

# Collaborative Filtering - An Introduction

# Online Recommendations

## Frequently Bought Together



Total price: ₹39,701.00

Add all three to Cart

- ✓ This item: Apple iPhone 6 (Gold, 16GB) ₹39,499.00
- ✓ Premium Tempered Glass Screen Guard for Apple iPhone 6 ₹114.00
- ✓ Grabmore Ultra Thin Silicon Case For Apple iPhone 6 (Transparent) ₹88.00

## Customers Who Bought This Item Also Bought

Page 1 of 10



Premium Tempered Glass Screen Guard for Apple iPhone 6  
★★★★☆ 158  
₹114.00



Grabmore Ultra Thin Silicon Case For Apple iPhone 6 (Transparent)  
★★★★☆ 369  
₹88.00



Apple iPhone 5s (Gold, 16GB)  
★★★★☆ 2,431  
₹28,605.00



Apple iPhone 5s (Gold, 32GB)  
★★★★☆ 2,431  
₹38,914.00



Camera Lens Protective Case Cover Ring Installed for Apple iPhone 6 (gold)  
★★★★☆ 79  
₹70.00



Sony Xperia M2 (Black, 8GB)  
★★★★☆ 202  
₹12,545.00



OneAssist Premium Protection Plan for Mobile Protection and Assistance Services (Rs. 34001-45000)  
★★★★☆ 5



Apple iPhone 6 Plus (Gold, 16GB)  
★★★★☆ 348  
₹48,298.00



Spigen iPhone 6 4.7-Inch Case Thin Fit A (Champagne Gold) (SGP10943)  
★★★★☆ 58  
₹899.00

## Sponsored Products Related To This Item (What's this?)

Page 1 of 11



Snugg&#8482; iPhone 6 Plus Case - Leather Flip Case with Lifetime Guaran...  
★★★★☆ (54)  
₹1,395.00 **Fulfilled**



NAS Crystal Clear Transparent Ultra Thin Silicon Back Case Cover For Iph...  
★★★★☆ (6)  
₹299.00



iPhone 6 Leather Wallet Case - Red, Executive Style With Convenient Card...  
₹149.00 **Fulfilled**



NAS iPhone 6 Screen Guard Protector Premium Quality Crystal Clear Transp...  
★★★★☆ (8)  
₹199.00



Tech Armor Ballistic Glass Screen Protector with Anti-Fingerprint Coat...  
★★★★☆ (8)  
₹1,450.00 **Fulfilled**



Bushbuck Baronage Classical Edition Genuine Leather Case for iPhone 6 (4...  
★★★★☆ (1)  
₹1,600.00 **Fulfilled**



id America Liquid Rigid-Flex Case iPhone 6 (4.7) - Black  
₹1,100.00 **Fulfilled**

