Usporedba kontekstno utemeljenih i kolaborativnih sustava za davanje preporuka te implementacija hibridnog sustava

Petrak, Dina

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 699

USPOREDBA KONTEKSTNO UTEMELJENIH I KOLABORATIVNIH SUSTAVA ZA DAVANJE PREPORUKA TE IMPLEMENTACIJA HIBRIDNOG SUSTAVA

Dina Petrak

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 699

USPOREDBA KONTEKSTNO UTEMELJENIH I KOLABORATIVNIH SUSTAVA ZA DAVANJE PREPORUKA TE IMPLEMENTACIJA HIBRIDNOG SUSTAVA

Dina Petrak

Zagreb, veljača 2025.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

Zagreb, 30. rujna 2024.

DIPLOMSKI ZADATAK br. 699

Pristupnica:	**Dina Petrak (0036523849)**
Studij:	Računarstvo
Profil:	Znanost o podacima
Mentor:	izv. prof. dr. sc. Marko Đurasević

Zadatak: **Usporedba kontekstno utemeljenih i kolaborativnih sustava za davanje preporuka te implementacija hibridnog sustava**

Opis zadatka:

Opisati funkcioniranje sustava za davanje preporuka te objasniti razliku u načinu rada kontekstno utemeljenih i kolaborativnih postupaka filtriranja. Razmotriti i primijeniti različite pristupe implementacije za svaku od ovih metoda filtriranja. Analizirati prednosti i nedostatke njihovih performansa na problemima s različitim karakteristikama. Potom istražiti mogućnosti kombiniranja tih metoda u hibridni sustav preporučivanja. Testirati i evaluirati različite pristupe s ciljem poboljšanja točnosti i relevantnosti preporuka Na temelju dobivenih rezultata izvesti zaključke o učinkovitosti pojedinih pristupa. Izvorni kod programa, postignute rezultate, prigodno vizualizirane uz zaključke te korištenu literaturu uključiti u rad.

Rok za predaju rada: 14. veljače 2025.

Zahvaljujem mentoru izv. prof. dr. sc. Marku Đuraseviću na korisnim savjetima i pomoći pri izradi ovog rada.

Hvala mami i tati na svim mogućim oblicima podrške koju su mi pružili, a posebno vam hvala za trenutke u kojima ste vjerovali u mene, čak i onda kada ja nisam.

Hvala Anjici što me uvijek znala oraspoložiti i nasmijati.

Hvala bratićima što su uskočili u pomoć nebrojeno puta, u studentskom životu, ali i izvan njega.

Hvala prijateljima i Svim ostalima bez kojih studentski period ne bi prošao ovako zabavno.

Sadržaj

1. Uvod

Razvoj tehnologije uzrokovao je značajni rast količine podataka koju svakodnevno generiramo i pohranjujemo. Takav nagli rast rezultirao je neizbježnim izazovom učinkovitog upravljanja informacijama. Upravljanje tim podacima postalo je ključno za maksimiziranje njihove vrijednosti. Jedan od glavnih problema u tom procesu jest, u ogromnoj količini informacija, prepoznati koje su zaista važne i korisne, a koje nisu. Ovaj problem postaje sve izraženiji kako se količina podataka povećava te učinkovito upravljanje informacijama postaje sve složenije.

Ovakav problem filtriranja izražen je i u svakodnevnom životu gdje je dostupni sadržaj izuzetno širok a ljudi često nemaju vremena za pretraživanje i pronalaženje najrelevantnijih informacija. U tom kontekstu, izazov upravljanja informacijama nije samo u tehničkom organiziranju i pohranjivanju podataka, već i u prepoznavanju i izdvajaju sadržaja koji ima stvarnu vrijednost za korisnike. U takvim situacijama, postaje ključno razviti učinkovite metode za filtriranje informacija, kako bi i korisnici mogli brže i lakše pronaći ono što im je zaista važno.

Korisnici sve više ovise o platformama koje nude širok spektar proizvoda i sadržaja, zbog čega su suočeni s neprestanim izborom. Među tim platformama nalaze se internetske stranice za kupovinu, aplikacije za slušanje glazbe, gledanje filmova, te mnoge druge usluge. Iako raznolikost proizvoda i sadržaja nudi prednosti, ona također donosi određene izazove. Na prvi pogled može se činiti da je široka ponuda opcija prednost. Međutim, preveliki izbor može dovesti do pojave poznate kao paraliza odlučivanja. Kada se ljudi suoče s prevelikim brojem opcija, često se osjećaju preopterećeno, što otežava izbor i može dovesti do potpunog izbjegavanja donošenja odluke.

