Data Mining

Recommenders – Collaborative Filtering

Examples from "Data Science Concepts and Practice" 2018 MK : Vijay Kotu and Bala Deshpande

Terri Hoare – November 2023

Recommendation Engines
Collaborative Filtering

Collaborative Filtering is based on a simple idea—a user will prefer an item if it is recommended by their like-minded friends. Suppose there is a new restaurant in town. If a friend, who happens to have the same interests as the user, raves about it—the user might like it as well. The idea makes intuitive sense. Collaborative filtering leverages this insight into an algorithm that predicts a user's preference of an item by finding a cohort of users who happen to have the same preference as the user. The preference or predicted rating for an item can be deduced by the rating given by a cohort of similar users.

Recommendation Engines
Collaborative Filtering

The distinguishing feature of the collaborative filtering method is that the algorithm considers only the ratings matrix that is, the past user–item interactions. Hence, the collaborative filtering method is item domain independent. The same algorithm can be applied to predicting ratings for movies, music, books, and gardening tools. In fact, the algorithm does not have any knowledge about the items except the ratings given by the users.

Table Known Ratings Matrix

	The Godfather	2001: A Space Odyssey	The Hunt for Red October	Fargo	The Imitation Game	...
Josephine	5	4	1		1	
Olivia			2	2	4	
Amelia	5	1	4	4	1	
Zoe	2	2	5	1	1	
Alanna	5	5		1	1	
Kim		4	1	2	5	
...						

Recommendation Engines
Collaborative Filtering

There are two distinct approaches in processing the ratings matrix and extracting the rating prediction. The neighbourhood method finds a cohort of similar users or items. The latent factor method explains the ratings matrix through a set of dimensions called latent factors.

Table Known Ratings Matrix

	The Godfather	2001: A Space Odyssey	The Hunt for Red October	Fargo	The Imitation Game	...
Josephine	5	4	1		1	
Olivia			2	2	4	
Amelia	5	1	4	4	1	
Zoe	2	2	5	1	1	
Alanna	5	5		1	1	
Kim		4	1	2	5	
...						

Recommendation Engines
Collaborative Filtering – Neighbourhood Method

Neighbourhood methods calculate how similar the users (or the items) are in the known ratings matrix. The similarity scores or the measure of proximity, discussed in the k-Nearest Neighbour Classification techniques, are used in neighbourhood-based methods to identify similar users and items.

Commonly used similarity measures are: Jaccard similarity, Cosine similarity, and Pearson correlation coefficient.

The general approach for neighbourhood-based methods consists of two steps to find the predicted rating for a user-item:

1. Find the cohort of other similar users (or the items) who have rated the item (or the user) in question.

2. Deduce the rating from the ratings of similar users (or the items).

Recommendation Engines
Collaborative Filtering – Neighbourhood Method

The two methods for the neighbourhood-based systems, that is, user-based and item-based, are extremely similar. The former starts to identify similar users and the latter starts by identifying similarly rated items by the same users. For the ratings matrix shown in Table below, both the methods share the same technique, where one starts with finding similar rows and the other finds similar columns.

Table Known Ratings Matrix

	The Godfather	2001: A Space Odyssey	The Hunt for Red October	Fargo	The Imitation Game	...
Josephine	5	4	1		1	
Olivia			2	2	4	
Amelia	5	1	4	4	1	
Zoe	2	2	5	1	1	
Alanna	5	5		1	1	
Kim		4	1	2	5	
...						

Collaborative Filtering
Neighbourhood Method – User Based

The user-based collaborative filtering method operates on the assumption that similar users have similar likes. The two-step process of identifying new unseen user-item preferences consists of filtering similar users and deducing the ratings from similar users. The approach for user-based collaborative filtering is quite similar to the k-NN classification algorithm.

1. For every user x, find a set of N other users who are similar to the user x and who have rated the item i.

2. Approximate the rating of the user x for the item i, by aggregating (averaging) the rating of N similar users.

Collaborative Filtering
Neighbourhood Method – User Based

Step 1: Identifying Similar Users

The users are similar if their rating vectors are close according to a distance measure. Consider the rating matrix shown in Table below as a set of rating vectors. The rating for the user Amelia is represented as (5,1,4,4,1). The similarity between the two users is the similarity between the rating vectors.