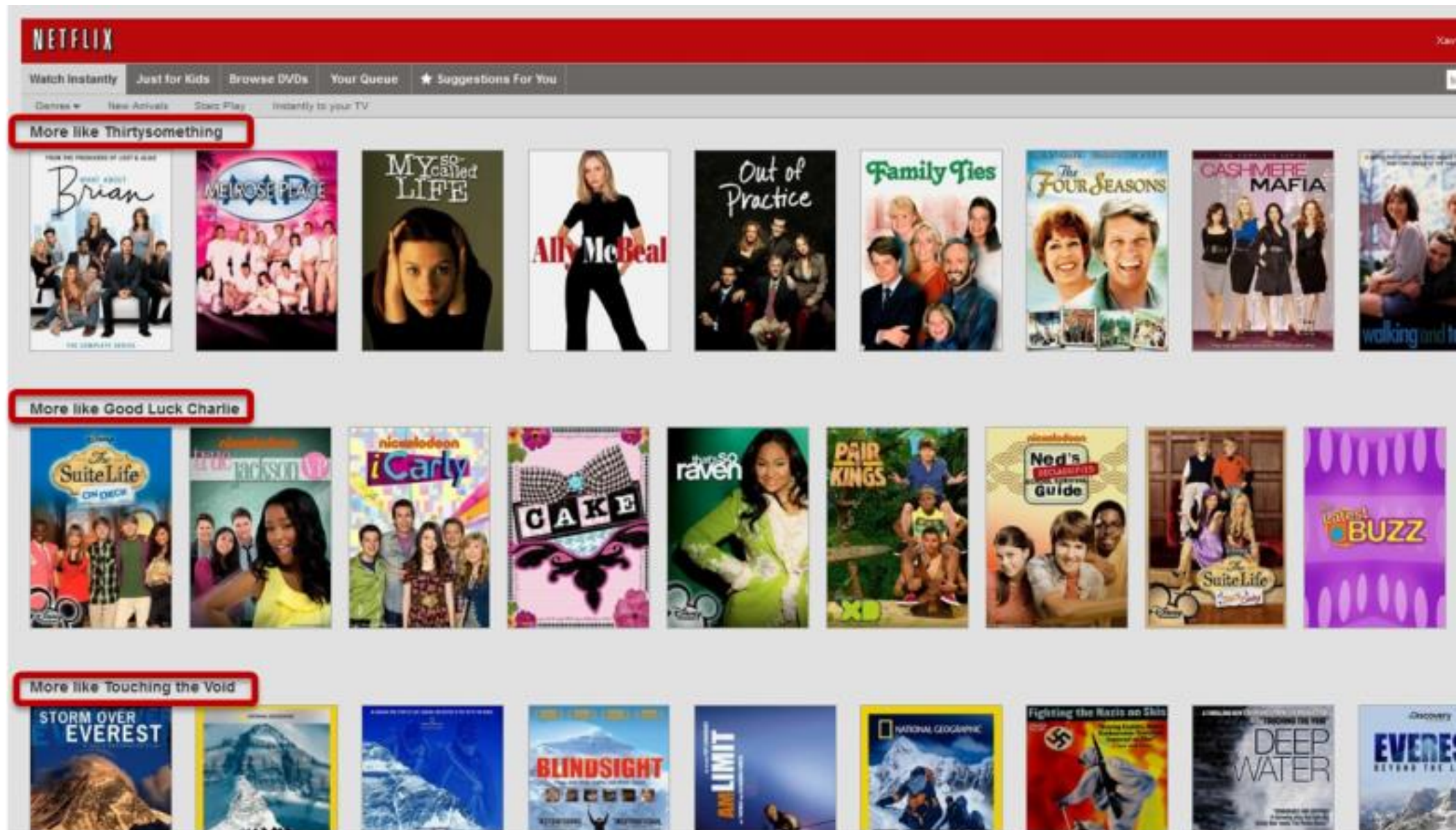


Snugg&#8482; iPhone 6 Plus Case - Leather Flip Case with Lifetime Guaran...  
★★★★☆ (53)  
₹1,395.00 **Fulfilled**



Tech Armor Ballistic Glass Screen Protector with Anti-Fingerprint Coat...  
★★★★☆ (4)  
₹1,199.00 **Fulfilled**

# Netflix







# Data Mining Methods and Nature of Data

**TABLE 1.1**

**ORGANIZATION OF DATA MINING METHODS IN THIS BOOK, ACCORDING TO THE NATURE OF THE DATA\***

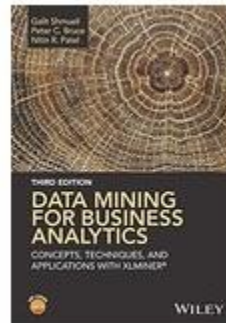
	Supervised		Unsupervised
	Continuous Response	Categorical Response	No Response
Continuous predictors	Linear regression (6) Neural nets (11) $k$ -Nearest neighbors (7)	Logistic regression (10) Neural nets (11) Discriminant analysis (12)	Principal components (4) Cluster analysis (15) Collaborative filtering (14) 
	Ensembles (13)	$k$ -Nearest neighbors (7) Ensembles (13)	
Categorical predictors	Linear regression (6) Neural nets (11)	Neural nets (11) Classification trees (9)	Association rules (14) Collaborative filtering (14) 
	Regression trees (9) Ensembles (13)	Logistic regression (10) Naïve Bayes (8) Ensembles (13)	

\* Numbers in parentheses indicate chapter number.