U takvim okolnostima, sustavi za preporuke postaju neophodni. Sustavi za davanje

preporuka, ili *recommender systems*, su vrsta algoritama strojnog učenja koji analiziraju korisničke podatke kako bi predvidjeli njihove preferencije. Oni igraju ključnu ulogu u smanjenju preopterećenosti informacijama kroz filtriranje, rangiranje i učinkovito dostavljanje relevantnih podataka. Ovi sustavi uspješno rješavaju problem velike količine proizvoda i sadržaja, pružajući korisnicima personalizirane prijedloge i usluge. Upravo zbog toga, takvi sustavi postali su sastavni dio mnogih platformi. Istovremeno, takvi sustavi pružateljima usluga donose veliku korist. Omogućuju ciljano plasiranje ponuda, pridonose pozitivnom iskustvu korisnika i time poboljšavaju ukupnu učinkovitost platforme.

U ovom radu analizirani su sustavi za davanje preporuka. Dana je njihova osnovna podjela te je opisan način rada pojedinih podvrsta. Primijenjeni su različiti pristupi implementacije te su performanse ispitane na problemima različitih karakteristika.

2. Podjela sustava za davanje preporuka

Sustavi za davanje preporuka su zapravo podvrsta strojnog učenja te se temelje na generiranju predikcija stavova o nekom proizvodu, većinom s ciljem da predvide što se korisniku sviđa. Iz velike količine generiranih informacija filtriraju ključne fragmente prema korisnikovim preferencijama, interesima ili promatranom ponašanju.

Postoji puno različitih vrsta sustava za davanje preporuka, što dovodi do raznih podjela i kategorizacija. Jedna od najosnovnijih podjela je ona na personalizirane i nepersonalizirane sustave.

Sustavi za davanje preporuka mogu biti vrlo jednostavni kada ne uzimaju u obzir individualne karakteristike svakog korisnika, već se oslanjaju na generalne podatke. Primjerice, takav sustav može uključivati preporuke proizvoda na temelju njihove popularnosti, ocjena korisnika ili kombinacije tih čimbenika. Prednost takvih sustava je njihova jednostavnost. Za generiranje preporuka nisu potrebni složeni algoritmi, a podaci se ne moraju analizirati posebno za svakog korisnika.

Glavni problem s nepersonaliziranim sustavima preporuka je taj što njihove preporuke često nisu precizne i korisne za korisnike. Uz to, takvi sustavi preporučuju već popularne proizvode, čime oni postaju još popularniji, dok manje popularni proizvodi ostaju nevidljivi. Ta pojava naziva se "problem dugog repa". Kako bi se izbjegla ova ograničenja, koriste se personalizirani sustavi preporuka.

Općenito, u današnje vrijeme, kada se govori o sustavima za davanje preporuka, naglasak se stavlja na personalizirane sustave. Oni omogućuju korisnicima da dobiju preporuke prilagođene njihovim specifičnim interesima i potrebama. No, postavlja se pitanje kako ovi sustavi zapravo rade.

Generalna ideja ovakvih sustava je razvoj funkcije koja uzima u obzir ciljane korisnike i određene stavke. Takva funkcija trebala bi mjeriti relevantnost stavke za pojedinog korisnika.

Početak razvoja takvog sustava sastoji se od prikupljanja podataka o korisnicima, stavkama te međusobnim interakcijama. Podaci se mogu prikupljati eksplicitno ili implicitno. Implicitni podaci mogu se prikupiti iz povijesti narudžbi, klikova na određene stavke ili broja puta kada se neka pjesma sluša, dok se eksplicitni podaci dobivaju iz ocjena i povratnih informacija korisnika. Sustav za davanje preporuka trebao bi efikasnije raditi s povećanjem dostupnih podataka.

Nakon prikupljanja, učitavanja i formatiranja podataka, slijedi proces izračuna korisnosti stavki za korisnika, koji se primjerice izražava u obliku predikcije ocjena. Za izračunavanje odnosno procjenu nepoznatih ocjena postoji nekoliko metoda a glavne će biti objašnjene u ovom radu.

