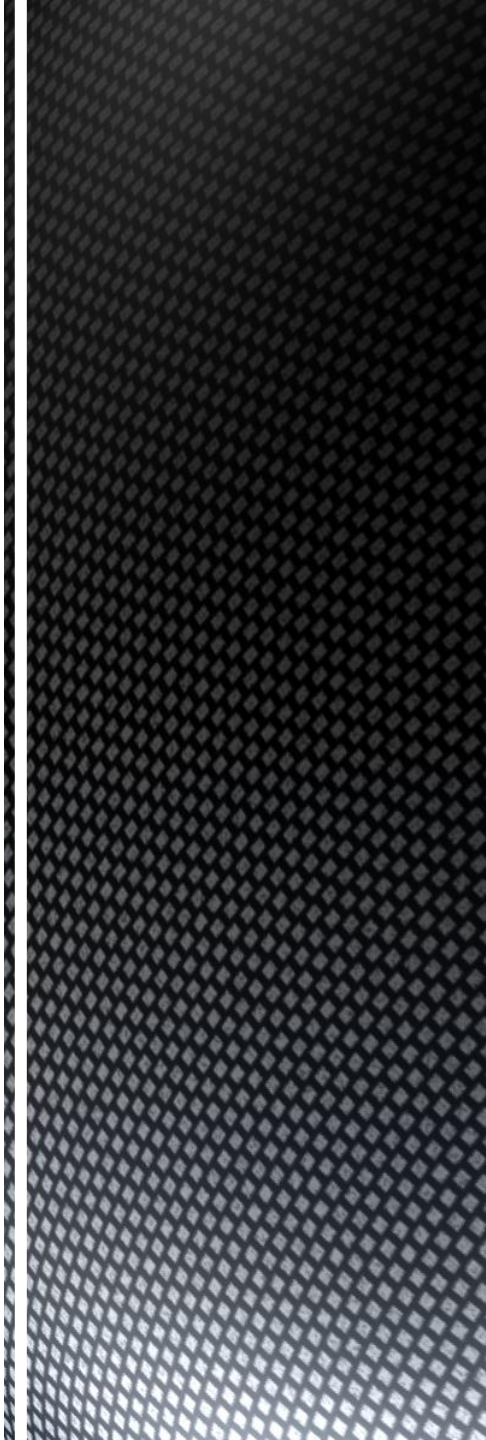
Movie Recommender Systems

Sistemi za preporuku filmova

Andrijana Marjanović 1043/2018



Sistemi za preporuku

Prikupljanje podataka o korisnicima usluga

Analiza podataka

Sprovođenje preporuka i reklamiranje

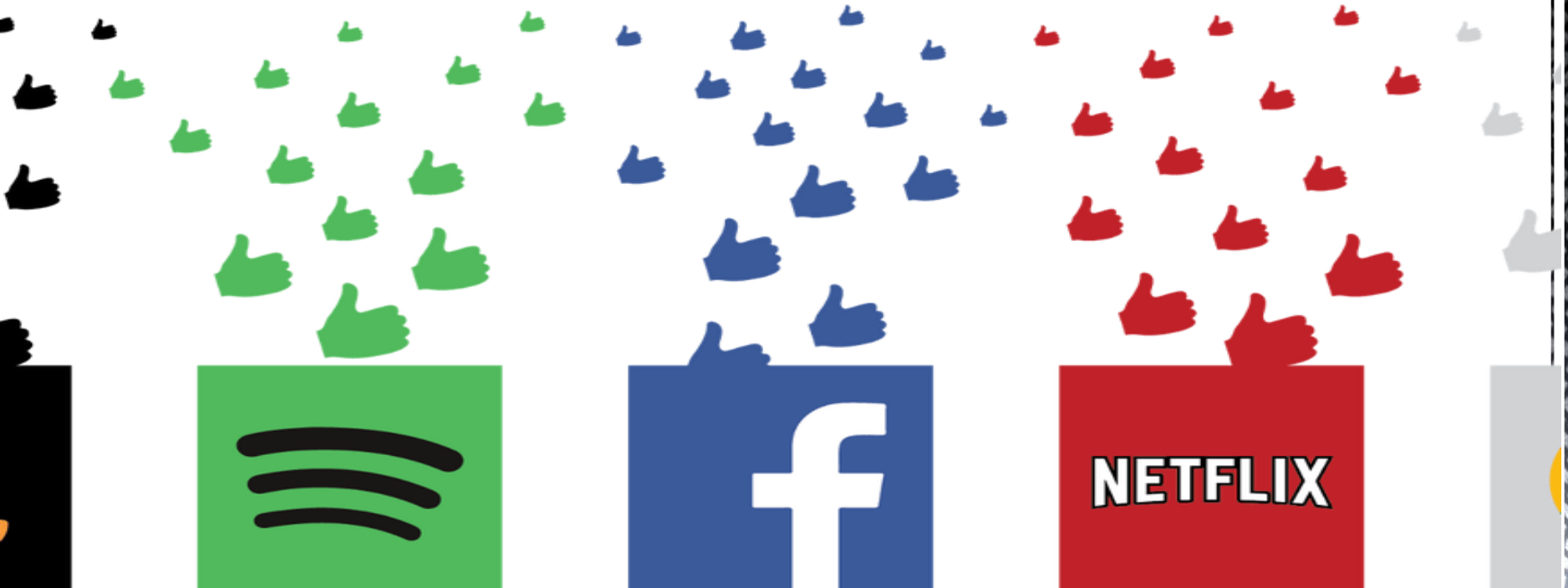
Ostvarivanje većeg prihoda

Sistemi za preporuku predstavljaju jedan od istaknutih primjera primjene mašinskog učenja u savremenom dobu.

