Collaborative Filtering on Movie Ratings

27.11.2018

—

Pournami Krishnan

Charles Hu

Overview

This paper documents the process and results for using the collaborative filtering method to provide movie recommendations to users. We will analyze and compare the quality of the final movie recommendations for a user with a dataset that has 25% missing data, and another sparser data set with 75% missing data.

What is collaborative filtering?

Collaborative filtering is a machine learning based system that recommends content to users, based on how similar in taste the users have with each other. For example, if 2 people rate things similarly, then the logic is that if person 1 likes something, then it is likely that person 2 will like it too. In contrast to content-based filtering, it is possible to provide recommendations using collaborative filtering without needing to understand the features of the items.

Dataset overview

We pulled a set of real movie ratings from real users from MovieLens.org (https://grouplens.org/datasets/movielens/latest/), a non-commercial organization that provides movie recommendations. From a set of 9,000 movies by 600 users, we cut it to a dataset that includes ratings for 20 movies by 50 users. We manually selected users based on who rated as many of the same movies as possible. Finally, we randomly deleted 25% of ratings data to create set 1, and randomly deleted 75% of ratings data to create set 2.

Set 1 (25% missing data)

	Aliza Smart	Kaylen Khan	Eesa Dotson	Eugene Franklin	Lynn Lott	Oakley Bautista	Nancie Decker	Lola-Rose Padilla	Kylan Whyte	Shaunie Hassan	Yasin Cousins
Toy Story	4		3.5	4	3	3	3	5		5	4
Grumpier Old Man	4	4		3		3	4		4	5	4
Heat	4	5	4		3.5	3	4	3			4
Seven (a.k.a. Se7en)	5	4	4.5	3	3	4	4	4	4	4	4
The Usual Suspects	5	1	5		3.5	5	5		5		5
From Dusk till Dawn	3		3.5	2			3			3	4
Bottle Rocket	5						3				4
Braveheart	4	5	4.5		3.5	5	5	4	4	3	5
Rob Roy	5	4					4	4		3	4
Canadian Bacon	5			2		3	3	3	3		4
Desperado	5	3	4		3	3	4	5	3		4
Billy Madison	5	4		2			3	1		3	5
Clerks	3		4	3		3	4	5	5		5
Dumb & Dumber	5	3	2.5	2			3	1	4	3	5
Ed Wood	4	3	4.5	4	3	3	3	4	3	3	4
Star Wars: Episode IV - A new hope	5		4	4	4	5	5	3	5		5
Pulp Fiction	3	2	4	3	4.5	5	5	3	5	4	5
Stargate	3	5		2	4	5	5	3	2	4	4
Tommy Boy	5	5		2			4		5	3	4
Clear and Present Danger	4	5			3	3	3		4	3	4

	Maxwell Mcneill	Jimmy Kimmel	Tonya Cunningham	Amaya Bass	Sommer Larsen	Chace Wilkerson	Irfan Ross	Merryn Mccullouch	Jamie Thomson	Alec Ridley	Yannis Villa	Zoha Hills	Kaylan Lim
Toy Story	5	3	4	2.5	2.5	4	3	4	2	4	5	4	2
Grumpier Old Man		3		2	3	3.5	3		2	3		4	
Heat	3	3		4	3.5	4.5	3		2	3		3	3
Seven (a.k.a. Se7en)	3	5	5	4	3.5	4.5	4	5	2	2	4.5	4	3
The Usual Suspects	5	5	4	3	3	3	5	4	3	4	4	3	
From Dusk till Dawn	1	3		4	3	4	3		3		2	3	3
Bottle Rocket		3					4				2	3	
Braveheart	4	5	5	2.5	4.5	4	3	5	2	4		5	3
Rob Roy	3	3				3	4				2	3	3
Canadian Bacon	3	3	2	3	3		4	3	2	3	2	3	2
Desperado		3	4	4.5	4	3.5	3	4	4	4	2	3	2
Billy Madison		3		3.5			4		3		2	3	
Clerks		3	3.5	1		3	4	3	1	3	2.5	3	2
Dumb & Dumber		5	3	3.5	4	4.5	4	3	3	2		3	
Ed Wood	5	3	4	3	3	3	3	3	2	3	3	3	2
Star Wars: Episode IV - A new hope	5	3	2.5	5	3.5	3.5	3	3	5	2	2.5	4	
Pulp Fiction	3	5	5	2	4	4.5	4	3	1	5	5	4	2
Stargate	3	3		3.5	3	3	3		4	2		3	3
Tommy Boy		5		4		4.5	3	3	2		2	3	
Clear and Present Danger	4	3	4				3		1	3		4	4