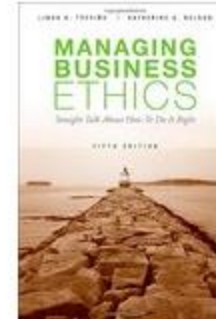
# Collaborative Filtering

- User based methods
- Item based methods

## Customers Who Bought This Item Also Bought



Practical Management  
Science (with Essential  
Textbook Resources...  
› Wayne L. Winston



Managing Business Ethics:  
Straight Talk about How to  
Do It Right  
› Linda K. Trevino



Data Science for Business:  
What You Need to Know  
about Data Mining and...  
› Foster Provost

# Item-user matrix

- Cells are user preferences,  $r_{ij}$ , for items
- Preferences can be ratings, or binary (buy, click, like)

User ID	Item ID			
	$I_1$	$I_2$	$\dots$	$I_p$
$U_1$	$r_{1,1}$	$r_{1,2}$	$\dots$	$r_{1,p}$
$U_2$	$r_{2,1}$	$r_{2,2}$	$\dots$	$r_{2,p}$
$\vdots$				
$U_n$	$r_{n,1}$	$r_{n,2}$	$\dots$	$r_{n,p}$

More efficient to store as rows of triplets

Each row has the user ID, the item ID, and the user's rating of that item

$(U_u, I_i, r_{ui})$

# User-based Collaborative Filtering

- Start with a single user who will be the target of the recommendations
- Find other users who are most similar, based on comparing preference vectors



# Measuring Proximity

- Like nearest-neighbor algorithm
- But Euclidean distance does not do well
- Correlation proximity does better (Pearson)
- For each user pair, find the co-rated items, calculate correlation between the vectors of their ratings for those items
  - Note that the average ratings for each user are across all products, not just the co-rated ones

$$\text{Corr}(U_1, U_2) = \frac{\sum (r_{1,i} - \bar{r}_1)(r_{2,i} - \bar{r}_2)}{\sqrt{\sum (r_{1,i} - \bar{r}_1)^2} \sqrt{\sum (r_{2,i} - \bar{r}_2)^2}}$$

## Example – Tiny Netflix subset

Customer ID	1	5	8	17	18	28	30	44	48
30878	4	1			3	3	4	5	
124105	4								
822109	5								
823519	3		1	4		4	5		
885013	4	5							
893988	3						4	4	
1248029	3					2	4		3
1503895	4								
1842128	4						3		
2238063	3								

TABLE 14.8

SAMPLE OF RECORDS FROM THE NETFLIX PRIZE CONTEST, FOR A SUBSET OF 10 CUSTOMERS AND 9 MOVIES

Consider users 30878 and 823519

# Correlation between users 30878 and 823519

First find average ratings for each user:

$$\bar{r}_{30878} = (4 + 1 + 3 + 3 + 4 + 5)/6 = 3.333$$

$$\bar{r}_{823519} = (3 + 1 + 4 + 4 + 5)/5 = 3.4$$

Find correlation using departure from avg. ratings for the co-rated movies (movies 1, 28 and 30):

$$\text{Corr}(U_1, U_2) = \frac{\sum (r_{1,i} - \bar{r}_1)(r_{2,i} - \bar{r}_2)}{\sqrt{\sum (r_{1,i} - \bar{r}_1)^2} \sqrt{\sum (r_{2,i} - \bar{r}_2)^2}}$$

$$\text{Corr}(U_{30878}, U_{823519}) =$$

$$\frac{(4 - 3.333)(3 - 3.4) + (3 - 3.333)(4 - 3.4) + (4 - 3.333)(5 - 3.4)}{\sqrt{(4 - 3.333)^2 + (3 - 3.333)^2 + (4 - 3.333)^2} \sqrt{(3 - 3.4)^2 + (4 - 3.4)^2 + (5 - 3.4)^2}}$$

$$= 0.6/1.75 = 0.34$$

Customer ID	1	5	8	17	18	28	30	44	48
30878	4	1			3	3	4	5	
124105	4								
822109	5								
823519	3		1	4		4	5		
885013	4	5							
893988	3						4	4	
1248029	3					2	4		3
1503895	4								
1842128	4						3		
2238063	3								

TABLE 14.8

SAMPLE OF RECORDS FROM THE NETFLIX PRIZE CONTEST, FOR A SUBSET OF 10 CUSTOMERS AND 9 MOVIES

# Cosine Similarity

Like correlation coefficient, except do not subtract the means

Use raw ratings instead of departures from averages

$$\begin{aligned}\text{Cos Sim}(U_{30878}, U_{823519}) &= \frac{4 \times 3 + 3 \times 4 + 4 \times 5}{\sqrt{4^2 + 3^2 + 4^2} \sqrt{3^2 + 4^2 + 5^2}} \\ &= 44/45.277 = 0.972\end{aligned}$$