Primjena

Sistemi za preporuku odlučuju:

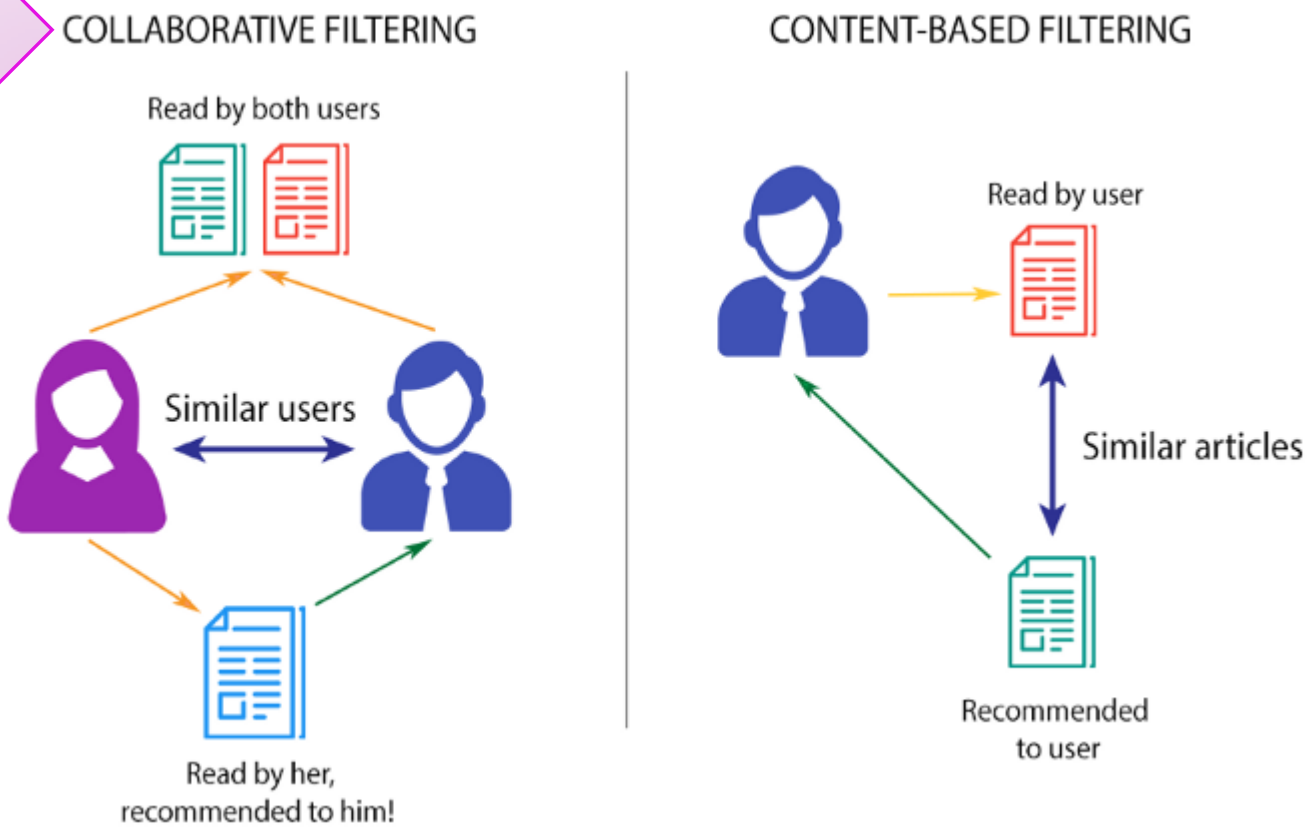
- Šta se prikazuje u novostima koje nam predstavlja **Facebook**
- U kom reposlijedu se proizvodi pojavljuju na sajtu elektronske platforme **Amazon**
- o preporukama videa na **YouTube**-u
- koji video treba da bude preporučen korisniku usloga kompanije **Netflix**...