	Mila Manning	Larissa Yu	Rudi Hebert	Klara Hurst	Gerald Haigh	Kasey Finley	Kierran Moore	Cleveland Wong	Elsie Mayer	Brittney French	Jem Betts	Barney Morse	Daanyaal Smyth
Toy Story	4	4	3	3	3.5	3	3.5	4	4	4	4	3.5	3.5
Grumpier Old Man			5	4		4	3	2	2	3	1		3.5
Heat	4.5	4	5	4		5			2	5	2	3.5	
Seven (a.k.a. Se7en)	4	4	5	3	4	4	4.5	3.5	5	4	2	3.5	5
The Usual Suspects	4.5	5	5	4		3		4	4	4	3	5	
From Dusk till Dawn	4.5	3		4	3	3	3.5	3	3	4		3	
Bottle Rocket			3			4		3		3		2	
Braveheart	3.5	3	3		4.5	4	4.5		4	4	2		
Rob Roy		4		3	3	4				3			3
Canadian Bacon	2	3	5		4	3	3		3	3	2	3	
Desperado	3.5	3	5	3	4	3	4	3	3	4			
Billy Madison					3	3				3	1		5
Clerks	4.5	5	3	3	4	3	4	4	3	3	4	4	5
Dumb & Dumber	0.5			3		4	3			3		1.5	5
Ed Wood	4.5	5	3	3	3	3	3.5	4	3	4	3	3	3
Star Wars: Episode IV - A new hope	3.5		5	4	5	4		1	5	4	5	5	4.5
Pulp Fiction	5	5	5	4	5	4	4	5	5	4	2	4	4.5
Stargate	2.5	1	5		5	3			2	4	3	2.5	2.5
Tommy Boy	3.5					3	3	3		3			4
Clear and Present Danger			3	3.5		4				3	4		

	Bridget Ellison	Beyonce Derrick	Amarah Trujillo	Mariyam Ibarra	Cheyenne Schaefer	Forrest Marquez	Rayan Kramer	Miranda O'Sullivan	Alvin Mcgrath	Tony Morrow	Filip Cook	Carley Schmidt	Firat Nolan
Toy Story	5	4	2	4		4.5		2	5		4	1	3
Grumpier Old Man	4	3				4	1				3.5	0.5	4
Heat		4	4	4			3		2	3.5	3	1	3
Seven (a.k.a. Se7en)	3	5	3	4	3	3.5	2	4	4	4	4	1	5
The Usual Suspects		4	4	4			3	3.5		5	4.5	2.5	5
From Dusk till Dawn	3	4	4	4.5			4			5	3.5	1	4
Bottle Rocket		3		3.5							4.5	1	5
Braveheart	5	5	5	4.5	5	5		3	5		3.5	1	3
Rob Roy		4				4			4		3	1	3
Canadian Bacon	2	3		3	3	3	3	3	3	3	2	1	4
Desperado		3		3.5		3	3	2.5		5	3	2	3
Billy Madison	5	3		3.5	4			3			4	1	4
Clerks		4	4	4	5			4		4	4	1	5
Dumb & Dumber	4	4		4.5	4	3		2.5		1	2.5	1	3
Ed Wood	3	4	4	3	3	5	3	3	3	4.5	3	1	3
Star Wars: Episode IV - A new hope		5	4	3		5	5	4	5	5	3.5	2.5	3
Pulp Fiction	2	4	5	5	4	5	4	4.5	5	5	4.5	1	5
Stargate	4	4	3	3.5	4		3		2	3	3	1	4
Tommy Boy	5	3		3.5	4		1	3	2		3	1	4
Clear and Present Danger	3	3			3	3	2		4		2	2	3

