

Recommender Systems

Mining Massive Datasets

Prof. Carlos Castillo — <https://chato.cl/teach>



Universitat
Pompeu Fabra
Barcelona

Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (Section 18.5) – slides by Lijun Zhang
- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al. (Chapter 9) – slides A, B



Recommendation algorithms
connect people, artists, and
creations, and to some
extent shape our taste, the
careers of those artists, and
the form of their creations

Recommender systems

- Product recommendation is perhaps one of the best known use cases:
 - Given data from user buying behaviors, profiles, interests, browsing behavior, buying behavior, and ratings about various items
 - Leverage such data to make **recommendations** to customers about possible buying interests

Recommender systems (general)

- In general, the idea is:
 - Given data from user interests, including profiles, browsing behavior, item interaction behavior, ratings about various items
 - Leverage such data to make **recommendations** to users about further **interesting** items

NETFLIX



amazon

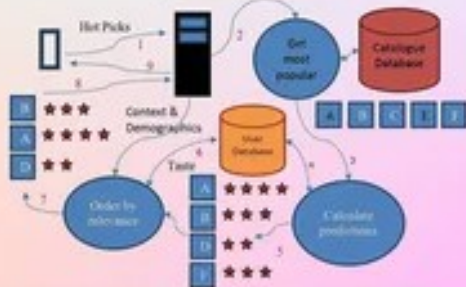


STEAM®

NETFLIX FILM RECOMMENDATION ALGORITHM



@TheInsaneApp



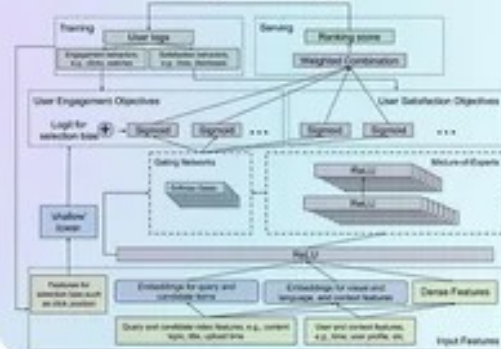
SPOTIFY RECOMMENDATION ALGORITHM



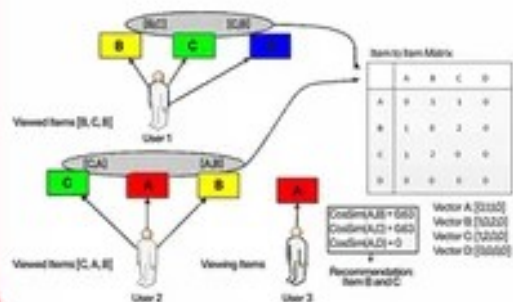
YOUTUBE VIDEO RECOMMENDATION ALGORITHM



@TheInsaneApp



AMAZON PRODUCT RECOMMENDATION ALGORITHM



Large scale engines for recommendation:

- are composed of **multiple layers**,
- use **online** and **offline** (batch) models,
- include complex **data pipelines** to move **behavioral** and **content** signals around.

Utility matrix

- Matrix D of size n (users) \times d (items)
 - The utility value for a user-item pair (D_{ij}) describe some relationship between user i and item j
 - Typically, a small subset of the utility values are known

Example utility matrices

	Dune	Soul	Nomadland	Avengers	RRR	Encanto
U_1	1			5		2
U_2		5			4	
U_3	5	3		1		
U_4			3			4
U_5				3	5	
U_6	5		4			

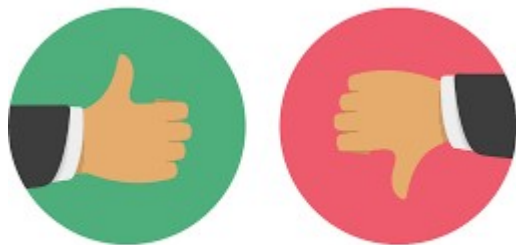
(a) Ratings-based

	Dune	Soul	Nomadland	Avengers	RRR	Encanto
U_1	1			1		1
U_2		1			1	
U_3	1	1		1		
U_4			1			1
U_5				1	1	
U_6	1		1			

(b) Positive preference, e.g., “like”

Types of utility

- **Explicit:** we ask users to rate items



- **Implicit:** we take watching/consuming/buying behavior as a positive signal, skip/hide as negative

Sources for a recommendation

- Content-based recommendation
 - Users and items are associated with features
 - Features are matched to infer interest
- Interaction-based recommendations
 - Leverage user preferences in the form of ratings or other behavior
 - Recommend through similarity or latent factors

