



Diplomski rad br. 2592

Hibridni sustav preporuke filmova

Dora Murk

Mentor: izv. prof. dr.sc. Goran Delač

06. srpnja 2021.

Sustavi preporuka

Što?

- Filtriranje i personaliziranje sadržaja

Tko?

- On-line trgovine, videoteke, knjižnice

Zašto?

- Lakše korištenje aplikacija, veća zarada

Kako?

- Podaci o sadržaju i korisnicima

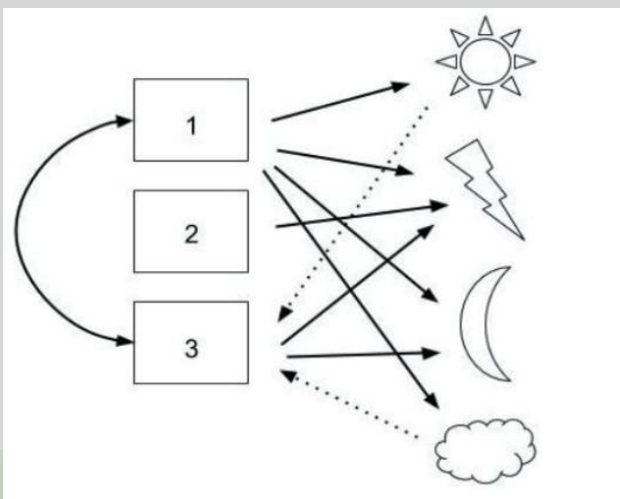
Zadatak

1. Sustav preporuke zasnovan na suradničkom filtriranju i sadržaju
2. Metrike za vrednovanje sustava preporuka
3. Skup podataka
4. Implementacija hibridnog sustava preporuke
5. Usporedba i vrednovanje sustava

Vrste sustava preporuka





Suradničko filtriranje

- Obrasci ponašanja korisnika
- Problem:
 - Hladni start
 - "siva ovca"
 - Rijetkost podataka



Filtriranje sadržaja

- Sličnost sadržaja prema obilježjima
- Problem:
 - Hladni start
 - Domensko znanje
 - Slične preporuke

	toplo	hladno	svijetlo	tamno	kišno
	x		x		
		x		x	
				x	x
	x		x		

Hibridni sustavi

- Kombinacija više tehnika preporučivanja
- Razne metode kombiniranja
- Prednost:
 - Eliminacija problema pojedinačnih tehnika
- $H = \mu S_1 + (1 - \mu) S_2$

Metrike za vrednovanje sustava

- Srednja apsolutna pogreška

- $MAE = \frac{1}{N} \sum (r - r')$

- Pogreška srednjih kvadrata

- $RMSE = \sqrt{\frac{1}{N} \sum (r - r')^2}$

- Stopa pogodaka

- Pokrivenost

- Povjerenje

Odabir skupa podataka

- *MovieLens (Kaggle)*
 - 9742 filma (*movies.csv*)
 - 610 korisnika
 - 100836 ocjena (*ratings.csv*)
- Mjera rijetkosti tablice korisnosti = 98.2%

movieId		title	genres
0	1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
1	2	Jumanji (1995)	Adventure Children Fantasy
2	3	Grumpier Old Men (1995)	Comedy Romance
3	4	Waiting to Exhale (1995)	Comedy Drama Romance
4	5	Father of the Bride Part II (1995)	Comedy
...

	userid	movieId	rating	timestamp
0	1	1	4.0	964982703
1	1	3	4.0	964981247
2	1	6	4.0	964982224
3	1	47	5.0	964983815
4	1	50	5.0	964982931
...

Implementacija

- **Hibridni sustav preporuke filmova**

- Sustav zasnovan na suradničkom filtriranju
 - Kosinusna sličnost korisnika
- Sustav zasnovan na sadržaju
 - Tf-idf vektorizacija filmova po žanrovima

	0	1	2	3	4	5	6	7	8	9	...	14	15	16	17	18	19	20	21	22	23
0	0.000000	0.416846	0.516225	0.504845	0.267586	0.0	0.0	0.000000	0.482990	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
1	0.000000	0.512361	0.000000	0.620525	0.000000	0.0	0.0	0.000000	0.593662	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2	0.000000	0.000000	0.000000	0.000000	0.570915	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.821009	0.0	0.0	0.0	0.0
3	0.000000	0.000000	0.000000	0.000000	0.505015	0.0	0.0	0.466405	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.726241	0.0	0.0	0.0	0.0
4	0.000000	0.000000	0.000000	0.000000	1.000000	0.0	0.0	0.000000	0.000000	0.0	...	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
...

- Težinska kombinacija navedenih sustava

	similarity	userId
300	0.124799	301
596	0.102631	597
413	0.101348	414
476	0.099217	477
56	0.099070	57
368	0.098295	369
205	0.096852	206
534	0.096493	535
589	0.095191	590
417	0.094153	418
119	0.092770	120
74	0.091987	75
576	0.089396	577
197	0.088883	198
159	0.088133	160
225	0.088068	226

Implementacija

- `get_utility_matrix(n_users, n_movies, df_ratings)`
- `normalize_utility_matrix(df_utility)`
- `sklearn.metrics.pairwise.cosine_similarity(X, Y=None)`
- `predict_CF(movieId, userId, df)`
- `_concatenate_genres_of_movies(genres)`
- `sklearn.feature_extraction.text.TfidfVectorizer()`
- `predict(MOVIE_ID, USER_ID)`

- Python biblioteke

- Pandas
- Scikit-learn



Vrednovanje sustava

- Slučajno uzorkovanje primjera
 - 30 slučajnih korisnika
 - 30 slučajnih filmova za svakog odabranog korisnika

- 5 uzoraka
 - Suradničko filtriranje
 - Filtriranje sadržaja
 - Hibridni sustavi:

$$H_1 = 0.5 * CF + 0.5 * CB$$
$$H_2 = 0.75 * CF + 0.25 * CB$$
$$H_3 = 0.25 * CF + 0.75 * CB$$

Sustav preporuke	Metrika	Uzorak1	Uzorak2	Uzorak3	Uzorak4	Uzorak5	Srednja vrijednost
CF	rmse	0.791915	0.819120	0.771614	0.811314	0.818518	0.802496
CF	mae	0.580407	0.584551	0.557123	0.581288	0.600645	0.580803
CB	rmse	0.938147	0.889874	0.884212	1.035040	0.960284	0.941511
CB	mae	0.715454	0.657268	0.660415	0.777619	0.728923	0.707936
H1	rmse	0.764795	0.751113	0.736063	0.815694	0.793135	0.772160
H1	mae	0.565100	0.532391	0.532492	0.589676	0.578401	0.559612
H2	rmse	0.756587	0.762917	0.734201	0.789728	0.784635	0.765614
H2	mae	0.554376	0.542970	0.530891	0.569659	0.576055	0.554790
H3	rmse	0.813670	0.777001	0.775289	0.880046	0.838486	0.816898
H3	mae	0.605032	0.549289	0.561057	0.636258	0.614383	0.593204

Zaključak

- Korisna uporaba velikih skupova podataka
- Razni pristupi i pretpostavke
- Lako vrednovanje i evaluacija rješenja
- Velike prednosti hibridnih pristupa preporučivanju