Table Known Ratings Matrix and a User-item Rating Prediction

	The Godfather	2001: A Space Odyssey	The Hunt for Red October	Fargo	The Imitation Game
Josephine	5	4	1		1
Olivia		**?**	2	2	4
Amelia	5	1	4	4	1
Zoe	2	2	5	1	1
Alanna	5	5		1	1
Kim		4	1	2	5

Collaborative Filtering
Neighbourhood Method – User Based

Step 1: Identifying Similar Users (cont.)

A quantifying metric is needed in order to measure the similarity between the user's vectors. Jaccard similarity, Cosine similarity, and Pearson correlation coefficient are some of the commonly used distance and similarity metrics. The cosine similarity measure between two nonzero user vectors for the user Olivia and the user Amelia is given by

$$\text{Cosine similarity } (|x \,.\, y|) = \frac{x\,y}{||x||\;||y||}$$

$$\text{Cosine similarity } (r_{\text{olivia}},\ r_{\text{amelia}}) = \frac{0\times5+0\times1+2\times4+2\times4+4\times1}{\sqrt{2^2+2^2+4^2}\times\sqrt{5^2+1^2+4^2+4^2+1^2}} = 0.53$$

Table Known Ratings Matrix and a User-item Rating Prediction

	The Godfather	2001: A Space Odyssey	The Hunt for Red October	Fargo	The Imitation Game
Josephine	5	4	1		1
Olivia		**?**	2	2	4
Amelia	5	1	4	4	1
Zoe	2	2	5	1	1
Alanna	5	5		1	1
Kim		4	1	2	5

Collaborative Filtering
Neighbourhood Method – User Based

Step 1: Identifying Similar Users (cont.)

Note that the cosine similarity measure equates the lack of ratings as zero value ratings, which can also be considered as a low rating. This assumption works fine for the applications for where the user has purchased the item or not. In the movie recommendation case, this assumption can yield the wrong results because the lack of rating does not mean that the user dislikes the movie. Hence, the similarity measure needs to be enhanced to take into consideration the lack of rating being different from a low rating for an item. Moreover, biases in the ratings should also be dealt with. Some users are more generous in giving ratings than others who are more critical. The user's bias in giving ratings skews the similarity score between users.

Collaborative Filtering
Neighbourhood Method – User Based

Step 1: Identifying Similar Users (cont.)

Centred cosine similarity measure addresses the problem by normalizing the ratings across all the users. To achieve this, all the ratings for a user is subtracted from the average rating of the user. Thus, a negative number means below average rating and a positive number means above average ratings given by the same user. The normalized version of the ratings matrix is shown in Table below. Each value of the ratings matrix is normalized with the average rating of the user and the similarity metric calculated accordingly.

Table Normalized Ratings Matrix

	The Godfather	2001: A Space Odyssey	The Hunt for Red October	Fargo	The Imitation Game
Josephine	2.3	1.3	−1.8		−1.8
Olivia			−0.7	−0.7	1.3
Amelia	2.0	−2.0	1.0	1.0	−2.0
Zoe	−0.2	−0.2	2.8	−1.2	−1.2
Alanna	2.0	2.0		−2.0	−2.0
Kim		1.0	−2.0	−1.0	2.0

Collaborative Filtering
Neighbourhood Method – User Based

Step 1: Identifying Similar Users (cont.)

The similarity score can be pre-computed between all the possible pairs of users and the results can be kept ready in a user-to-user matrix shown in sample Table below for ease of calculation in the further steps. At this point the neighbourhood or cohort size, k, has to be declared, similar to k in the k-NN algorithm. Assume k is 3 for this example. The goal of this step is to find three users similar to the user Olivia who have also rated the movie 2001: A Space Odyssey. From the table, the top three users can be found similar to the user Olivia, who are Kim (0.90), Alanna (-0.20), and Josephine (-0.20).

