13 Nov 2022

Introduction to Recommendation Systems

Senthilkumar Subramanian
Ford Motor Company
GDI&A
Global Data Insight & Analytics

Disclaimer

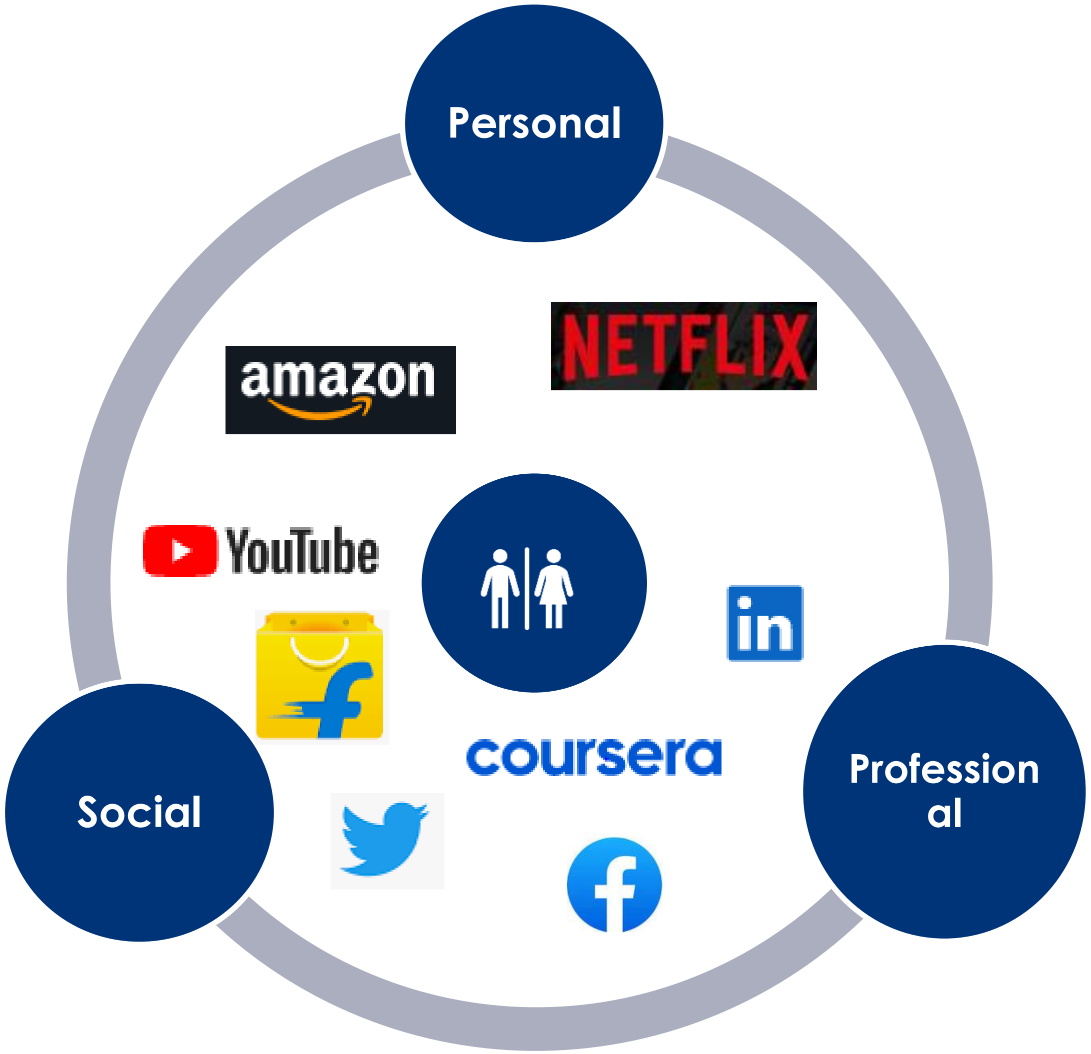
Author takes the complete responsibility

In case of any discrepancies seen in the content or point of view expressed

Contents

GDI&A
Global Data Insight & Analytics

A Slice of our Day-Today Routine



Shopping)

Entertainment

Socializing

Professional

What is Recommendation System (RS)?

- RS filters and suggests relevant content at individual level based on user's past preference which helps to improve the user current experience
- Used widely for recommending movies, articles, CPG* goods
- Effective for both service providers and users
- Helps business to upsell and cross sell through product recommendations
- E-commerce/Digital business derive maximum benefit through RS

Image Source: Shutterbox.com

CPG: Consumer Packaged Goods

Why RS got more popular in the recent decades?

