Metoda uzajamnog filtriranja i primene u elektronskoj trgovini

master rad

mentor: dr Aleksandar Kartelj

Jovana Pejkić
1033/2020

Matematički fakultet
Univerzitet u Beogradu

27. septembar 2021.

Sadržaj

Motivacija

- Rukovanje ogromnom količinom podataka uz korišćenje postojećih alata postalo je nemoguće;
- Javlja se potreba za naprednijim pristupima u pretraživanju i filtriranju informacija;
- Mnoge veb stranice za e-trgovinu koriste alate kao što su sistemi za preporuke;
- Glavni ishod dobrog sistema za preporuke je povećanje lojalnosti kupaca, a time i povećanje zarade.

Opis problema

- Sa početkom Veb 2.0 (oko 2004. god.) i prodavnice i kupci podjednako su se „preselili" na veb;
- Javljaju se pitanja:
  - Kako bi onlajn kupac trebalo da filtrira ogromnu količinu opcija koje se nude?
  - Kako otkriti nove proizvode?
  - Kako među milionima takvih proizvoda pronaći onaj koji se traži?

Korišćene metode

- Sistem za preporuke koji koristi uzajamno filtriranje i KNN;
- Sistemi za preporuke koji koriste tehnike KNN i SVD;
- Sistem za preporuke zasnovan na dubokom učenju;
- Hibridni sistemi za preporuke.

(a) Sistem za preporuke koji koristi uzajamno filtriranje i KNN

Mere sličnosti

- Pretpostavka: „sličnim korisnicima se sviđaju slične stavke";
- Pitanje: kako definisati i identifikovati slične korisnike, da bi se pronašle slične stavke koje mogu biti preporučene?
- U obzir treba uzeti matricu korisnik-stavka (uglavnom retka);
- Mere sličnosti:
  - **Euklidsko rastojanje** (eng. euclidean distance)
  - Pirsonova korelacija (eng. pearson correlation)
  - **Kosinusna sličnost** (eng. cosine similarity)
  - Prilagođena kosinusna mera sličnosti (eng. adjusted cosine similarity)
  - Žakardova sličnost (eng. jaccard similarity)
  - Srednje kvadratno rastojanje (eng. mean square distance)

Funkcije predviđanja

- Težinska suma (eng. weighted sum);
- Pristup zasnovan na centriranju proseka (eng. mean centering approach);
- Pristup zasnovan na z-rezultatu (eng. z-score approach);

Metrike za evaluaciju

- Mere koje se koriste za evaluaciju modela za klasifikaciju, uz određene modifikacije, mogu biti iskorišćene i kod sistema za preporuke;
- Sistemi za preporuke mogu biti evaluirani u pogledu:
  - tačnosti predviđenih ocena (MAE, RMSE)
  - tačnosti rangiranja stavki (proračuni zasnovani na korisnosti (eng. utility-based computations), koeficijenti korelacije ranga (eng. rank correlation coefficients) i ROC krive (eng. receiver operating characteristic curve).);
- U ovom radu, fokus je na merenju tačnosti predviđenih ocena.

Podaci

- Skup podataka sadrži informacije o proizvodima koji se prodaju na Amazonu.
- Sastoji iz dve tabele:
  - Prva tabela sadrži informacije o proizvodima, sastoji iz 19999 redova i 449 kolona.
  - Druga tabela sadrži informacije o interakcijama korisnik-stavka, astoji iz 19999 redova i 29 kolona.
- Atributi koji za sprovođenje pomenutih algoritama nisu od značaja su eliminisani u procesu pretprocesiranja.

Podaci

	asin	title	brand	dimensions	weight
0	0764443682	Slime Time Fall Fest [With CDROM and Collector...	Group Publishing (CO)	Product Dimensions:\n \n8.7...	Shipping Weight:\n \n2.4 po...
2	1940280001	Magical Things I Really Do Do Too!	Christopher Manos	Package Dimensions:\n \n8.5...	Shipping Weight:\n \n6.1 ou...