Ranges from 0 (no similarity) to 1 (perfect match)

Customer ID	1	5	8	17	18	28	30	44	48
30878	4	1			3	3	4	5	
124105	4								
822109	5								
823519	3		1	4		4	5		
885013	4	5							
893988	3						4	4	
1248029	3					2	4		3
1503895	4								
1842128	4						3		
2238063	3								

TABLE 14.8

SAMPLE OF RECORDS FROM THE NETFLIX PRIZE CONTEST, FOR A SUBSET OF 10 CUSTOMERS AND 9 MOVIES

# Using the similarity info to make recommendations

- Given a new user, identify k-nearest users
- Consider all the items they rated/purchased, except for the co-rated ones
- Among these other items, what is the best one? “Best” could be
  - Most purchased
  - Highest rated
  - Most rated
- That “best” item is the recommendation for the new user



# Cold Start

- Collaborative filtering suffers from what is called a *cold start*: it cannot be used as is to create recommendations for new users or new items.
- For a user who rated a single item, the correlation coefficient between this and other users (in user-generated collaborative filtering) will have a denominator of zero and the cosine proximity will be 1 regardless of the rating.
- In a similar vein, users with just one item, and items with just one user, do not qualify as candidates for nearby neighbors

# Item-based collaborative filtering

- When the number of users is huge, user-based calculations pose an obstacle (similarity measures cannot be calculated until user shows up)
- Alternative – when a user purchases an item, focus on similar items
  1. Find co-rated (co-purchased) items (by any user)
  2. Recommend the most popular or most correlated item

# Item-based collaborative filtering

- Similarity is now computed between items, instead of users. For example, in the Netflix sample, the correlation between movie 1 (with average  $r_1 = 3.7$ ) and movie 5 (with average  $r_5 = 3$ ) is:

$$\text{Corr}(I_1, I_5) = \frac{(4 - 3.7)(1 - 3) + (4 - 3.7)(5 - 3)}{\sqrt{(4 - 3.7)^2 + (4 - 3.7)^2} \sqrt{(1 - 3)^2 + (5 - 3)^2}} = 0$$

Customer ID	Movie ID								
	1	5	8	17	18	28	30	44	48
30878	4	1			3	3	4	5	
124105	4								
822109	5								
823519	3		1	4		4	5		
885013	4	5							
893988	3						4	4	
1248029	3					2	4		3
1503895	4								
1842128	4						3		
2238063	3								

TABLE 14.8

SAMPLE OF RECORDS FROM THE NETFLIX PRIZE CONTEST, FOR A SUBSET OF 10 CUSTOMERS AND 9 MOVIES

- Thus we can compute similarity between all the movies.
- This can be done offline.
- In real time, for a user who rates a certain movie highly, we can look up the movie correlation table and recommend the movie with the highest positive correlation to the user's newly rated movie.

# Summary – Collaborative Filtering

- User-based – for a new user, find other users who share his/her preferences, recommend the highest-rated item that new user does not have.
  - User-user correlations cannot be calculated until new user appears on the scene... so it is slow if lots of users
- Item-based – for a new user considering an item, find other item that is most similar in terms of user preferences.
  - Ability to calculate item-item correlations in advance greatly speeds up the algorithm
  - The disadvantage of item-based recommendations is that there is less diversity between items (compared to users' taste), and therefore, the recommendations are often obvious.

## Association Rules

- focus entirely on frequent (popular) item combinations.
- Data rows are single transactions.
- Ignores user dimension.
- Often used in displays (what goes with what).
- Binary Data
- Two or more items

## Collaborative Filtering

- focus is on user preferences.
- Data rows are user purchases or ratings over time.
- Can capture “long tail” of user preferences
- useful for recommendations involving unusual items
- Binary as well as Ratings data
- Between pairs of items or users



# Slide Contents - References

- The contents of this presentation were sourced and assembled from
  - Data Mining for Business Analytics: Concepts, Techniques and Applications in R, by Galit Shmueli et al., Wiley India, 2018.