- ❑ **Internet Penetration:** Massive expansion of internet across all levels helped to access digital contents at minimal to zero cost

- ❑ **E-Commerce Disruption:** There is exponential growth of E-Commerce sectors with the deeper penetration into remote towns/villages. Also e-commerce product portfolio expanded from selling fresh vegetables to delivering cars through online

- ❑ **Scarcity to Abundance:** Number of products available in an E-Commerce store or digital streaming platforms are much higher compared to Physical retail stores
  - ❑ More choice necessities better filters and recommendation engines for the business

- ❑ **Advancement in Data Science & Cloud Computing :** Lead to development of scalable and most efficient recommendation engines

GDI&A
Global Data Insight & Analytics

Personalization

Just sharing the recommendations for couple of profiles from the same household

Recommendations are personalised at the individual level even within same household

Types of Recommendation System

- **Recommendation System**
  - **Popularity based**
    - Ranked based on maximum viewers or buyers
  - **Collaborative Filtering**
    - Computed based on user ratings
  - **Content Based Filtering**
    - Computed using features like item descriptions or in case of movies using Director, Actor, Genres
  - **Hybrid Methods**
    - Ensemble of Few recommendation systems

Popularity Based Approach

- Recommends the top trending article or most selling items
- Relatively easy to compute, No customer data required
- But all customers get same suggestions, Not personalized

Recommendations for books as on 5th Nov 2022 at Amazon.com

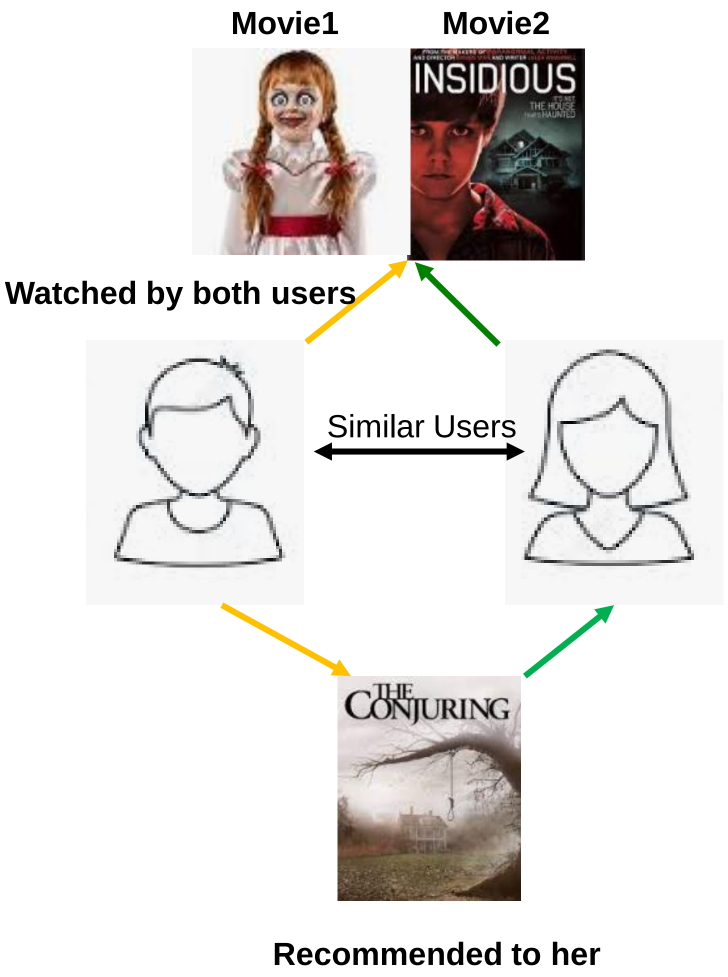
CONTENTS

Collaborative Filtering

- Computed using customer ratings, based on which personalized recommendations can be made

- Assumption with CF is that if users A and B have similar taste in a product, then A and B are likely to have similar taste in other products as well

- Similarity can be measured at user-user level or by item-item grouping



Find Similar Users / Similar movies?

