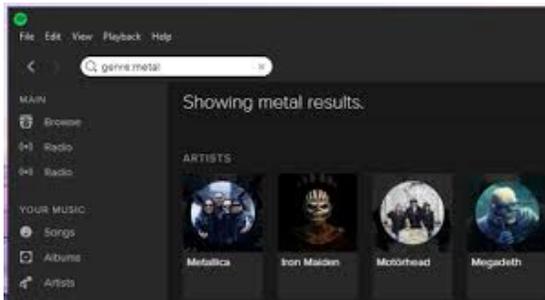


Movie Recommendation – Collaborative Filtering

CMPT 733

Spring 2017

Apala Guha



Users

Preferences

- rating, vote, follow
- review
- click, time spent

Products

- books, movies
- friends, people to follow
- playlists, service bundles

Recommendations

- ranking
- related to context/mood
- sequence



	Alice	Bob	Carol	Dave
Shanghai Triad	5	?	4	?
Usual suspects	?	?	1	1
In Love and War	2	3	2	1
Anna Karenina	2	5	?	?
Incognito	1	?	2	3



Recommender



- demographics



- index no.



- aggregate stats

Non-personalized recommender

Non-personalized

- Nuances not known or required
- Initial recommendation/profile-building stage
- The same recommendation fits many

The screenshot shows the TripAdvisor website interface for Berlin. At the top, there's a navigation bar with links for Home, Hotels, Flights, Restaurants, Vacation Rentals, Trip Ideas, and Write a Review. A search bar is present with the text 'City, hotel name, etc'. Below the navigation, there's a sidebar with various filters and categories like Berlin Tourism, Berlin Hotels, Vacation Rentals, Flights to Berlin, Berlin Deals, Ski Germany, More on Berlin, Restaurants, Things to Do, Travel Forum, Travel Guide, Photos, Videos, Map, Berlin Deals, Discount Hotels, Hotel & Air, and All Travel Cities. The main content area is titled 'Things to Do in Berlin' and features a 'Traveler recommended attractions' section. The first attraction listed is the 'Neues Museum', which is ranked #1 of 239 attractions in Berlin. It has 45 reviews and a 5-star rating. The description mentions that it's the Egyptian Collection, which was moved into the Neues Museum 4 weeks ago, and it was previously closed for 10 years. The attraction type is 'Museum', and it's located 0.8 miles from the city center. There are also quotes from travelers: 'History at its best!' (Feb 16, 2011) and 'Worth a visit' (Feb 22, 2011). On the right side, there's a section for 'Advice from real travelers' with links to Neighborhoods, First-time visitors, Recommended Reading, Public transportation, and Walking tour. At the bottom, there's a 'Free Newsletter' button and a footer with links for Cruise Cities, Destinations, Family Vacation Cities, and See all sites.

TRENDING

- **Kirby Misperton, North Yorkshire:** Councillors Approve Fracking Tests After 5-Year Pause in UK
- **Solar Impulse 2:** Solar-Powered Plane on Journey Around the World Lands in Dayton, Ohio
- **Dartford Crossing:** Problem With Tunnel's Safety System Causes Major Traffic Delays in Kent
- **Paris Hilton:** Socialite Reportedly Has Wardrobe Malfunction in London
- **World Turtle Day:** May 23 Marks Annual Observance to Raise Awareness for Turtles and Tortoises
- **Mansfield, Nottinghamshire:** Army Clears



- demographics
- Per-user model



- index no.
- Stable feature vector



- aggregate stats
- user-item stats

Content-based filtering

CBF

- Stable item description e.g. publisher summary
- Item feature vectors e.g. TF-IDF
- Learn per-user models to map item feature vectors => user ratings
- Assumes user preferences remain stable and are not highly nuanced
- Product attributes suitably captured by summary/review

NEWS Search News
Advanced news search

[U.S. edition](#)