Osnovni algoritam za izračun predikcija, odnosno davanje preporuka, temelji se na matricama. Te matrice predstavljaju prikupljene podatke koji najčešće uključuju informacije o proizvodima, korisnicima ili različitim atributima povezanim s njima. Ključna ideja sustava preporuka je prepoznavanje uzoraka sličnosti unutar tih matrica, na temelju čega se generiraju preporuke korisnicima.

Primjerice, zamislimo matricu u kojoj redci predstavljaju modne artikle, dok stupci sadrže njihove karakteristike, unesene kao numeričke vrijednosti. Prvi stupac može označavati vrstu artikla, gdje broj 0 označava hlače, a broj 1 majice. Drugi stupac može predstavljati cjenovni rang artikla, prikazan skalom od 1-5. dok ostali stupci mogu obuhvaćati dodatne atribute, poput boje, materijala ili stila. Nakon što je matrica značajki stvorena, sljedeći korak je izrada profila korisnika. To zapravo znači pogledati koje je artikle korisnik već kupio i ocijenio pozitivno. Na primjer, ako korisnik često ocjenjuje hlače u srednjem cjenovnom rangu visoko, njegov profil će odraziti preferencije za te specifične karakteristike. Nakon toga, pristupa se koraku izračuna sličnosti proizvoda, gdje se proizvodi koji su se svidjeli korisniku uspoređuju s ostalim dostupnim artiklima. Artikli koji su slični preporučuju se korisniku.

S druge strane, problemu možemo pristupiti tako da u matricu postavimo korisnike

kao retke, a artikle koje su ocijenili kao stupce. U takvoj matrici gledamo koji korisnici imaju slične ukuse i preferencije.

Upravo ova dva pristupa čine glavnu podjelu personaliziranih sustava za davanje preporuka na *Content Based Recommendation Systems* odnosno sustave temeljene na sadržaju objekta ili *Collaborative Filtering Recommendation Systems* odnosno sustave temeljene na suradnji. Osim toga postoje i hibridni sustavi koji koriste kombinaciju *content based* i *collaborative filtriranja*.

2.1. Sustavi temeljeni na sadržaju

Prva opisana metoda spadala bi u ovu podvrstu sustava za davanje preporuka. Sustavi temeljeni na sadržaju koriste informacije o karakteristikama proizvoda ali i korisnika. Riječ sadržaj u samom nazivu ovih sustava odnosi se na značajke koji opisuju proizvode i korisnike. Te značajke mogu biti bilo što što stavku izdvaja od drugih stavki. U slučaju našeg jednostavnog primjera, ti atributi su bili vrsta artikla, cjenovni rang i boje. Inače, primjeri ovih atributa mogu uključivati autora knjige, izvođača pjesme, temu novinarskog članka ili žanr filma.

Osim karakteristika stavki, drugi ključan element sustava temeljenih na sadržaju je korisnički profil. Ideja je da se preporuke generiraju na osnovu karakteristika artikala ali i korisničkog profila. Sustav sprema podatke koje korisnik pruža na platformi, primjerice ocjenjivanjem, te na temelju tih podataka generira korisnički profil.

Može se lako zaključiti da se ova metoda preporuka uvelike oslanja na sposobnost opisivanja karakterističnih značajki stavki. Ne trebamo veliku zajednicu korisnika, detaljnu povijest ocjenjivanja ili puno povratnih informacija da bismo izračunali ove preporuke. Iz tog razloga, sustavi temeljeni na sadržaju imaju puno manjih problema sa hladnim startom.

Hladni start je problem koji se javlja kada sustav preporuka nema dovoljno informacija za generiranje relevantnih preporuka. Iako sustavi temeljeni na sadržaju zahtijevaju neke početne ulazne podatke kako bi mogli davati preporuke, kvaliteta ranijih predikcija općenito je bolja od ostalih sustava koji zahtijevaju korelacije iznimno velikog broja korisnika. U slučaju sustava temeljenih na sadržaju, za davanje uspješnih preporuka nisu

potrebni nikakvi podatci od drugih korisnika, što značajno smanjuje utjecaj problema hladnog starta.