User	Movie1	Movie2	Movie3	Movie4	Movie5	Movie6
Nikita	4	3	4	4	5	?
Krisha	4	3	4	4	4	5
Nikhil	3	1	1	3	2	4
Nithin	1	2	4	1	3	5
Amit	3	1	1	3	3	3

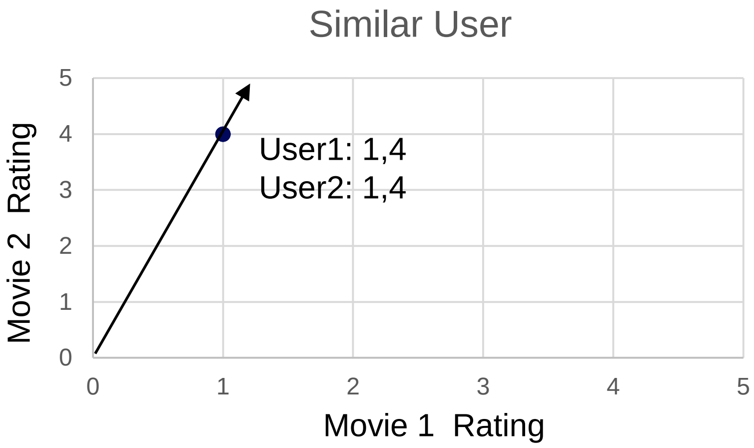
Rating 1- Low: 5 Highest

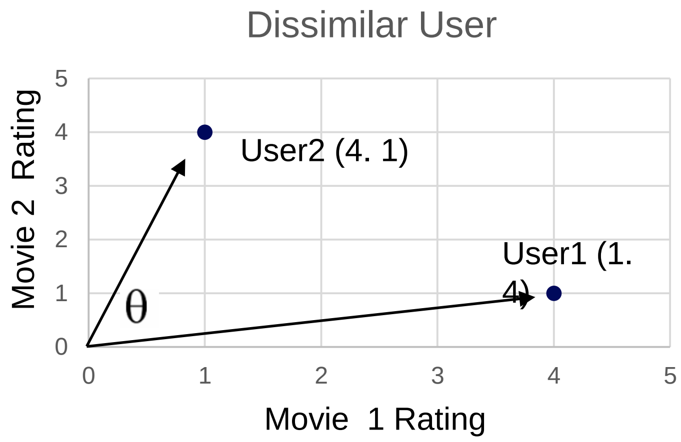
Unboxing Cosine Similarity (1/2)

	Movie1	Movie2
User1	1	4
User2	1	4

	Movie1	Movie2
User1	1	4
User3	3	5

	Movie1	Movie2
User1	1	4
User2	4	1





- ❑ Similarity b/n two users or items can be measured using the angular distance
- ❑ Cosine of the angle between the lines/vectors defines how similar two users in a scale of 0 to 1

Unboxing Cosine Similarity (2/2)

❑ **Let us compute similarity value for the given two user manually**

	Movie1	Movie2
User1	1	4
User2	1	4

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

A	User1
B	User2
i=1	Movie1
i=2	Movie2

Similarity = ((User1 Movie1*User2Movie1) + (User1Movie2*User2Movie2))/
(sqrt (User1Movie1^2+User1Movie2^2))+ (sqrt (User2Movie1^2+ User2Movie2^2))

= ((1*1) + (4*4))/(Sqrt (17) * Sqrt(17))
= 17/17
= 1

- **Similarity values can take values from 0 to 1, In this case 1 implies both the users have similar interest**

Collaborative Filtering – Step by Step Approach

1. Import raw data with relevant fields
2. Create user-item table
3. Standardize the Data
4. Compute Cosine Similarity
5. Based on Similarity, find the closest neighbors
6. Predict the Ratings for given user / items
7. Recommend items

Collaborative Filtering – Step by Step Approach (1/2)

In the given dataset, let us find similar users and predict the ratings for the given user

1.Import raw data with relevant field

User	MovieId	Rating
Nikita	Movie1	4
Nikita	Movie2	3
Nikita	Movie3	4
Nikita	Movie4	4
Nikita	Movie5	5
Krisha	Movie1	4
Krisha	Movie2	3
Krisha	Movie3	4
Krisha	Movie4	4
Krisha	Movie5	4
Krisha	Movie6	5
Nikhil	Movie1	3
Nikhil	Movie2	1
Nikhil	Movie3	1
Nikhil	Movie4	3
Nikhil	Movie5	2

2.Create a Cross Tab of User-Movie Table

User	Movie1	Movie2	Movie3	Movie4	Movie5	Movie6
Nikita	4	3	4	4	5	?
Krisha	4	3	4	4	4	5
Nikhil	3	1	1	3	2	4
Nithin	1	2	4	1	3	5
Amit	3	1	1	3	3	3