Top Stories

[Russia »](#)
Russian spy ring suspect jumps bail in Cyprus
The Guardian - 30 minutes ago
The Russian espionage drama intensified tonight as one of the suspects in the alleged "deep cover" spy ring failed to answer bail in Cyprus.
[Alleged Russian spy ring members led typical American lives](#) Los Angeles Times
[American beauty](#) BBC News
[AFP - New York Times - Reuters - BusinessWeek](#)
[all 5,461 news articles »](#)

[Elena Kagan »](#)
Day 3: Leahy predicts confirmation
The Associated Press - [David Espo](#) - 45 minutes ago
WASHINGTON - Supreme Court nominee Elena Kagan neared the end of a grueling turn in the Senate Judiciary Committee witness chair Wednesday, and the senator presiding over the proceedings predicted her confirmation.
[ABC News - MiamiHerald.com - CBS News - Christian Science Monitor](#)
[all 8,691 news articles »](#)

[Afghanistan »](#)
Record casualties in June as Petraeus takes helm in Afghanistan
Los Angeles Times - [Laura King](#) - 25 minutes ago
At least 101 Western troops died in June; 58 were US service members. Although buried bombs pose a significant hazard, other threats are growing as insurgents become bolder in their attacks.
[FOXNews - USA Today - msnbc.com - The Associated Press](#)
[all 4,361 news articles »](#)

News for you - Edit personalization

[BP »](#)
Waves From Storm Hinder Spill Effort
New York Times - [Henry Fountain](#) - 2 hours ago
Oil cleanup workers were evacuated from the beach in Port Fourchon, La., on Tuesday because of high winds and lightning. By HENRY FOUNTAIN The first major storm of the season in the Gulf of Mexico continued to disrupt oil spill cleanup and containment ...
[all 7,804 news articles »](#)

[Kristen Stewart »](#)
'Eclipse' Fans Love Movie's Tent Scene, Faster Pace
MTV.com - [Jocelyn Vena](#), [James Lacsina](#) - 2 hours ago

Recent

[Vultures circle BP over fears it may](#)
The Guardian - 18 minutes ago

[Wal-Mart worker, fired for legal pot s](#)
[sues](#)
Reuters - [Jonathan Stempel](#), [Andre](#) - 15 minutes ago

[Doc Rivers coming back](#)
USA Today - 12 minutes ago

2010 FIFA World Cup South Africa

[FIFA.com - Schedule - Standings - T](#)

Round of 16

 Spain	1 : 0	
 Paraguay	0 (5) : (3) 0	
 Brazil	3 : 0	

Upcoming matches

Quarter-finals

Jul 2 7:00 AM (Pacific Time) on ESF
 Netherlands vs.  Brazil

Jul 2 11:30 AM (Pacific Time) on ESF
 Uruguay vs.  Ghana

San Francisco Bay Area - Edit

57°F ☀️ Thu ☁️ 75°F | 54°F Fri ☁️ 75°F | 57°F 74°F |

[San Francisco »](#)
[Man charged in S.F. pride event sho](#)
[longer believed to be the gunman](#)
San Jose Mercury News - 3 hours ago
[all 720 articles »](#)

[San Jose International Airport »](#)