Preporuke ovakvih sustava obično su vrlo prilagođene interesima korisnika te su iz tog razloga visoko relevantne. Negativna strana takvog pristupa je činjenica da sam korisnik može pretpostaviti da će mu se stavka svidjeti. Primjerice, kada korisnik ima omiljenu marku odjeće, vjerojatno će sam potražiti nove artikle te marke, bez obzira na preporuke sustava. Dakle, u sustavima temeljenim na sadržaju može se pojaviti nedostatak novosti i raznolikosti.

Osim toga, velik nedostatak tih sustava je potreba za konstantnim definiranjem i dodavanjem atributa. Svaki put kada se pojavi nova stavka potrebno je detektirati njene karakteristike i efikasno ih pohraniti, što može biti zahtjevan proces.

2.2. Sustavi temeljeni na suradnji

Druga velika skupina sustava za davanje preporuka su sustavi temeljeni na suradničkom filtriranju. Ti sustavi koriste metode koje se temelje isključivo na zabilježenim interakcijama između korisnika i stavki te uz pomoć njih stvaraju nove preporuke. Glavna ideja ove metode je da su dosadašnje interakcije između korisnika i stavki dovoljne za detektiranje sličnih korisnika pa samim time i sličnih stavki. Na temelju procijenjenih sličnosti između korisnika, sustav može identificirati obrasce u ukusima i na taj način generirati predikcije. Ključna pretpostavka ove metode jest da će korisnici koji pokazuju sličan ukus prema određenom proizvodu vrlo vjerojatno imati slične preferencije i za druge proizvode.

Sustavi temeljeni na suradnji obično se dijele u dvije glavne kategorije: pristupi temeljeni na modelima i pristupi temeljeni na memoriji.

Pristupi temeljeni na modelima koriste metode strojnog učenja za izradu modela te preporuke temelje na predikcijama tog modela. Podaci iz skupa koriste se kao ulazne vrijednosti za treniranja modela, a za generiranje predikcija koriste se različite metode. Te metode obuhvaćaju tehnike poput stabala odlučivanja, neuronskih mreža i bayesovske statistike.

Pristupi temeljeni na memoriji izravno koriste zabilježene podatke o interakcijama, ne oslanjajući se na model. U osnovi, daju predikcije temelje na pretraživanju najbližih susjeda. Ti sustavi se mogu dalje podijeliti na filtriranje temeljeno na korisnicima i filtriranje temeljeno na stavkama. Filtriranje temeljeno na korisnicima procjenjuje koju bi ocjenu neki korisnik mogao dati određenom proizvodu na temelju ocjena koje su toj stavci dali slični korisnici. S druge strane, filtriranje temeljeno na stavkama predviđa ocjene proizvoda ovisno o njihovoj sličnosti s drugim proizvodima na temelju prethodnih ocjena.

Sustavi temeljeni na suradničkom filtriranju imaju svoje prednosti i nedostatke. Kao što je spomenuto, ključna ideja ovih sustava je da koriste isključivo interakcije između korisnika i stavki, što donosi velike prednosti, ali također ima i svoje probleme.

Glavna prednost takvih sustava je da ne zahtijevaju informacije o karakteristikama korisnika ili stavki, što ih čini primjenjivima u raznim situacijama. Osim toga, tijekom vremena, kako korisnici imaju sve više interakcija s različitim stavkama, sustav postaje precizniji u davanju novih preporuka. Svaka nova recenzija donosi dodatne informacije koje pomažu sustavu da bolje identificira korisničke preferencije, čime se poboljšava učinkovitost i relevantnost preporuka. Dakle, kontinuirana upotreba sustava pridonosi njegovom poboljšanju.

S druge strane, takvi sustavi suočavaju se s problemom hladnog starta prilikom početnih faza korištenja, kada nemaju dovoljno podataka. Naime, bez dovoljnog broja interakcija ili recenzija, sustav se oslanja na ograničene ili potpuno nepostojeće podatke, što otežava davanje relevantne preporuke. Takve preporuke mogu biti generičke ili čak potpuno promašene u odnosu na korisničke interese.

3. Implementacija sustava za davanje preporuka

U sklopu ovog rada, implementirana su dva sustava za davanje preporuka, sustav temeljen na suradnji i sustav temelje na sadržaju. Kao izvor podataka korišten je *MovieLens* [1]skup podataka.