Različiti pristupi

- Jednostavni sistemi za davanje preporuka – *Simple Recommender*
- Sistemi zasnovani na sadržaju – *Content-based Filtering*
- Sistemi zasnovani na kolaborativnom filtriranju – *Collaborative Filtering*

State-of-the-art



Jednostavni sistemi

Različiti kriterijumi za izbor filmova koji su potencijalni da se dopadnu korisniku

Drugačiji skup filmova za preporuku, što predstavlja nedostatak ovog pristupa

Prikaz 10 filmova koji su dobili najveci broj ocjena

```
movies_df.sort_values(by='count_votes', ascending=False).head(10)
```

	movieId	title	genres
2339	3097	Shop Around the Corner, The (1940)	Comedy|Drama|Romance
6518	53972	Live Free or Die Hard (2007)	Action|A
9175	149011	He Never Died (2015)	Comedy
7118	71033	Secret in Their Eyes, The (El secreto de sus o...	Crime|Dr
1801	2401	Pale Rider (1985)	Western
1678	2259	Blame It on Rio (1984)	Comedy
7093	70305	Race to Witch Mountain (2009)	Adventu Fi|Thrille
1578	2117	1984 (Nineteen Eighty-Four) (1984)	Drama|S
1445	1968	Breakfast Club, The (1985)	Comedy
1409	1929	Grand Hotel (1932)	Drama|R

Prikaz 10 filmova sa najvisom prosjecnom ocjenom

```
movies_df.sort_values(by='average_votes', ascending=False).head(10)
```

	movieId	title	genres
8658	120919	Man on High Heels (2014)	Action|Comedy
9414	165529	Flowers for Algernon (2000)	Drama
294	336	Walking Dead, The (1995)	Drama|War
8686	122892	Avengers: Age of Ultron (2015)	Action|Adventure|Sci-Fi
9407	165101	Inferno (2016)	Mystery|Thriller
8698	122924	X-Men: Apocalypse (2016)	Action|Adventure|Fantasy|Sci-Fi
290	332	Village of the Damned (1995)	Horror|Sci-Fi
483	551	Nightmare Before Christmas, The (1993)	Animation|Children|Fantasy|Musical
260	300	Quiz Show (1994)	Drama
9147	147376	Doctor Who: A Christmas Carol (2010)	Sci-Fi

Jednostavni sistemi

- Najpopularniji filmovi među velikim brojem korisnika imaju veliku vjerovatnoću da se dopadnu i većem dijelu prosječne publike
- Pružaju generalizovane preporuke
- Ne uzima se u obzir različitost korisnika koja potiče od istorije ocjenjivanja prethodnog pogledanih filmova

Izračunavanje težinske ocjene (*weighted rating):*

- Izvor *TMDB*
- Postavljanje praga za broj potrebnih dobijenih ocjena da bi film mogao ući u uži skup iz kog se biraju filmovi za preporuku
- Film A koji je dobio samo jednu ocjenu 5 – film B koji je dobio više ocjena 5

Sistemi zasnovani na sadržaju

- Informacije o filmovima kao što su pripadanje žanru, oznake (etikete) koje asociraju na određeni film odnosno *tag*, glumačka ekipa, režiser itd…
- Dobar izbor u slučaju ***cold-start*** problema (kada korisnik nije ocjenio nijedan film), ali u slučaju velikih baza zahtjevno u smislu potrebe opisivanja cijelog sadržaja
- Sličnost među filmovima određena nekom odabranom metrikom – *cosine similarity*, *Jaccard similarity, Pearson*-ov koeficijent korelacije, Euklidovo rastojanje…