- demographics
- Per-user model
- User attributes unknown



- index no.
- Stable feature vector
- Item attributes unknown

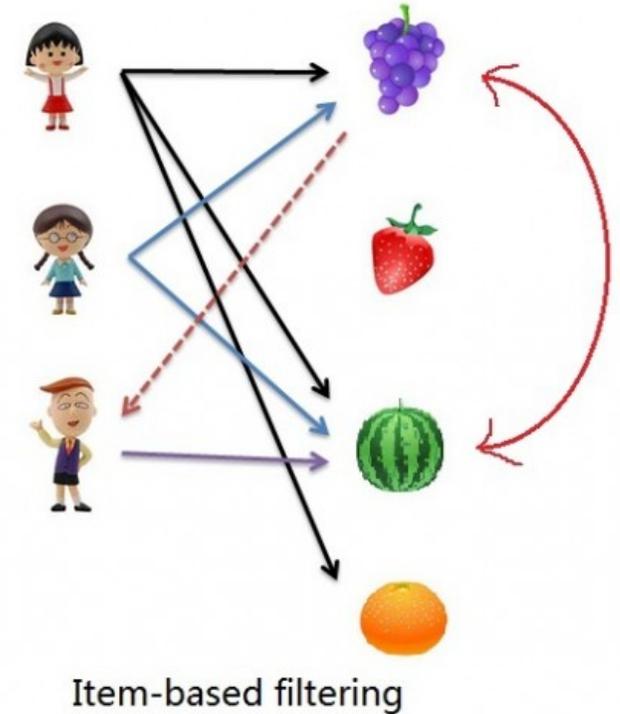
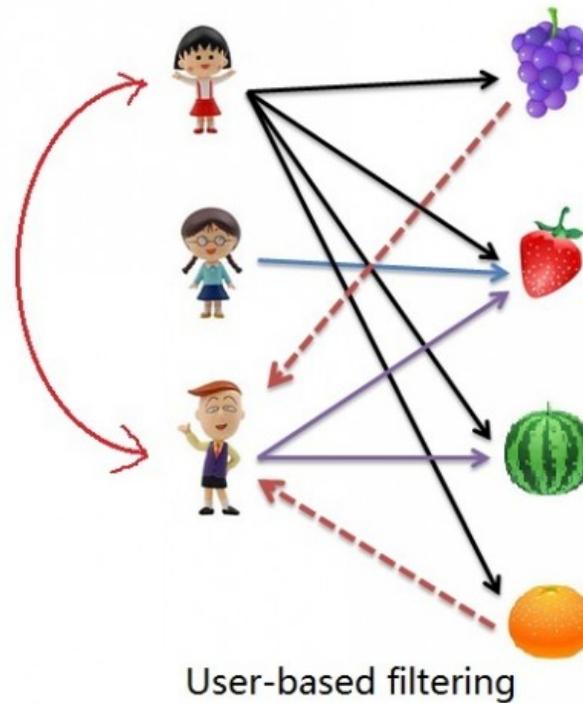


- aggregate stats
- user-item stats

Nearest-neighbor Collaborative filtering

Nearest-neighbor CF

- User-user CF: find the users most similar to this user and how these similar users rated the item
- Item-item CF: find the items most similar to this item and how this user rated these similar items
- No assumed attributes – appropriate when it is difficult to manually enumerate attributes of items



User-user CF

- We want to find $r_{u,i}$
- similarity between two users = $w_{u_1,u_2} = \text{similarity}_i(R_{u_1,i}, R_{u_2,i})$
 - R is the vector of items that both users rated
 - Pearson/rank/Jaccard/cosine
- Discard users who are not similar – their opinion is not important
- $r_{u,i} = \text{sum}_{u'}(r_{u',i} * w_{u,u'}) / \text{sum}_{u'}(w_{u,u'})$
- Correct for user rating scale: compute deviation from user mean
 $r_{u,i} = r_u + \text{sum}_{u'}(d_{u',i} * w_{u,u'}) / \text{sum}_{u'}(w_{u,u'})$

Item-item CF

- We want to find $r_{u,i}$
- similarity between two items = $w_{i_1,i_2} = \text{similarity}_u(R_{u,i_1}, R_{u,i_2})$
 - R is the vector of ratings by users that rated both items
 - Pearson/rank/Jaccard/cosine
- Discard items that are not similar
- $r_{u,i} = \text{sum}_{i'}(r_{u,i'} * w_{i,i'}) / \text{sum}_{i'}(w_{i,i'})$
- Correct for user rating scale: compute deviation from user mean
 $r_{u,i} = r_u + \text{sum}_{i'}(d_{u,i'} * w_{i,i'}) / \text{sum}_{i'}(w_{i,i'})$
- Which one is better and why?