3.Standardise the Data

User	Movie1	Movie2	Movie3	Movie4	Movie5
Nikita	0.00	-0.50	0.00	0.00	0.50
Krisha	0.20	-0.80	0.20	0.20	0.20
Nikhil	0.50	-0.50	-0.50	0.50	0.00
Nithin	-0.40	-0.07	0.60	-0.40	0.27
Amit	0.40	-0.60	-0.60	0.40	0.40

Standardized value= (Given value- Row Mean)/Range

GDI&A
Global Data Insight & Analytics

Collaborative Filtering – Step by Step Approach (2/2)

In the given dataset, we will find similar users and predict the ratings for the given user

4.User-User Similarity

Simillarity	Nikita	Krisha	Nikhil	Nithin	Amit
Nikita	1	0.791	0.354	0.271	0.645
Krisha	0.791	1	0.559	0.086	0.612
Nikhil	0.354	0.559	1.000	-0.767	0.913
Nithin	0.271	0.086	-0.767	1.000	-0.560
Amit	0.645	0.612	0.913	-0.560	1.000

5.Find closest neighbors for Nikita

Similar to Nikita

User	Rank
Krisha	1
Amit	2
Nikhil	3

6.Predicting the Ratings

Predicted Ratings for Movie 6 for Nikita	Ratings
Based on Top 2 users	4
Based on Top 3 users	4

- **Recommendation decisions are made if the rating crosses a certain threshold**
- **In this case, Movie 6 can be recommended to Nikita since the predicted rating observed to be on the higher scale**
- **Practically in many cases, Item-Item CF works much better than User-User CF since User preferences may change over time, however Item features does not change over time**

Evaluating Recommendations Systems (1/2)

K Fold Cross Validation

- ❑ Create K randomly assigned training and test sets. Develop RS using individual training sets and apply it to test set and measure the accuracy
- ❑ Take the average of accuracy score to see how well the recommendation system is learning. This method is beneficial to prevent model from overfitting

MAE (Mean Absolute Error):

- ❑ It is the absolute average of Actual – Predicted Rating. Lower the MAE value, more accurate is the prediction.

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

y_i = prediction

x_i = true value

n = total number of data points

Evaluating Recommendations Systems (2/2)

Root Mean Square Deviation/Error (RMSD)

- Like MAE but penalize more when the prediction is very far from the true value and penalize lesser for when the prediction is closer to the true value

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \hat{x}_i\right)^2}{N}}$$

RMSD = root-mean-square deviation
i = variable i
N = number of non-missing data points
x_i = actual observations time series
$\hat{x}_i$ = estimated time series

Limitations of Collaborative Filtering

- ❑ **Cold Start problem**
  - ❑ We can not compute CF for the Users or items with no historical ratings
- ❑ **Data Sparsity**
  - ❑ Sparse availability of ratings for certain users or items makes the predictions less accurate
- ❑ **Scalability**
  - ❑ If the number of items or users are massive then it becomes computationally intensive
- ❑ **Dynamic updates**

Contents

GDI&A
Global Data Insight & Analytics

Content Based Filtering

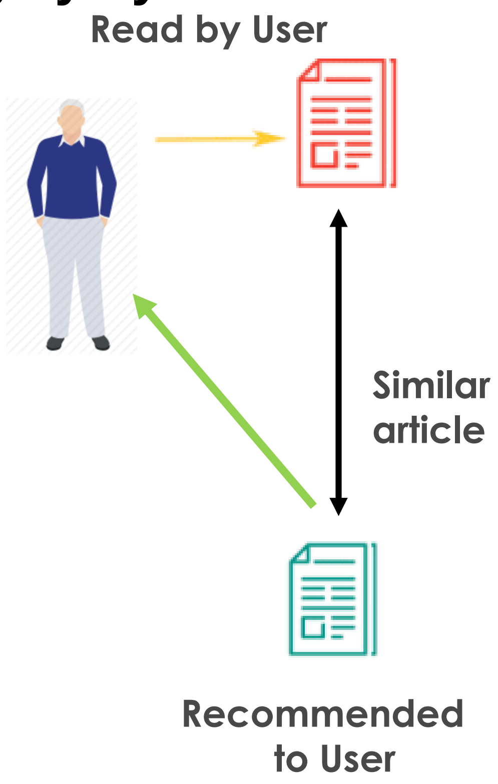
- Content based recommendations are made based on the item profiles using features extracted from the content of the items the user has evaluated in the past

- Recommend items to customer x similar to previous items rated highly by x

- Examples:

  - Recommend Movies from the same actor, genre, casts

  - Recommend New articles with similar content, same author



Content Based Filtering

- ❑ **From the features, create item profile for example:**
  - ❑ **Movies: Author, Title Cast, Genre | Articles: Domain, Publishers**
- ❑ **Ideally item profile can be created as Vector using real counts or Boolean**
- ❑ **Item profile can be created by using words with high TF-IDF Score**
- ❑ **How to create important features for the profile**
  - ❑ **TF-IDF Score:** $W_{ij} = TF_{ij} * IDF_i$
  - ❑ **TF-IDF:** (Term Frequency, Inverse Doc Frequency) **– Used for information retrieval**
    - ❑ **Term Frequency => Total Frequency of given word in the article/total number of words in the article**
    - ❑ **IDF: Log (Total number of articles in the given corpus/Number of articles containing given word**

Content based Filtering – Step by Step Approach

1. **Import raw data**
   (Can be Text count or Boolean)

2. **Create Item Profile**
   e.g: For movies: Genre, Cast, Director

3. **Convert Word to Vec (TF-IDF)**

4. **Compute Cosine Similarity**

5. **Based on Similarity, find the closest neighbors**

6. **Predict the Ratings for given user / items**

7. **Recommend items**

Computing TF-IDF

- Let us compute TF-IDF for the given example
  - TF-IDF Score: $W_{ij} = TF_{ij} * IDF_i$
    - Term Frequency => Total Frequency of given word in the article/total number of words in the article
    - IDF: Log (Total number of articles in the given corpus/Number of articles containing given word

Step 1: Raw Data

Sentence1	Best	Actress	
Sentence2	Best	Actor	
Sentence3	Best	Actor	Actress

Step 2: Creating Feature List

Article	Best	Actress	Actor
Sentence1	1	1	0
Sentence2	1	0	1
Sentence3	1	1	1

Step 3a: Computing TF

Article	Best	Actress	Actor
Sentence1	1/2	1/2	0
Sentence2	1/2	0	1/2
Sentence3	1/3	1/3	1/3

Step 3b: Computing IDF

	Best	Actress	Actor
IDF	0	0.176	0.176

Step 3c: TF*IDF Scores

	Best	Actress	Actor
Sentence1	0	0.088	0.000
Sentence2	0	0.000	0.088
Sentence3	0	0.059	0.059

Once words are converted to vectors, output of TF-IDF scores can be used to compute similarity score

Pros & Cons of Content Based Filtering

- **Content Based can be deployed even if there is no explicit rating provided by users. Very effective in finding similar articles, recommending blogs, posts.**
- **Cons:**
  - **Cold Start problem:** We can not compute CB for the Users or items with no historical ratings
  - **Item Description:** Rich item metadata is required for creating feature list/item profile
  - **Overspecialization:** Users are restricted to get recommendations similar to items already defined in their profiles

Hybrid Approach

- Most of the product companies leverage multiple Recommendation systems in certain combinations to arrive at the final prediction
- In addition to Popularity based, Collaborative Filtering, Content Based, Product companies also leverages clustering, modelling approaches, association rule mining and customer demographic info to arrive at a hybrid approach to recommend products to increase the scalability and accuracy of the recommendations

Practice/Assignment

- Using Collaborative Filtering approach create Item-Item Similarity measure for the given data.
- Find top 2 movies similar to Movie 5 from the given list

User	Krisha	Nikhil	Nithin	Amit	Nikita
Movie1	4	3	1	3	4
Movie2	3	1	2	1	3
Movie3	4	1	4	1	4
Movie4	4	3	1	3	4
Movie5	4	2	3	3	5


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
спасибо
bedankt
dziękuję
go raibh maith agat
mochchakkeram
sukriya
kop khun krap
terima kasih
감사합니다
ありがとう
ευχαριστώ
arigatô
takk
dakujem
grazie
tapadh leat
мерси
рахмат
Баярлалаа
dank je
hvala
mauruuru
sagolun
enkosi
trugarez
хвала
asante
manana
obrigada
shukriya
dhanyavadagalu
diolch
rahmet
tanemirt
xièxie
najis tuke
kam sah hamnida
mesi
sobodi
dëkuji
chnorakaloutioun
gratias ago
gràcies
sulpay
mersi
barka
welalin
tack
spas
faafetai lava
vinaka
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
kiitos
dankie
dhanyavad
nandri
gracie
köszönöm
akun
dankon
djiere dieuf
tau
mamnun
murakoze
chokrane
tenki