3.1. Analiza skupa podataka

MovieLens skup podataka obuhvaća oko milijun ocjena. Radi se o ocjenama koje je dalo 6000 anonimnih korisnika za približno 4000 različitih filmova. Svaka ocjena prikazuje zadovoljstvo korisnika određenim filmom na ljestvici od 1 do 5, pri čemu više ocjene označavaju veću razinu zadovoljstva.

Skup podataka organiziran je u tri datoteke: *movies*, *ratings* i *users*. Ispisi podataka tih datoteka prikazan su slikama 3.1., 3.2., 3.3. Datoteka movies uključuje podatke o karakteristikama filmovima, poput naslova i žanra. Datoteka users sadrži demografske podatke korisnika, uključujući spol, dob, poštanski broj i zanimanje. Datoteka ratings sadrži identifikatore korisnika i filmova, pripadajuću ocjenu koju je korisnik dodijelio filmu, kao i dodatne informacije poput *timestampa*.

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

Slika 3.1. Datoteka *movies*

[1]`https://www.kaggle.com/datasets/odedgolden/movielens-1m-dataset`

	user_id	gender	zipcode	age_desc	occ_desc
0	1	F	48067	Under 18	K-12 student
1	2	M	70072	56+	self-employed
2	3	M	55117	25-34	scientist
3	4	M	02460	45-49	executive/managerial
4	5	M	55455	25-34	writer

Slika 3.2. Datoteka *users*

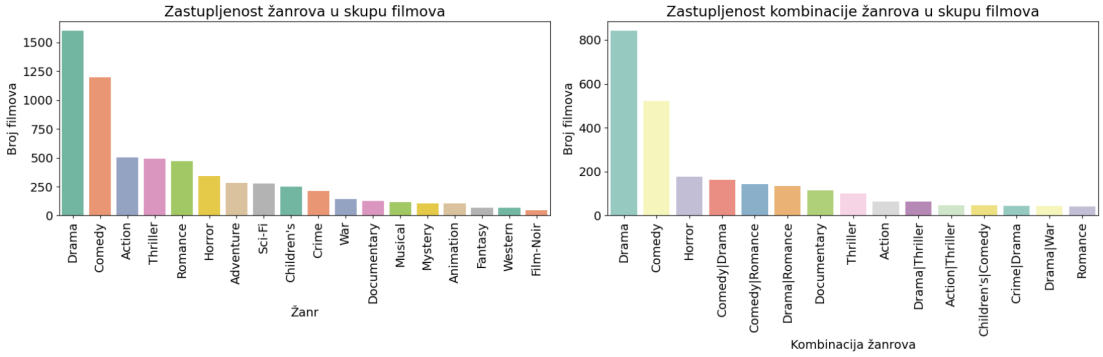
	user_id	movie_id	rating	timestamp
0	1	1193	5	978300760
1	1	661	3	978302109
2	1	914	3	978301968
3	1	3408	4	978300275
4	1	2355	5	978824291

Slika 3.3. Datoteka *ratings*

U ovom poglavlju analiziran je ovaj skup podataka. Glavni cilj analize je razumjeti sadržaj podataka, te procijeniti njihovu raznolikost i količinu. Broj i karakteristike podataka su izuzetno važni za učinkoviti rad sustava za davanje preporuka.

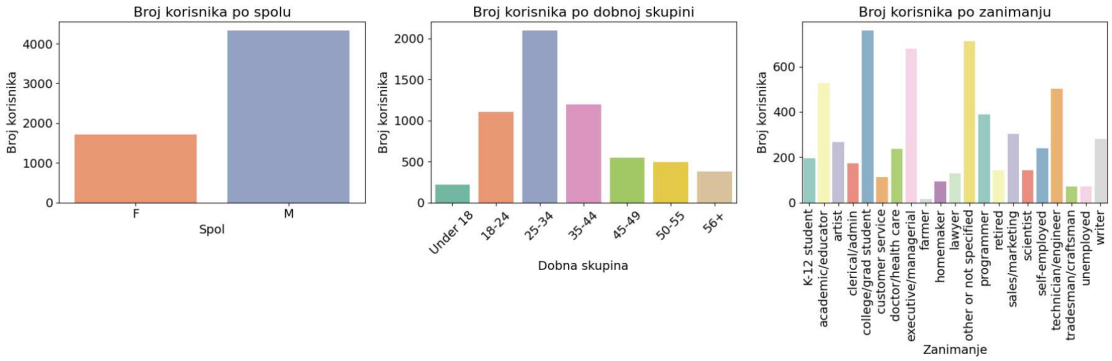
Velika količina podataka važna je karakteristika skupa podataka koji se koristi. Kao i kod svih drugih sustava, veća količina relevantnih podataka omogućuje efikasniji rad. Raznovrsnost podataka, odnosno različiti žanrovi ili karakteristike korisnika, omogućuju sustavu da generira raznolike preporuke. Osim toga, bitna je i ravnoteža u podatcima, prevelika količina određenog žanra ili demografske skupine, može dovesti do pristranih preporuka. Kada podaci nedostaju, primjerice u situaciji kada neki korisnici ocjenjuju samo mali broj filmova, sustavi mogu imati poteškoća s radom. O tome će se govoriti u kasnijim poglavljima.