Issues

- Typically user set has much higher cardinality than item set
- Items on average have more ratings than users
 - Therefore item similarity is more meaningful
 - Also there are fewer pairs to calculate



- demographics
- Per-user model
- User attributes unknown
- Latent user attributes



- index no.
- Stable feature vector
- Item attributes unknown
- Latent item attributes



- aggregate stats
- user-item stats

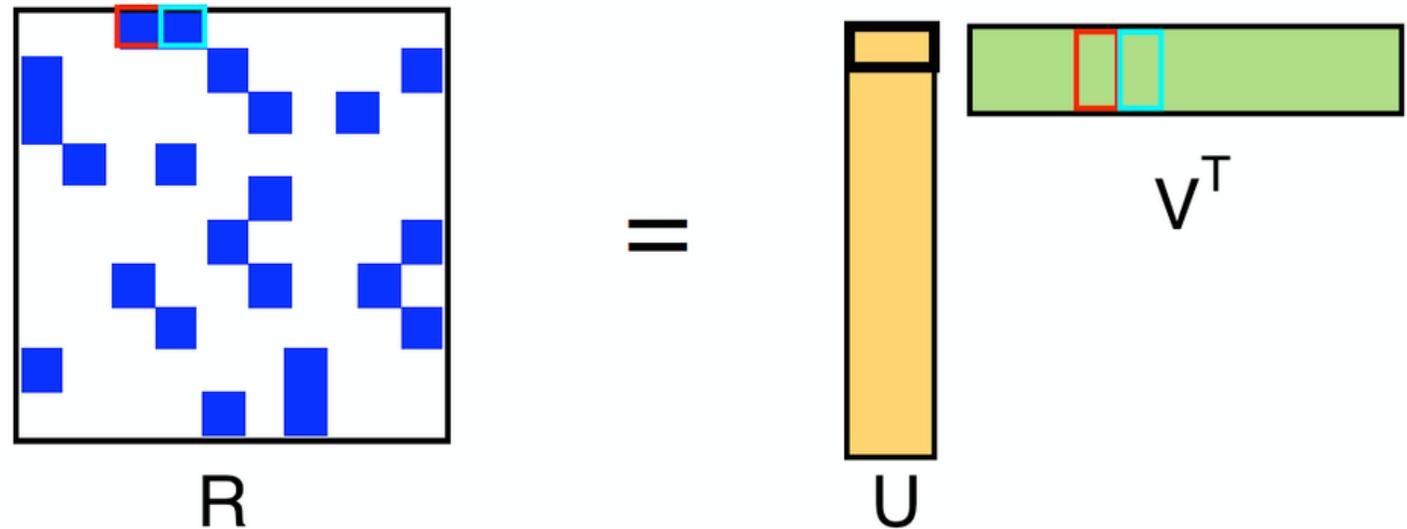
Matrix Factorization-based Collaborative filtering

	Alice	Bob	Carol	Dave
Shanghai Triad	5	?	4	?
Usual suspects	?	?	1	1
In Love and War	2	3	2	1
Anna Karenina	2	5	?	?
Incognito	1	?	2	3

$$M_{U \times I} = X_{U \times K} * Y^T_{I \times K}$$

Matrix factorization-based CF

- $K = \text{\#latent attributes}$
- Once factorized, any user-item pair can be predicted by multiplying the user vector with the item vector
- Similar to regression learning, known matrix cells are used as examples to learn the latent attributes
- However there are two sets of attributes to learn (X and Y) – learning starts with initial guesses for both X and Y, current values of X are used to drive each step in learning Y values and vice-versa => known as alternating least squares (ALS) method
- Why use it?



Matrix-factor CF

- Computationally efficient
- Users may not be entirely similar to each other, so also for items – some users may be similar on some aspects but differ on others, some products may appear similar to some users but different to others
- Dimensionality reduction for high-dimension problems e.g. in Instagram the user set and item set have the same cardinality

Assignment

- Use PySpark ML recommendation module for building ALS models
- Use PySpark SQL grouping and aggregation API for item-item CF