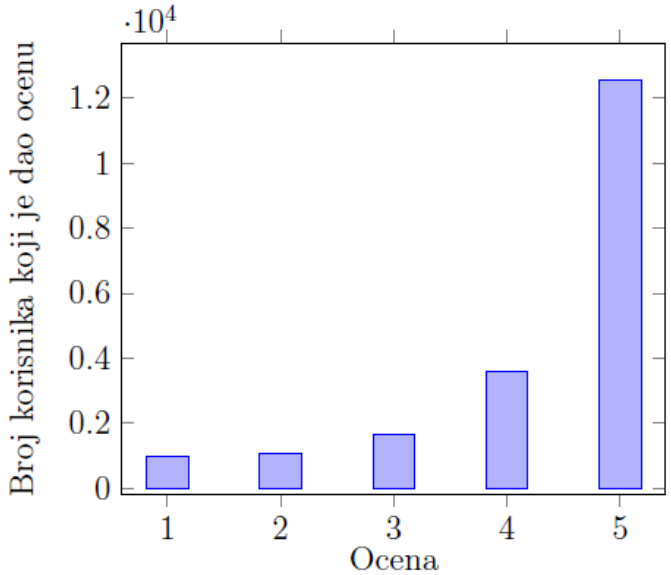
Slika: Tabela 1 - o proizvodima

	reviewerID	reviewerName	asin	reviewText	summary	overall
0	A1D4G1SNUZWQOT	Tracy	7106116521	Exactly what I needed.	perfect replacements!!	5
1	A3DDWDH9PX2YX2	Sonja Lau	7106116521	I agree with the other review, the opening is ...	I agree with the other review, the opening is ...	2

Slika: Tabela 2 - o interakcijama korisnik-stavka

Podaci

- Vrednosti atributa koji predstavljaju ocenu:
  - su iz intervala [1, 5]
  - mogu biti samo celi brojevi



Slika 5.1: Ocenjivanje proizvoda

Slika: Prikaz koliko korisnika je dalo koju ocenu

Sistem za preporuke koji koristi uzajamno filtriranje i KNN

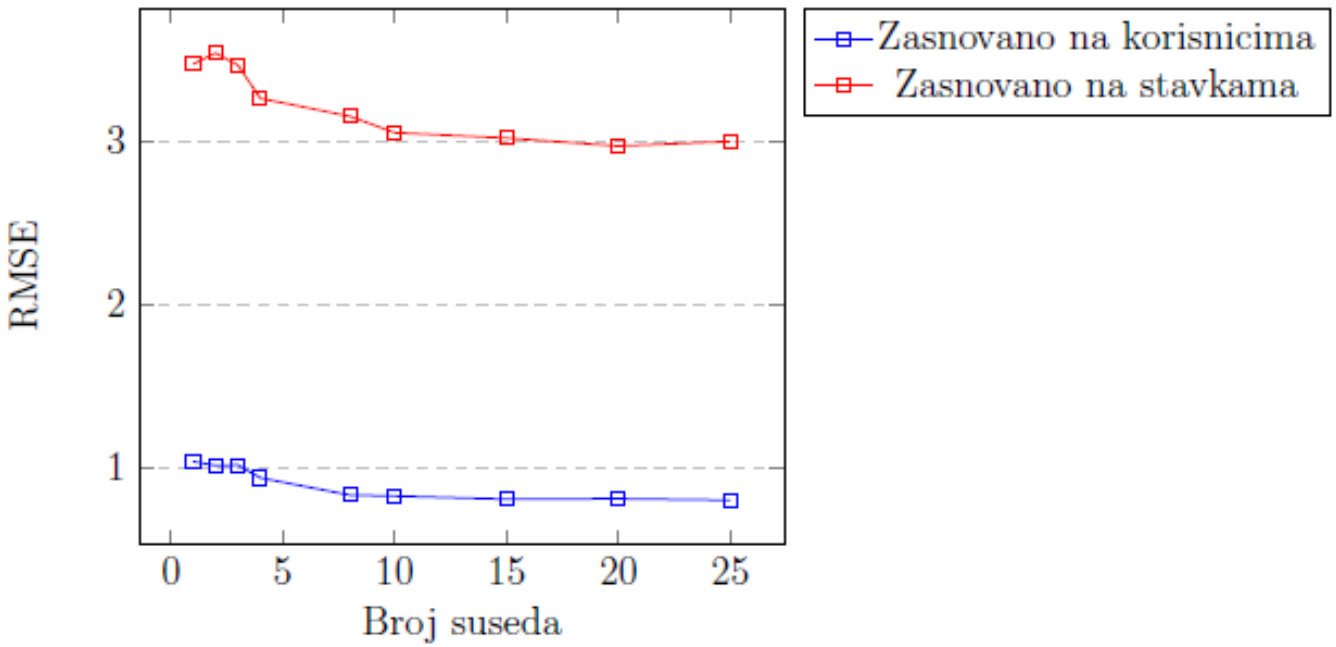
- Skup podataka koji je korišćen za potrebe ove implementacije sastoji se samo iz tabele sa ocenama proizvoda koje su im korisnici dodelili;
- Mera za merenje tačnosti predviđenih ocena je RMSE;
- Sistem je evaluiran za različite vrednosti parametra k (broja suseda), za pristup zasnovan na korisnicima i pristup zasnovan na stavkama;
- U formiranoj matrica ocena sve nedefinisane vrednosti su zamenjene nulom.