Table User-to-User Similarity Matrix

	Josephine	Olivia	Amelia	Zoe	Alanna	Kim
Josephine	1.00	−0.20	0.28	−0.30	0.74	0.11
Olivia		1.00	−0.65	−0.50	−0.20	0.90
Amelia			1.00	0.33	0.13	−0.74
Zoe				1.00	0.30	−0.67
Alanna					1.00	0.00
Kim						1.00

Collaborative Filtering
Neighbourhood Method – User Based

Step 2: Deducing Rating From Neighbourhood Users

Once the cohort of users are found, deducing the predicted rating is straight forward. The predicted rating for Olivia for the movie 2001: A Space Odyssey is the average of the ratings given by Kim, Alanna, and Josephine, for the same movie.

The user-based neighbourhood technique provides an intuitive way to fill the ratings matrix. The steps for finding similar users and rating deductions is repeated for every blank ratings cell in the ratings matrix. However, the collaborative filtering process can be time consuming. One way to speed-up the process is to pre-compute the user-to-user similarity matrix. The user-to-user similarity matrix should be updated for any new user.

However, a new user similarity with other users can be computed only when their preference information is known—cold start problem!

Collaborative Filtering
Neighbourhood Method – User Based

Step 2: Deducing Rating From Neighbourhood Users (cont.)

The cold start problem can be mitigated by a couple of strategies. First, the system can ask all the new users to select or enter their preferred items once they have signed up. This can be selecting a few movies from a curated list of movies or tagging them from the movie catalogue. The new user onboarding process incorporates this step, so the recommendation engines have some seed information about the new users. Second, the system can rely heavily on implicit data collection through search or clickstream activity until a solid item preference profile can be build.

Popular items are preferred by a large number of users. Collaborative filtering tends to recommend popular items to users because the cohort selection might be skewed towards higher ratings for popular items. Recommending the popular items is not necessarily a bad decision. However, one of the objectives of recommendation engine is to discover the personalized idiosyncratic and unique items of the user. Moreover, the non-personalized popular items are usually shown by best seller or trending now lists, which have the same content for all the users.

Collaborative Filtering
Neighbourhood Method – Item Based

The **item-based** neighbourhood method operates on the assumption that users prefer items that are similar to previously preferred items. In this context, items that are similar tend to be rated similarly by the same users. If a user liked the dark comedy crime movie Fargo, then the user might like the crime thriller movie No Country for Old Men, provided that both movies are rated similarly by the same users.

The two-step process of identifying new user-item preference using item-based collaborative filtering include identifying similar items and deducing a rating from similar items.

1. For every item i, find a set of N other items which have similar ratings when rated by the same user.

2. Approximate the rating for the item i by aggregating (averaging) the rating of N similar items rated by the user.

Collaborative Filtering
Neighbourhood Method – Item Based

The ratings matrix can be used to compute the predicted rating for the same unseen cell—for the user Olivia for the movie 2001: A Space Odyssey–using the item–based method. To realize the item–based neighbourhood method, the rating matrix has to be transposed (swapping rows and columns) and continued with the same steps as the user–based (or row based) neighbourhood method. below shows the transposed version of the original ratings matrix.

Table Transposed Ratings Matrix

	Josephine	Olivia	Amelia	Zoe	Alanna	Kim
The Godfather	5		5	2	5	
2001: A Space Odyssey	4	?	1	2	5	4
The Hunt for Red October	1	2	4	5		1
Fargo		2	4	1	1	2
The Imitation Game	1	4	1	1	1	5

Collaborative Filtering
Neighbourhood Method – Item Based

The centred cosine or Pearson correlation coefficient metric is used to calculate the similarity between movies based on the ratings pattern. Since the objective is to find the rating for 2001: A Space Odyssey, the similarity score would need to be found for all the movies with 2001: A Space Odyssey.