THE COLD START PROBLEM

New items have no ratings

and

New users have no history

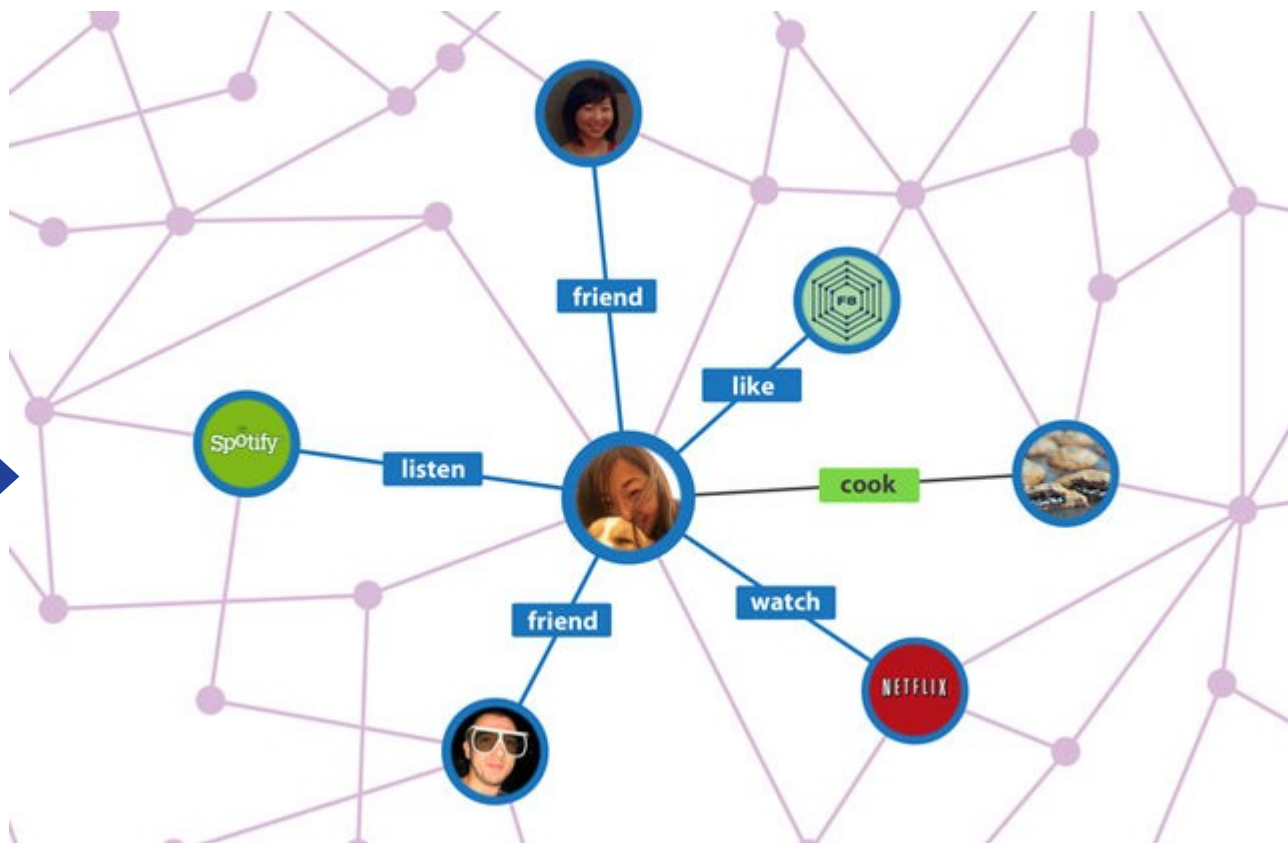


Photo: Torque News

THE COLD START PROBLEM

Solution 1. "Side information"

 Login with Facebook



Choose some artists you like.

Choose at least 3. We'll make some special playlists for you.



Taylor Swift



Ed Sheeran



Drake



Calvin Harris



Kendrick Lamar



MORE
FOR YOU



The Chainsmokers



Lorde



ODESZA



MORE
ELECTRONICA

THE COLD START PROBLEM

Solution 2. “On-boarding” users

Touch the genres you like



Alternative/Indie



Blues



Christian/Gospel



Classical

SKIP THE QUIZ

NEXT

Content-based recommendations

General idea of content-based recommendations

- Movies: recommend other movies with **same** director, actor, genre, as viewed ones
- Products: recommend other products in **same** category, brand, color, as purchased ones