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

Sistemi zasnovani na sadržaju

(a) Ratings

userId	movieId	rating
1	1	5
2	2	4
3	3	2
4	1	4
5	2	3
6	3	2
7	1	5
8	3	1
9	4	4

(b) Movies

movieId	title	genres
1	Toy Story	Animation|Children's|Comedy
2	Jumanji	Adventure|Children's|Fantasy
3	Grumpier Old Man	Comedy|Romance
4	Waiting to Exhale	Comedy|Drama
5	Father of the Bride Part II	Comedy

- ✓ TF-IDF vektorizacija primjenjena na *tag*-ove filmova koji su dati u obliku tekstualnih podataka
- ✓ Kodiranje žanrova iz tekstualnih podataka u vektorski oblik pomoću funcije *get_dummies()*
- ✓ Mjera sličnosti određena na osnovu *cosine similarity*

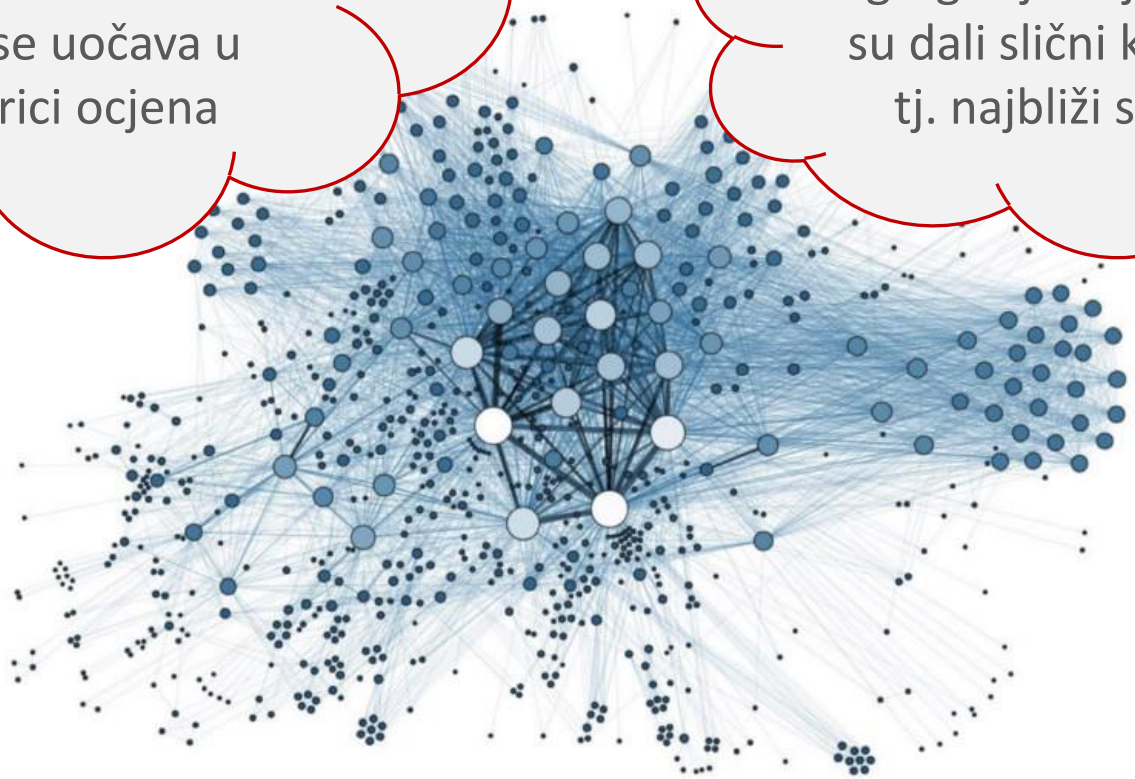
Kolaborativno filtriranje

Item-based

Tehnika za davanje preporuka bazirana na vezi između filmova koja se uočava u matrici ocjena