View sheet Ratings (25%) for expanded view

Set 2 - Sparse (75% missing data)

	Aliza Smart	Kaylen Khan	Eesa Dotson	Eugene Franklin	Lynn Lott	Oakley Bautista	Nancie Decker	Lola-Rose Padilla	Kylan Whyte	Shaunie Hassan	Yasin Cousins	Maxwell Mcneill	Jimmy Kimmel	Tonya Cunningham	Amaya Bass	Sommer Larsen	Chace Wilkerson
Toy Story				4		3		5									
Grumpier Old Man	4	4		3					4	5			3		2		3.5
Heat	4	5	4			3					4	3			4	3.5	
Seven (a.k.a. Se7en)					3					4							
The Usual Suspects			5			5	5		5					4			
From Dusk till Dawn	3			2							4	1			4		4
Bottle Rocket																	
Braveheart	4						5		4	3			5	5		4.5	
Rob Roy		4					4	4				3					3
Canadian Bacon																	
Desperado			4					5	3							4	
Billy Madison	5	4		2				1			5				3.5		
Clerks	3		4				4						3	3.5	1		3
Dumb & Dumber		3						1	4	3	5		5			4	4.5
Ed Wood																	
Star Wars: Episode IV - A new hope			4		4	5						5		2.5			
Pulp Fiction		2				5	5							5			
Stargate		5								4		3					
Tommy Boy	5			2							4		5		4		4.5
Clear and Present Danger					3				4	3		4	3				

	Irfan Ross	Merryn Mcculloch	Jamie Thomson	Alec Ridley	Yannis Villa	Zoha Hills	Kaylan Lim	Mila Manning	Larissa Yu	Rudi Hebert	Klara Hurst	Gerald Haigh	Kasey Finley	Kierran Moore	Cleveland Wong	Elsie Mayer	Brittney French
Toy Story				4					4			3.5	3	3.5	4	4	4
Grumpier Old Man						4											
Heat								4.5	4		4		5			2	5
Seven (a.k.a. Se7en)	4	5	2		4.5	4	3				3	4	4	4.5		5	
The Usual Suspects	5			4	4					5							4
From Dusk till Dawn								4.5									
Bottle Rocket																	
Braveheart	3	5		4		5		3.5		3							4
Rob Roy									4								
Canadian Bacon							2										
Desperado							2	3.5					3	4			
Billy Madison					2												
Clerks					2.5				5		3				4		
Dumb & Dumber	4	3				3								3			
Ed Wood	3		2												4		
Star Wars: Episode IV - A new hope			5							5	4	5			1	5	
Pulp Fiction					5		2						4	4	5	5	4
Stargate								2.5	1								
Tommy Boy																	
Clear and Present Danger				3		4	4				3.5		4				