Creating a recommendation

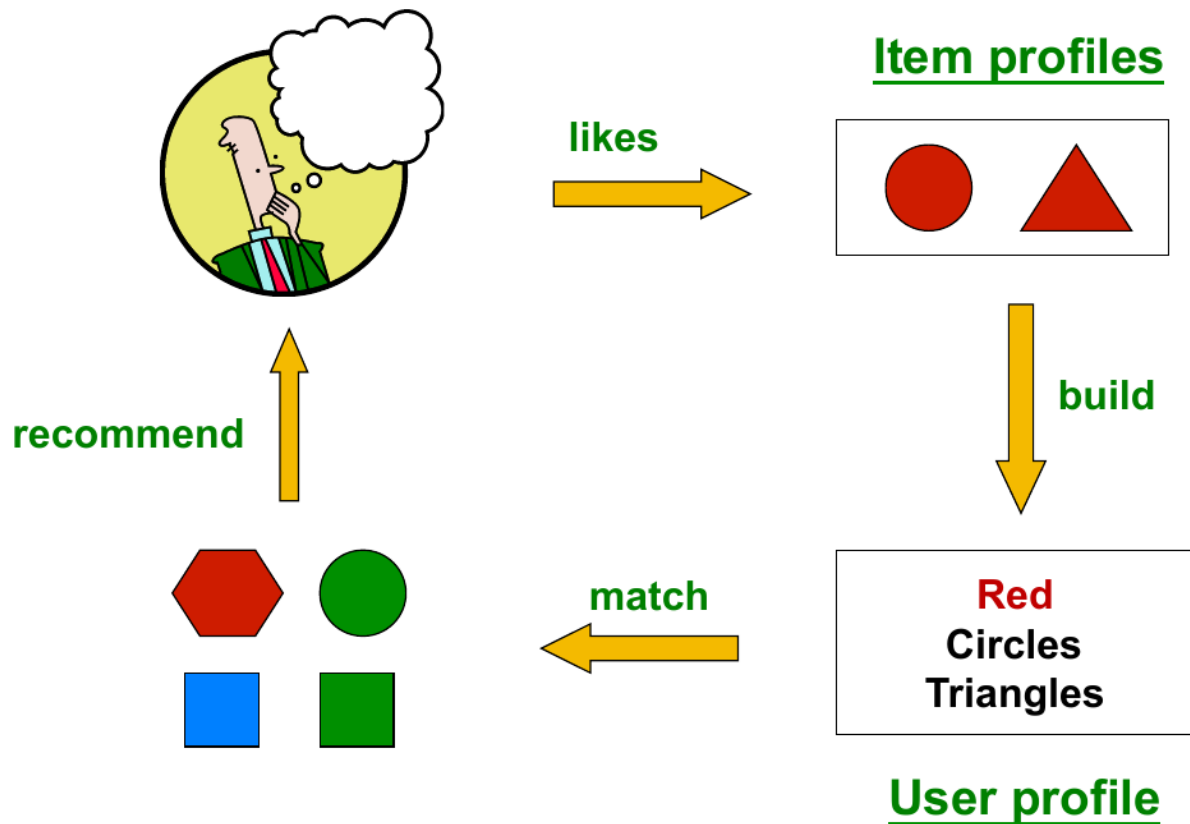
- User is associated with some documents that describe his/her interests
 - Specified demographic profile
 - Specified interests at registration time
 - Descriptions of the items bought
- Items are also associated with semi-structured descriptions



JBL GO lleva el sonido de calidad JBL a todas partes. GO es su solución de altavoz todo en uno y reproduce música en tiempo real vía Bluetooth desde smartphones y tabletas, gracias a su batería recargable. También cuenta con un práctico manos libres.

Potencia	3 W
Respuesta de Frecuencia	180Hz – 20 kHz
Tipo de altavoz	Portátil
Amplificador de sonido	Integrado

Creating a recommendation (cont.)



Possible recommendation methods

- **If no utility matrix is available**
 - k-nearest neighbor approach
 - Find the top-k items that are closest to the user (when items and users can be represented in the same space, e.g., dating apps)
 - The cosine similarity with tf-idf can be used
- **If a utility matrix is available**
 - Classification-based approach: training documents are those for which the user has specified utility, labels are utility values
 - Regression-based approach in the case of ratings
- Limitations: depends on the quality of the features

Example: regression-based approach for content-based recommendation

Movie	Adventure	Action	Science-Fiction	Drama	Crime	Thriller		User 1	User 2
Star Wars IV	1	1	1	0	0	0		1	-1
Saving Private Ryan	0	0	0	1	0	0			
American Beauty	0	0	0	1	0	0			
City of Gold	0	0	0	1	1	0		-1	1
Interstellar	0	0	1	1	0	0		1	
The Matrix	1	1	1	0	0	1			1

...

We would do two regressions: one for the ratings of user 1 and another for user 2.
(We can also do this for groups of users, e.g., by city and age)

How many rated movies would we need, as a minimum, to be able to do this?

Exercise

Content-based recommender based on regression



- Database of ~ 100 electric scooters, of which 12 have been rated on a scale 1-5
- We have done linear regression on:
price [\$], battery capacity [Wh], range [km]
- Which would be your top-3 recommended scooter among the remaining ones?

Spreadsheet links: <https://upfbarcelona.padlet.org/chato/hogch321o6pws1fd>



Pros and Cons of content-based recommendations

- Pros:
 - No cold-start problem if no utility needed
 - Able to recommend to users with very particular tastes
 - Able to recommend new and obscure items
 - Able to provide explanations that are easily understandable

Pros and Cons of content-based recommendations

- Cons:
 - Finding the correct features might be hard
 - Recommending for new users still challenging if user features are different from item features
 - Overspecialization/"bubble": might reinforce user interests
 - Does not exploit ratings of other users!

Summary

Things to remember

- Content-based recommendations
- Regression-based method

Exercises for TT16-TT18

- Mining of Massive Datasets 2nd edition (2014) by Leskovec et al. Note that some exercises cover advanced concepts:
 - Exercises 9.2.8
 - Exercises 9.3.4
 - Exercises 9.4.6