Sistem za preporuke koji koristi uzajamno filtriranje i KNN

Tabela: Matrica ocena

Id stavke / Id korisnika	1	2	3	...	193585	193587	193609
1	4.0	0.0	4.0	...	0.0	0.0	0.0
2	0.0	0.0	0.0	...	0.0	0.0	0.0
3	0.0	0.0	0.0	...	0.0	0.0	0.0
4	0.0	0.0	0.0	...	0.0	0.0	0.0
5	4.0	0.0	0.0	...	0.0	0.0	0.0

Sistem za preporuke koji koristi uzajamno filtriranje i KNN



(a) Sistem za preporuke koji koristi uzajamno filtriranje i KNN

Sistem za preporuke koji koristi uzajamno filtriranje i KNN

Tabela: Tabelarni prikaz greške RMSE za pristup zasnovan na stavkama i pristup zasnovan na korisnicima

Broj suseda k	1	2	3	4	8
Zasnovano na korisnicima	1.036	1.012	1.014	0.935	0.830
Zasnovano na stavkama	3.470	3.538	3.465	3.261	3.151

Broj suseda k	10	15	20	25
Zasnovano na korisnicima	0.824	0.805	0.808	0.799
Zasnovano na stavkama	3.051	3.018	2.968	2.997

Sistem za preporuke koji koristi uzajamno filtriranje i KNN

Prema datim rezultatima može se zaključiti da pristup zasnovan na korisnicima daje bolja predviđanja.

Sistemi za preporuke koji koriste tehnike KNN i SVD

- Dva različita pristupa:
  - pristup zasnovan na memoriji (varijacije algoritma KNN),
  - pristup zasnovan na modelu (SVD tehnika);
- Za implementaciju oba pristupa korišćena je biblioteka Surprise;

Sistemi za preporuke koji koriste tehnike KNN i SVD

- Korišćena je tabela koja sadrži atribute:
  - id korisnika,
  - naziv proizvoda i
  - ocenu;
- Eliminisani su svi duplikati;
- Svi modeli (SVD i varijacije KNN) su obučavani korišćenjem 5-slojne unakrsne provere (eng. 5 fold cross validation);
- Za evaluaciju su korišćene mere MAE i RMSE.

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

Tabela: Performanse modela

Metod	MAE	RMSE	Parametar *fit_time*	Parametar *test_time*
SVD	0.846150	1.066627	0.142742	0.002011
KNNBasic	0.852506	1.076491	0.120428	0.004301
KNNBaseline	0.850317	1.082411	0.091381	0.003928
KNNWithMeans	0.859457	1.097607	0.296933	0.005781
KNNWithZScore	0.865696	1.101871	0.243386	0.005185

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

- Među isprobanim algoritmima, najmanju grešku RMSE ima SVD;
- Što se tiče varijacija KNN algoritama, najmanju grešku RMSE ima KNNBasic;
- Algoritam KNNWithMeans ima najveću vrednost parametra fit_time, dok algoritam SVD ima najmanju vrednost parametra test_time.

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

- Skup podataka je podeljen slučajnim uzorkom na skup podataka za obučavanje (70%) i skup podataka za testiranje (30%);
- Za modele KNNBasic i SVD izvršeno je podešavanje parametara pri čemu je opet korišćena 5-slojna unakrsna provera;

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

Tabela: Vrednosti parametara za koje se postižu najbolje performanse modela KNNBasic

Metod	mera sličnosti	parametar min_support	parametar user_based
KNNBasic	MSD	3	False

- Tada MAE iznosi 0.87106, a RMSE 1.08917.

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

Tabela: Vrednosti parametara za koje se postižu najbolje performanse modela SVD da bi se postigla najveća vrednost mere MAE

Metod	parametar n_factors	parametar n_epochs	parametar lr_all	parametar reg_all
SVD	50	5	0.005	0.2

- Tada MAE iznosi 0.85317.

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

Tabela: Vrednosti parametara za koje se postižu najbolje performanse modela SVD da bi se postigla najveća vrednost mere RMSE

Metod	parametar n_factors	parametar n_epochs	parametar lr_all	parametar reg_all
SVD	90	5	0.002	0.5

- Tada RMSE iznosi 1.079925.

