining. There are 4 main types of recommender system– content based, Collaborative, Demographic and hybrid filtering.[1] All of them are delicate to recommend new items to users or items to new users based on the relationship between users and items discovered from history data. This paper is a mini survey to introduce these types of recommender systems.

Keywords-; Recommender system, content-based filtering, collaborative filtering, hybrid filtering, data mining.

### I. INTRODUCTION

Recommender system [6] plays an important role in e-commerce[7] like Amazon, eBay and some movie websites like Netflix, Hulu and so on. Such websites always have huge amounts of users and when they want to recommend some new products or movie to their users based on users' interest, they need to build recommender systems. This mini survey will briefly introduce several main types of recommender system and the comparison of them.

### II. RECOMMENDER SYSTEM

#### 1 Content based system

Content-based filtering [2] methods are based on a description of the item and a profile of the user's preferences. These types of profiles and description are created at the beginning, when the user creates the account and starts using the system. Item's description is the features content-based system used to classify user like or dislike this item. In this system, keywords are used to describe the items and a user profile is built to indicate the type of item this user likes. In other words, these algorithms try to recommend items that are similar to those that a user liked in the past or is examining in the present. It does not rely on a user sign-in mechanism to generate this often temporary profile. In particular, various candidate items are compared with items previously rated by the user and the best-matching items are recommended.

#### 2 Collaborative system

Instead of using description of item and user file, collaborative system [4][5] based on collaborative filtering and using the rating, description and preferences given by users to find the correlation between users and items. Given the history rating data given by users and some similarity algorithms, we can get a list of similar users or items of one specific user or item. Then we can predict the unknown rating of this user or item by the rating of similar users or items. Collaborative filtering (CF) methods produce user-specific recommendations of items based on patterns of ratings or usages like purchases without the demand for data about either items or users.

3 Demographic system

Just as what it is called, demographic system uses demographic information such as gender, age, job, living area and so on to find similar users. The difference between demographic and content-based method is demographic system only need information on users and it does not care the description of items. Demographic usually divides users into several groups based on their gender, age, job and living area and predict the rating for each group by the history data in that group.[3]

4 Hybrid system

Hybrid system combines content-based system and collaborative system to get the best advantage and gaining better result and reduce the issues and challenges of these applications. Hybrid system includes several methods: Weighted, Switching, Mixed, Feature combination, Feature augmentation, Cascade, Meta-level. Hybrid system mixes these methods to alleviate the issues and problems in system either content-based or collaborative used only.

## III. Comparison

1 Content-based RS

Advantages: Doesn’t need users’ data, predicts based on items similarity
Disadvantages: Need description to detect item features, doesn’t depend on users’ rate

2 Collaborative RS

Advantages: The system doesn’t use demographic information to recommend items, matches similar items between users, able to recommend to the user items outside their preferences and may like this item.
Disadvantages: Suffer from cold start problem, popular user and highest rating decide the quality

3 Demographic RS

Advantages: Doesn’t based on user-item ratings, gives recommendation before user rated any item.   Disadvantages: Gathering of demographic data leads to privacy issues, Stability vs. plasticity problem

4 Hybrid Approaches

Advantages: Combine all advantages between content based and collaborative filtering, based on items’ description and user's evaluation, solve over specialization, increase customer satisfaction rate. Disadvantages: Suffer from the cold start problem, Early Rater problem for products, sparsity problem.

## Reference

[1] Ponnam, Lakshmi Tharun, et al. "Movie recommender system using item-based collaborative filtering technique." Emerging Trends in Engineering, Technology and Science (ICETETS), International Conference on. IEEE, 2016.

[2] Goswami T., Vaisshnavi Y. (2020) A Case Study on Correctness Evaluation of Content Based Recommender System Based on Text, Semantic Text and Visual Similarity. In: Kumar

A., Paprzycki M., Gunjan V. (eds) ICDSMLA 2019. Lecture Notes in Electrical Engineering, vol 601. Springer, Singapore.

[3] M. H. Mohamed, M. H. Khafagy and M. H. Ibrahim, "Recommender Systems Challenges and Solutions Survey," *2019 International Conference on Innovative Trends in Computer Engineering (ITCE)*, Aswan, Egypt, 2019, pp. 149-155, doi: 10.1109/ITCE.2019.8646645.

[4] J. Breese, D.. Heckerman, and C. Kadie, Empirical analysis of predictive algorithms for collaborative filtering, *Proc. Conf. Uncertainty in Artificial Intelligence*, (UAI98) 1998

[5] J.L. Herlocker, J.A. Konstan, J.R.A. Borchers, and J. Riedl, An algorithmic framework for performing collaborative filtering, Proc. International on ACM SIGIR Research and Development in Information Retrieval, (SIGIR98) 1998

[6] P. Resnick and H.R. Varian, Recommender Systems, Special Issue of Communications of the ACM, 40(3), 1997

[7] Kulkarni P.V., Rai S., Kale R. (2020) Recommender System in eLearning: A Survey. In: Bhalla S., Kwan P., Bedekar M., Phalnikar R., Sirsikar S. (eds) Proceeding of International Conference on Computational Science and Applications. Algorithms for Intelligent Systems. Springer, Singapore.

## 5 Open Issues

### Grey sheep

Grey sheep occurs in collaborative filtering and I also met this problem in project that is a user’s preference do not match any group, which means we cannot find any similar user of this specific user. Thus, the similarity and correlation cannot be calculated on that specific user and we cannot predict rating on it.

One suggested solution is to use one-class classification to generate a prediction list for these users, where decision boundaries are learned that distinguish between normal and grey-sheep users.

### Scalability

Scalability measure the ability of the system to work effectively with high performance while growing in the information. Recommender system needs to recommend items to the users without no change while the number of users increased or the number of items increased too.

Solution: More computations and cost more, also a large database to store huge amount data of rating .