	Jem Betts	Barney Morse	Daanyaal Smyth	Bridget Ellison	Beyonce Derrick	Amarah Trujillo	Mariyam Ibarra	Cheyenne Schaefer	Forrest Marquez	Rayan Kramer	Miranda O'Sullivan	Alvin Mcgrath	Tony Morrow	Filip Cook	Carley Schmidt	Firat Nolan
Toy Story												5				
Grumpier Old Man	1			4					4	1						
Heat	2					4	4			3			3.5		1	3
Seven (a.k.a. Se7en)		3.5			5											
The Usual Suspects															2.5	
From Dusk till Dawn					4					4			5			4
Bottle Rocket							3.5							4.5		5
Braveheart					5	5					3	5			1	
Rob Roy					4				4			4				
Canadian Bacon	2									3				2		
Desperado							3.5				2.5		5			3
Billy Madison	1		5	5				4						4		
Clerks	4	4	5			4	4				4		4		1	5
Dumb & Dumber		1.5	5	4				4	3				1	2.5		
Ed Wood						4		3								
Star Wars: Episode IV - A new hope		5			5						4	5			2.5	
Pulp Fiction						5	5		5				5			5
Stargate			2.5													
Tommy Boy			4	5							3			3		
Clear and Present Danger	4			3				3	3	2		4				

View sheet Sparsity Rating (75%) for expanded view

Calculating closest neighbors based on Set 1 and Set 2 using Pearson Correlation

To determine closest neighbors based on data from both Set 1 and 2, we first used the Pearson correlation equation to calculate the user similarity between all the users based on their ratings.

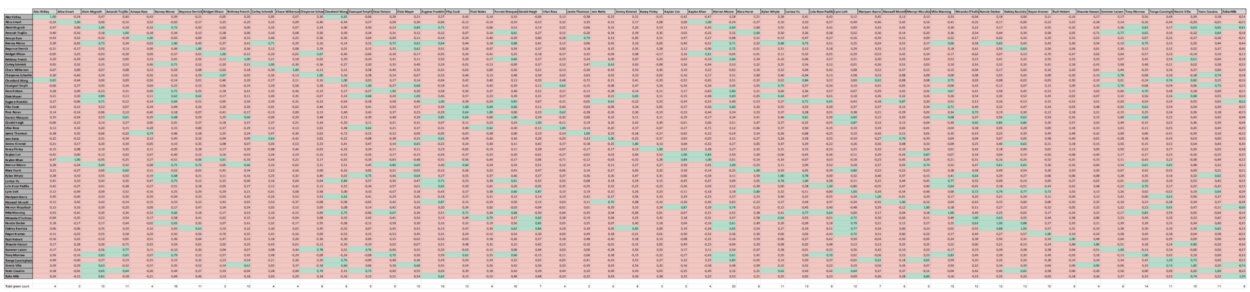
$$\mathrm{PC}(u,v) = \frac{\sum_{i \in \mathcal{I}_{uv}} (r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in \mathcal{I}_{uv}} (r_{ui} - \bar{r}_u)^2 \sum_{i \in \mathcal{I}_{uv}} (r_{vi} - \bar{r}_v)^2}}.$$

We used the below Python code to perform the above equation and determine the correlation between the ratings of 2 users. We repeated this for all the permutations of 2 different users.

```
# PEARSON CORRELATION USING FORMULA

def average(x):
    assert len(x) > 0
    return float(sum(x)) / len(x)

def pearson_def(x, y):
    assert len(x) == len(y)
    n = len(x)
    if (n > 0):
        avg_x = average(x)
        avg_y = average(y)
        diffprod = 0
        xdiff2 = 0
        ydiff2 = 0
        for idx in range(n):
            xdiff = x[idx] - avg_x
            ydiff = y[idx] - avg_y
            diffprod += xdiff * ydiff
            xdiff2 += xdiff * xdiff
            ydiff2 += ydiff * ydiff

        return diffprod / math.sqrt(xdiff2 * ydiff2)
    else:
        return None
```