Table Transposed Ratings Matrix

	Josephine	Olivia	Amelia	Zoe	Alanna	Kim
The Godfather	5		5	2	5	
2001: A Space Odyssey	4	**?**	1	2	5	4
The Hunt for Red October	1	2	4	5		1
Fargo		2	4	1	1	2
The Imitation Game	1	4	1	1	1	5

Collaborative Filtering
Neighbourhood Method – Item Based

Table below shows the centred rating values along with the similarity score of all the movies with 2001: A Space Odyssey. The similarity score is calculated using Eq. below on the centred rating values. Since the centred ratings can be negative, the similarity score can be positive or negative. Depending on the number of neighbours specified, the top k neighbours to the movie 2001: A Space Odyssey can now be narrowed down using the magnitude of the similarity score. Assume k is 2.

Table Normalized Ratings and Similarity With a Movie

	Josephine	Olivia	Amelia	Zoe	Alanna	Kim	Similarity with the Movie 2001: A Space Odyssey
The Godfather	2.3		2.0	−0.2	2.0		**−0.10**
2001: A Space Odyssey	1.3		−2.0	−0.2	2.0	1.0	**1.00**
The Hunt for Red October	−1.8	−0.7	1.0	2.8		−2.0	**−0.36**
Fargo		−0.7	1.0	−1.2	−2.0	−1.0	**0.24**
The Imitation Game	−1.8	1.3	−2.0	−1.2	−2.0	2.0	**−0.43**

Collaborative Filtering
Neighbourhood Method – Item Based

From Table below the nearest two movies to 2001: A space Odyssey can be concluded, rated by Olivia, are Fargo and The Hunt for Red October. The predicted centred rating for the 2001: A space Odyssey for Olivia, using Eq. is:

$$\frac{(0.24 \times -0.7) + (-0.36 \times -0.7)}{(0.24 - 0.36)} = -0.67$$

The normalized rating for Olivia and 2001: A space Odyssey is -0.67 and the real ratings for 2001: A space Odyssey by Olivia is 2.

Table Normalized Ratings and Similarity With a Movie

	Josephine	Olivia	Amelia	Zoe	Alanna	Kim	Similarity with the Movie 2001: A Space Odyssey
The Godfather	2.3		2.0	−0.2	2.0		**−0.10**
2001: A Space Odyssey	1.3		−2.0	−0.2	2.0	1.0	**1.00**
The Hunt for Red October	−1.8	−0.7	1.0	2.8		−2.0	**−0.36**
Fargo		−0.7	1.0	−1.2	−2.0	−1.0	**0.24**
The Imitation Game	−1.8	1.3	−2.0	−1.2	−2.0	2.0	**−0.43**

Collaborative Filtering
Neighbourhood Method – Item Based

User-Based or Item-Based Collaborative Filtering?

The neighbourhood technique for predicting a rating for a user-item combination, either with user-based or item-based, is very similar. After all, if the ratings matrix is transposed at the beginning, the item-based approach is exactly the same as the user-based approach. However, the predicted rating is different when these two approaches are used on the same ratings matrix.

Conceptually, finding similar items is relatively easier than finding similar users. Items tend to get aligned with specific genres or types of the items. A movie can belong to either Classics or Science fiction genres or, less likely, in both the genres. However, a user may like both Classics and Science fiction genres. It is common for the users to have interests in multiple genres and develop unique taste profiles.

Collaborative Filtering
Performance – Neighbourhood Method – Item Based

Mean absolute error (MAE) The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction. Expressed in words, the MAE is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

Root mean squared error (RMSE) The RMSE is a quadratic scoring rule which measures the average magnitude of the error. Expressing the formula in words, the difference between forecast and corresponding observed values are each squared and then averaged over the sample. Finally, the square root of the average is taken. Since the errors are squared before they are averaged, the RMSE gives a relatively high weight to large errors. This means the RMSE is most useful when large errors are particularly undesirable.

Collaborative Filtering
Performance – Neighbourhood Method – Item Based

The MAE and the RMSE can be used together to diagnose the variation in the errors in a set of forecasts. The RMSE will always be larger or equal to the MAE; the greater difference between them, the greater the **variance** in the individual errors in the sample. If the RMSE=MAE, then all the errors are of the same magnitude.

Both the MAE and RMSE can range from 0 to ∞. They are negatively-oriented scores: Lower values are better.