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

- Modeli su testirani na skupu podataka za testiranje;
- Za evaluaciju su korisćene mere MAE i RMSE;

Tabela: Vrednosti mera MAE i RMSE za metode KNNBasic i SVD

Metod	MAE	RMSE
KNNBasic	0.9197	1.1572
SVD	0.8760	1.0996

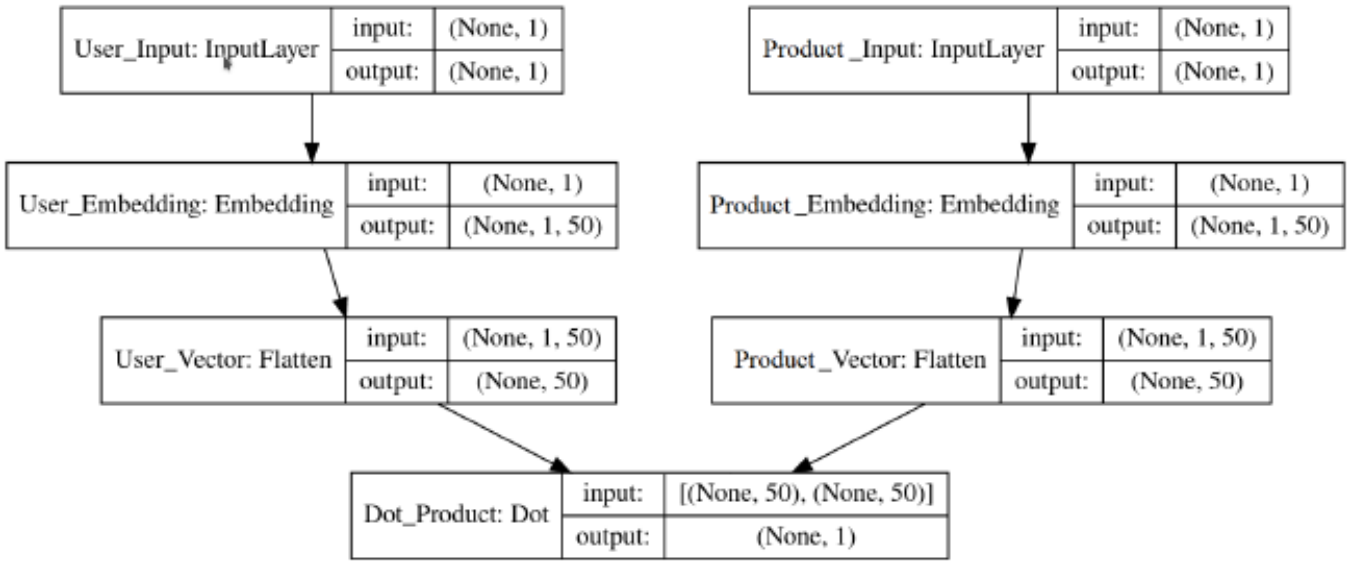
- Vrednosti mera MAE i RMSE za oba metoda su uporediva sa vrednostima dobijenim na skupu za obučavanje;
- Modeli KNNBasic i SVD dobro generalizuju.

Evaluacija modela - Sistemi za preporuke koji koriste tehnike KNN i SVD

- Preporuke dobijene koristeći algoritme KNNBasic i SVD su iste, samo je redosled (prioritet) drugačiji;
- Očekivano s obzirom na to da su u pitanju različiti algoritmi;
- Modeli se mogu unaprediti implementiranjem hibridnog pristupa;
  - Uzajamno filtriranja nije dobra opcija u slučaju novog korisnika ili nove stavke;
  - U takvim slučajevima mogao bi da se iskoristi model zasnovan na sadržaju,