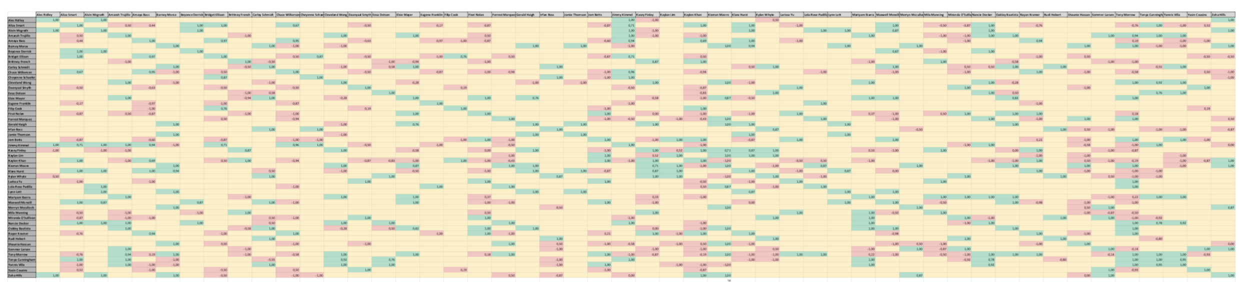
The results for both 2 sets of data are below:

Correlation of Users from Set 1 (25% missing data)



View sheet Pearson Correlation (25%) for expanded view

Correlation of Users from Set 2 (75% missing data)



View sheet Pearson Correlation (75%) for expanded view

Conclusions of Pearson correlation

We filtered the excel spreadsheet to show any pearson correlation equal to or above 0.6 to be green to indicate that it is strong. Anything below 0.6 was colored as red to indicate weak.

From the 25% missing data set, every user typically had 10 or less users with strong similarity. There were almost no users with strong negative correlations.

We also noticed that many correlations could not be determined and are therefore empty when using the Pearson correlation on the data set with 75% missing data. This is because there isn't enough data to determine the correlation, thus it is 0 and correlation is not derived. The results were therefore more extreme. For every user, roughly more than half of the 50 users did not have a determinable correlation. For the ones that did have a correlation, they tended to either be very strong (above 0.6) or very negative (below -0.6).

Determining user based rating prediction using Set 1 and Set 2

To give rating based predictions, we must first determine a user's top 10 nearest neighbors. Because the 75% missing data set has far less correlations, we needed to use this set first to pick a user with enough correlations. If a user has enough correlations for this 75% missing data set, then they would also therefore surely have enough correlations in the 25% missing set. We ultimately picked user Kierran Moore, because he had the highest number of positive correlations from the 75% missing data set, 14.

Kierran Moore's top 10 neighbors from both data sets and their respective correlations are below:

	PC Ratings (25%)	PC Ratings (75%)
	Kierran Moore	Kierran Moore
Barney Morse	0,71	1,00
Cleveland Wong	0,51	1,00
Forrest Marquez	0,34	1,00
Gerald Haigh	0,51	1,00
Kasey Finley	0,27	0,71
Merryn Mcculloch	0,74	1,00
Oakley Bautista	0,65	1,00
Shaunie Hassan	-0,04	1,00
Tony Morrow	0,61	1,00
Zoha Hills	0,53	1,00

Looking at this table, we can easily see the discrepancies when there are not enough correlations. The system tends to give extreme results when there is a lack of data, and in our case, extreme positive correlation results. For example, Shaunie Hassan's correlation with Kierran was very negligable at -0.04, but the system thought she was perfectly similar with Kierran at 1.0 when there was 75% missing data.

Now that we know the top 10 nearest neighbors, their respective correlations, and their rating, we can predict Kierran's movie ratings using the below formula:

- The predicted rating $\hat{r}_{ui}$ can be calculated as the average of the ratings by neighbors.
- $\mathcal{N}_i(u)$ stands for the set of neighbors that have rated i.
- But we also need to take into account the similarity of *u* to each neighbor *v* (w_{uv}) and get a right value in the allowed range of ratings.

$$\hat{r}_{ui} = \frac{\sum_{v \in \mathcal{N}_i(u)} w_{uv} r_{vi}}{\sum_{v \in \mathcal{N}_i(u)} |w_{uv}|}.$$

We chose to predict the rating for Kierran for 2 movies, Braveheart and Rob Roy because they are similar, historical war-action movies. See predicted ratings on the next page:

Prediction for Braveheart, 25% and 75% missing data

Kierran actually already rated Braveheart with a rating of 4.5. Thus, predicting his rating for Braveheart would help us understand how accurate our collaborative filtering model is. We plugged in the data for the formula into excel below, based on the correlation of top 10 nearest neighbors and their ratings for Braveheart:

25% missing data

```
=((0.71*0)+(0.51*0)+(0.34*5)+(0.51*4.5)+(0.27*4)+(0.74*5)+(0.65*5)+(-0.04*3)+(0.61*5)+(0.53*5))/(4.83)
```

75% missing data

```
=((0.71*0)+(0.51*0)+(0.34*0)+(0.51*0)+(0.27*0)+(0.74*5)+(0.65*0)+(-0.04*3)+(0.61*5)+(0.53*5))/(4.83)
```

Note: the 4.83 in the denominator is the sum of the top 10 neighbor correlations from the 25% missing data set (see Excel for calculation)

	Braveheart	
	Ratings (25%)	Ratings (75%)
Barney Morse		
Cleveland Wong		
Forrest Marquez	5	
Gerald Haigh	4.5	
Kasey Finley	4	
Merryn Mcculloch	5	5
Oakley Bautista	5	
Shaunie Hassan	3	3
Tony Morrow	5	5
Zoha Hills	5	5
Kierran Moore predicted rating	3.64	1.92

Seeing the results, we see that using 25% missing data, our model predicts a reasonably close 3.64 rating when compared to Kierran's real 4.5 rating for Braveheart. With 75% missing data however, our model is much less accurate and predicts a rating of only 1.92.

Prediction for Rob Roy, 25% and 75% missing data

Similar to Braveheart, we predict the Kierran's ratings for Rob Roy by plugging the correlations and ratings for Rob Roy for the top 10 neighbors into the formula in Excel.

25% missing data

```
=((1*0)+(1*0)+(1*4)+(1*3)+(0.71*4)+(1*0)+(1*0)+(1*3)+(1*0)+(1*3))/9.71
```

75% missing data

```
=((1*0)+(1*0)+(1*4)+(1*0)+(0.71*0)+(1*0)+(1*0)+(1*0)+(1*0)+(1*0))/9.71
```

Note: The 9.71 in the denominator is the sum of the top 10 neighbor correlations from the 75% missing data set (See excel for calculation).

	Rob Roy	
	Ratings (25%)	**Ratings (75%)**
Barney Morse		
Cleveland Wong		
Forrest Marquez	4	4
Gerald Haigh	3	
Kasey Finley	4	
Merryn Mcculloch		
Oakley Bautista		
Shaunie Hassan	3	
Tony Morrow		
Zoha Hills	3	
Kierran Moore predicted rating	1.63	0.41

Here, we see that the results from the model are not very accurate. Kierran is predicted to have a rating of 1.63 with 25% missing data, although the 5 other ratings are 3 or above. Similarly, Kierran is predicted to have a rating of 0.41 with 75% missing data, when the only other rating provided is a 4. These discrepancies can be explained due to the general lack of ratings available for Rob Roy. This causes the numerator of the formula to grow smaller, dragging down the overall recommended rating score.

Conclusion

I. More rating data means more accurate pearson correlations

As seen from the results of the pearson correlation calculation using the 2 sets of data, using 25% missing data provides a more accurate and varied set of correlations. WIth 75% missing data however, the correlation overestimates the strength of the relationship between users.

II. More rating data means more accurate rating suggestions

When there is a lack of sufficient ratings, the collaborative filtering rating suggestions tends to be lower. This phenomena can be seen through the application of the ratings formula, as the numerator and overall rating will be lower and lower.

III. Collaborative filtering still has its benefits

Despite some of these issues with collaborative filtering, it still has value because it is able to provide ratings when there is no information available about the content itself. It can also provide novel recommendations that aren't extremely similar content wise.