User-based

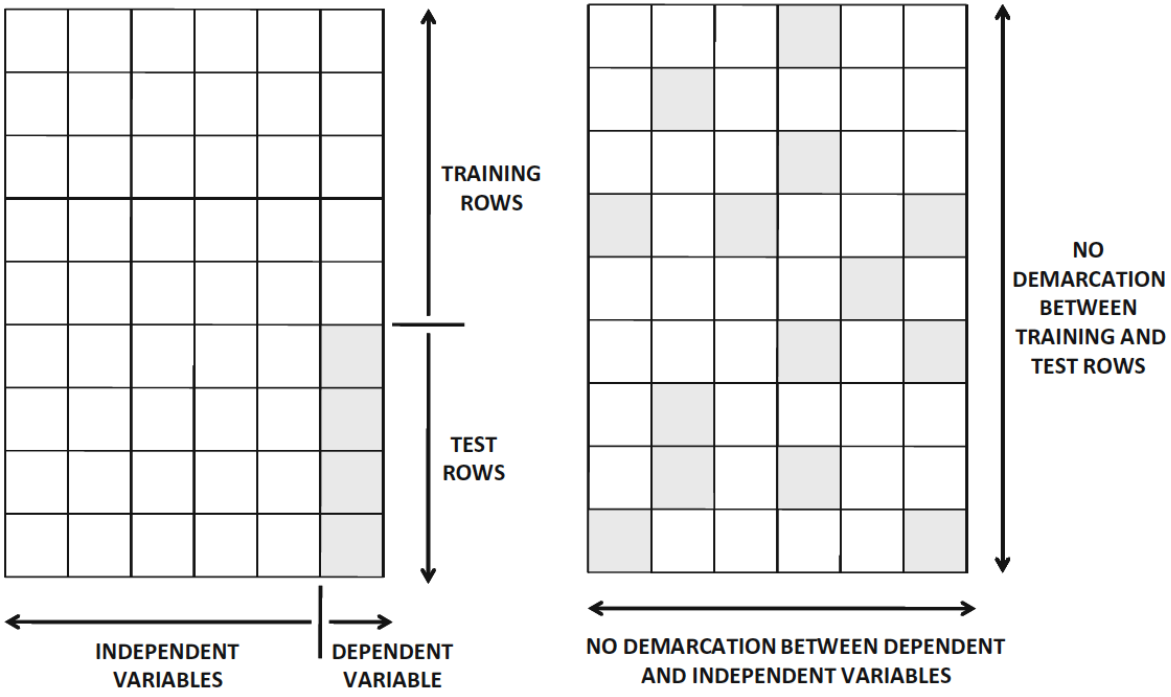
Formiranje predviđanja na osnovu agregacije ocjena koje su dali slični korisnici, tj. najbliži susjedi



Kolaborativno filtriranje

Podjela na skup za obučavanje i testiranje

Podaci u obliku matrice ocjena koje je ***i***. korisnik dao ***j***. filmu



(a) Classification (b) Collaborative filtering

Oba skupa su iste dimenzije, svaka vrijednost iz početne matrice ocjena se nalazi samo u jednom od ta dva skupa

Kolaborativno filtriranje

Baza podataka

- 9 742 filma
- 610 korisnika
- 100 836 ocjene

- Ocjene na skali od 0.5, 1, 1.5, …, 5
- ID filma = cijeli broj između 1 i 193609

Algoritmi su sprovedeni nad podskupovima podataka (iz baze podataka) koji sadrže:

- Sve filmove sa maksimalnim ID brojem 100 – s obzirom na to da u bazi ID brojevi filmova nisu dodjeljeni redom (neki brojevi nedostaju), ovakvih filmova ima 89, a broj korisnika koji je ocjenio te filmove je 503
- Sve filmove sa maksimalnim ID brojem 300 – ovakvih filmova ima 261, a broj korisnika koji je ocjenio te filmove je 571
- Sve filmove sa maksimalnim ID brojem 1000 – 761 film, 603 korisnika
- Sve filmove sa maksimalnim ID brojem 1500 – 1146 filmova, 608 korisnika
- Sve filmove sa maksimalnim ID brojem 3000 – 2259 filmova, 610 korisnika