Evaluacija modela - Sistemi za preporuke zasnovani na dubokom učenju

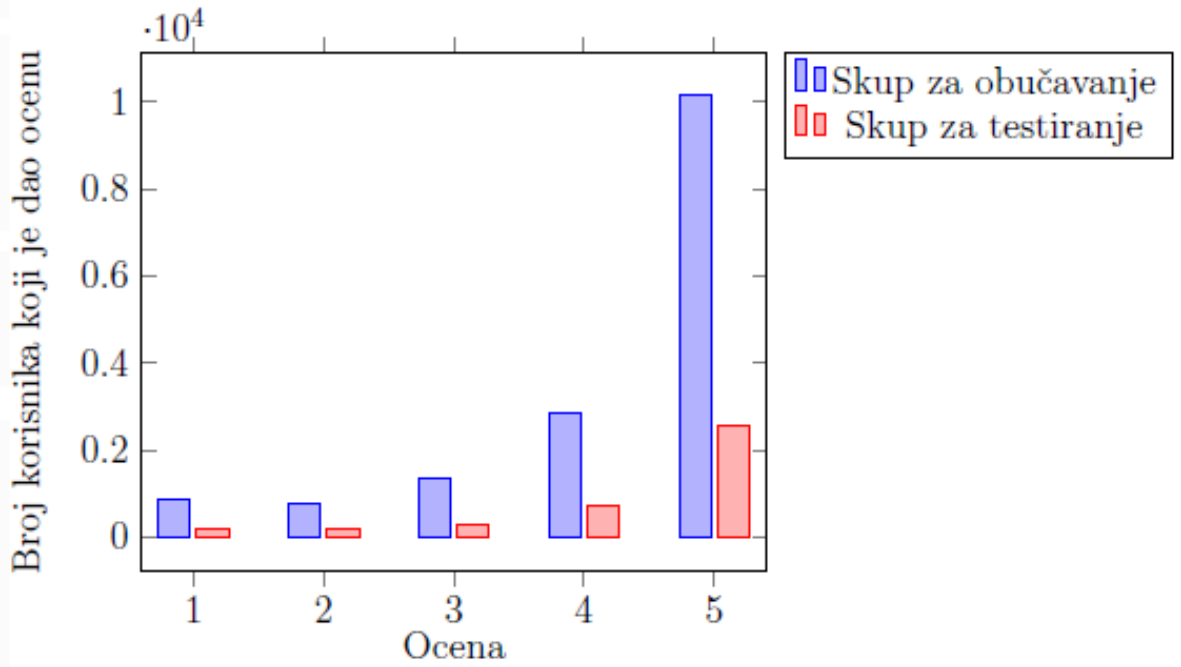
- Za implementaciju je korišćena biblioteka Keras;
- Arhitekturu modela čine tri sloja.



(a) Arhitektura mreže

Evaluacija modela - Sistemi za preporuke zasnovani na dubokom učenju

- Skup podataka je podeljen na skupove za obučavanje i testiranje u razmeri 70:30.



(a) Vizuelni prikaz podele skupa podataka

Evaluacija modela - Sistemi za preporuke zasnovani na dubokom učenju

Tabela: Vrednosti funkcije gubitka

Parametar *epochs*	Parametar *batch_size*	Funkcije gubitka – obučvanje	Funkcije gubitka – testiranje
10	64	0.02155	17.58496
10	128	0.06493	17.59478
20	64	0.02150	17.59012
20	128	0.52304	17.63714
30	64	0.01992	17.57510
30	128	0.00501	17.56958
100	64	0.01391	17.54213
100	128	0.00964	17.55469

Evaluacija modela - Hibridni sistemi za preporuke

- Dve vrste hibridnih sistema:
  - hibrid zasnovan na težinama;
  - hibrid sa smenjivanjem.
- Evaluacija: izračunate su razlike između predviđene i stvarne ocene;
- Tri tabele: users_all, items_all, ratings_all.

Evaluacija modela - Hibridni sistem sa težinama

Tabela: Hibridni sistem zasnovan na težinama – rezultati

Predviđena ocena	Greška
4.5625	0.4375

Evaluacija modela - Hibridni sistem sa smenjivanjem

Tabela: Hibridni sistem sa smenjivanjem – rezultati

Pristup	Predviđena ocena	Greška
Uzajamno filtriranje	4.125	0.875
Zasnovan na sadržaju	2.000	3.000

Poređenje sa rezultatima postojećeg rada

- Predviđanje ocena za filmove koristeći uzajamno filtriranje;

Tabela: Približne vrednosti rezultata za prvi skup podataka

Pristup zasnovan na	Veličina skupa za obučavanje	RMSE
stavkama	90	3.55
stavkama	75	3.60
stavkama	50	3.60
korisnicima	90	3.30
korisnicima	75	3.35
korisnicima	50	3.45

Tabela: Približne vrednosti rezultata za drugi skup podataka

Pristup zasnovan na	Veličina skupa za obučavanje	RMSE
stavkama	90	3.40
stavkama	75	3.45
stavkama	50	3.55
korisnicima	90	3.00
korisnicima	75	3.10
korisnicima	50	3.30

Diskusija i zaključak

- Nedostaci i zaključci:
  - Nedovoljna količina podataka;
  - Neki pristupi daju odlične rezultate, dok su rezultati dobijeni drugim pristupima nezadovoljavajući;
  - Nije razmatran "efekat promene preferencije korisnika";
- Moguća poboljšanja:
  - Kombinovanje više različitih modela u okviru hibridnog sistema;
  - Rad sa većim skupom podataka, obogaćivanje baze podataka, dodavanje novih atributa.

Hvala na pažnji.
Pitanja?