Slika 3.4. predstavlja analizu podataka iz datoteke *movies*. Važno je napomenuti kako ova analiza prikazuje zastupljenost žanrova u broju dostupnih filmova, a ne njihovu popularnost među korisnicima. Kasnije će se ovi podatci usporediti s podacima o popularnosti žanrova među korisnicima. Drama se ističe kao najzastupljeniji žanr, slijedi ju komedija, dok su ostali žanrovi manje zastupljeni. Kombinacije žanrova pokazuju zanimljivu činjenicu da drama, komedija i horor često stoje kao zasebni žanr. Osim toga, česte kombinacije žanrova su komedije, drame romantike.



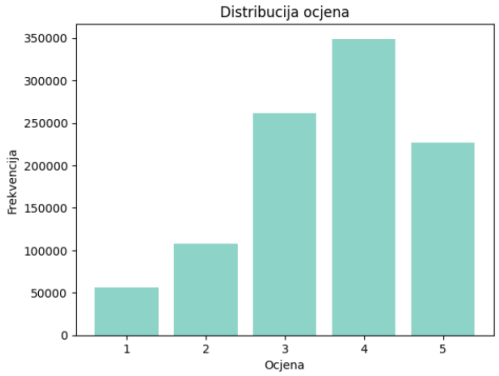
Slika 3.4. Analiza distribucije žanrova

Analizom korisničkih podataka iz datoteke *users* uočava se da su muškarci više zastupljeni među korisnicima te da većina korisnika pripada dobnoj skupini od 25 do 34 godine. Osim toga, zanimljiva je i činjenica da ovaj skup podataka sadrži i zanimanja korisnika, što se može iskoristiti u određenim implementacijama sustava za davanje preporuka. Analiza datoteke *users* prikazana je slikom 3.5.



Slika 3.5. Analiza distribucije korisnika

Slika 3.6. prikazuje distribuciju ocjena. Korisnici češće dodjeljuju više ocjene, pri čemu je ocjena 4 najčešća, a ocjene 1 i 2 znatno manje zastupljene. Ovo može ukazivati na pojavu da korisnici više ocjenjuju one filmove koje im se sviđaju.



Slika 3.6. Analiza distribucije ocjena

Osim toga bitno je napomenuti kako je ocijenjeno oko 95% filmova u ukupnom skupu dostupnih filmova. Distribucija popularnosti žanrova filmova razlikuje se od broja žanrova u skupu podataka. Kao što je ranije spomenuto, drama je najzastupljeniji žanr u skupu filmova, no prema broju ocjena, komedija zauzima prvo mjesto, dok je drama na drugom mjestu. Ostali žanrovi, poput akcije i trilera, značajno su manje zastupljeni od drame, no broj ocjena za te žanrove nije proporcionalno niži. Statistike o broju ocjena

filmova prikazane su u tablici 3.1. U ovu analizu, odnosno tablicu, uključeni su samo filmovi koji imaju barem jednu ocjenu.