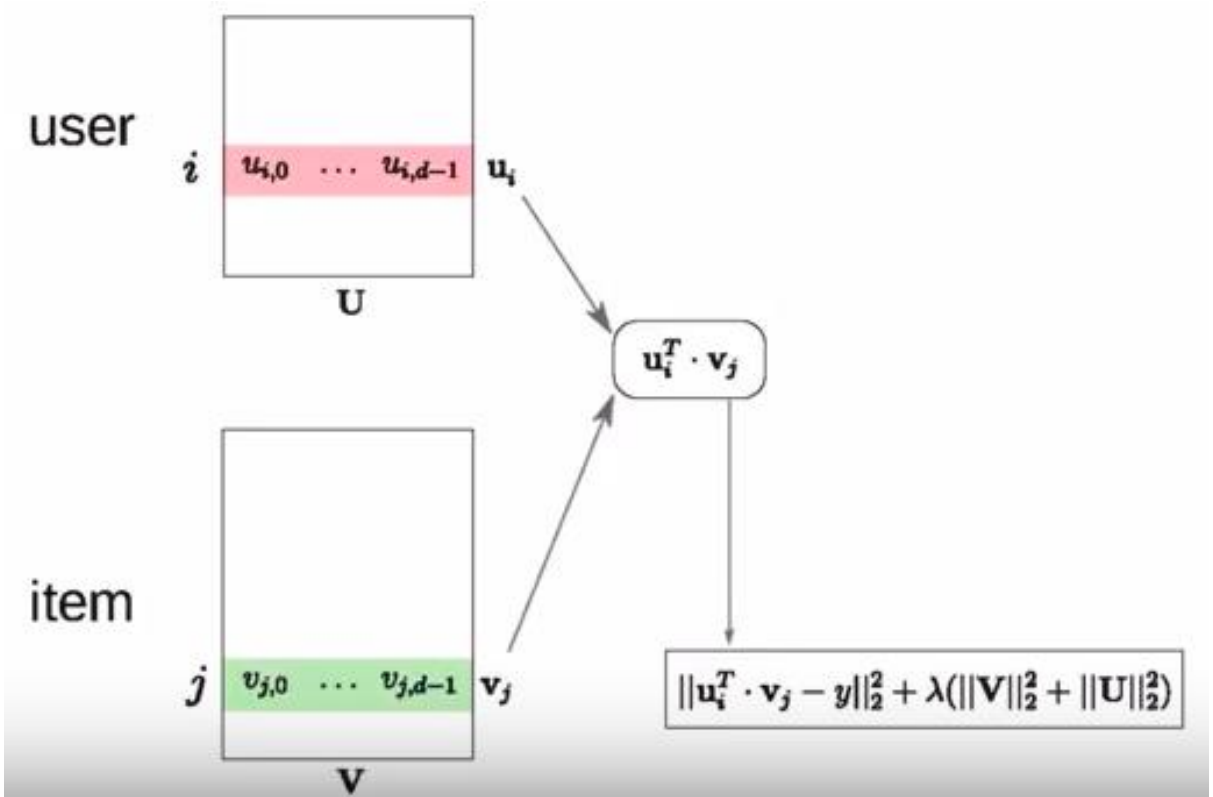
Kolaborativno filtriranje

Rezultati u slučaju Item-based i User-based CF

	maxId=100	maxId=300	maxId=1000	maxId=1500	maxId=3000
Item-based CF RMSE train	2.61	2.48	2.32	2.34	2.33
Item-based CF RMSE test	2.97	2.79	2.63	2.71	2.75
Item-based CF MAE train	2.34	2.19	2.02	2.03	2.02
Item-based CF MAE test	2.72	2.52	2.34	2.41	2.46
User-based CF RMSE train	2.09	2.28	2.36	2.43	2.46
User-based CF RMSE test	3.26	3.03	2.77	2.72	2.64
User-based CF MAE train	1.78	1.96	2.05	2.12	2.15
User-based CF MAE test	3.08	2.79	2.50	2.43	2.36

Kolaborativno filtriranje

- ✓ ***Singular Value Decomposition*** – nalazi matrice čiji proizvod daje matricu ocjena – smanjenje broja atributa, zadržava se informacija
- ✓ Problem jako prorijeđenih matrica
- ✓ **Metod latentnih faktora** uz optimizaciju Stohastičkim gradijentnim spustom



Kolaborativno filtriranje

Rezultati u slučaju Metoda latentnih faktora uz optimizaciju Stohastičkim gradijentnim spustom

	maxId=100	maxId=300	maxId=1000	maxId=1500	maxId=3000
SGD CF RMSE train	0.95	0.97	0.96	0.96	0.98
SGD CF RMSE test	0.95	0.95	0.94	0.93	0.98
SGD CF MAE train	0.76	0.79	0.78	0.79	0.83
SGD CF MAE test	0.77	0.76	0.76	0.76	0.80

Hvala na pažnji

Andrijana Marjanović 1043/2018