Statistika	**Vrijednost**
Najviše ocjena za jedan film	3428
Najmanje ocjena za jedan film	1
Prosječan broj ocjena po filmu	270
Medijan broja ocjena po filmu	124

Tablica 3.1. Statistika ocjenjivanja filmova po korisnicima

Ako se podatci o ocjenama promatraju iz perspektive interakcija korisnika, može se uočiti visoka aktivnost. Statistike o broju ocjena od strane korisnika prikazane su u tablici 3.2. Izuzetno je bitan podatak da je korisnik koji je ocijenio najmanje filmova, ipak ocijenio čak 20 filmova. Broj ocjena prema demografskim karakteristikama prati distribuciju korisnika prema tim karakteristikama. Primjerice, najviše korisnika pripada dobnoj skupini od 25 do 34 godine te upravo iz te skupine dolazi i najviše ocjena.

Statistika	**Vrijednost**
Najviše ocijenjenih filmova od strane jednog korisnika	2314
Najmanje ocijenjenih filmova od strane jednog korisnika	20
Prosječan broj ocjena po korisniku	166
Medijan broja ocjena po korisniku	96

Tablica 3.2. Statistika ocjenjivanja filmova po korisnicima

Analiza skupa podataka pokazuje da su podaci raznovrsni te da postoji dovoljan broj interakcija. Takav skup je temelj implementacije osnovnog sustava za davanje preporuka. Međutim, iako su ovi podaci gotovo idealni, u stvarnom svijetu sustavi često moraju raditi s nepotpunim podacima. To otvara prostor za daljnje istraživanje prilagodljivosti i otpornosti sustava u složenijim uvjetima.

3.2. Metrike za izračun sličnosti

Kao što je već spomenuto, u obje vrste sustava za davanje preporuka, mjerenje sličnosti između stavaka ili klijenata ima veliku važnost. U ovom poglavlju detaljnije se govori o različitim vrstama mjera koje se koriste za određivanje sličnosti. U oba sustava, onom temeljenom na sadržaju i onom temeljenom na suradnji, bit će implementirana mogućnost korištenja kosinusne, euklidske i pearsonove metrike. Kosinusa sličnost je mjera sličnosti između dva vektora koja se temelji na kut između njih. Dakle, ta mjera uspoređuje sličnost dva vektora neovisno o njihovoj veličini, uzimajući u obzir samo orijentaciju. Zbog tih karakteristika, kosinusna mjera sličnosti, otporna je na razlike u skalama vektora koje uspoređuje. Osim toga, kosinusna sličnosti nije osjetljiva na prisutnost ili odsutnost specifičnih vrijednosti. Izraz za dobivanje kosinusne sličnosti prikazan je formulom 3.1

$$\text{Kosinusna sličnost} = \frac{\mathbf{A} \cdot \mathbf{B}}{||\mathbf{A}||||\mathbf{B}||} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2}\sqrt{\sum_{i=1}^{n} b_i^2}} \tag{3.1}$$

Euklidska udaljenost mjeri stvarnu udaljenost između dvije točke odnosno dva vektora prostoru. U ovoj implementaciji, kada se izračuna euklidska udaljenost, ona se normalizira te pretvara u sličnost oduzimanjem od jedinice. Formula 3.2 prikazuje izračun euklidske udaljenosti između dva vektora, normalizaciju udaljenosti te dobivanje mjere sličnosti u rasponu od 0 do 1.

$$\begin{aligned} \text{Euklidska udaljenost} &= ||\mathbf{A} - \mathbf{B}|| = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \\ \text{Euklidska sličnost} &= 1 - \frac{\text{Euklidska udaljenost}}{\text{Maksimalna udaljenost}} \end{aligned} \tag{3.2}$$

Pearsonova korelacija mjeri linearni odnos između dva vektora podataka. Dakle, mjeri koliko dva vektora imaju sličan obrazac vrijednosti. Ova metrika pokazuje pokazuje kako se vrijednosti komponente jednog vektora mijenjaju u odnosu na odgovarajuću komponentu drugog vektora. Vrijednost pearsonove korelacije ide od jedan do minus jedan, gdje jedan označava savršenu pozitivnu korelaciju a minus jedan savršenu

negativnu. Izračun Pearsonovog koeficijenta korelacije prikazan je formulom 3.3

$$\text{Pearsonov koeficijent korelacije} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.3}$$

Utjecaj ovih mjera na skup podataka i njihova primjena u različitim implementacijama sustava bit će obrađen u kasnijim poglavljima. U njima će se također analizirati i uspoređivati rezultati različitih sustava za davanje preporuka.

3.3. Implementacija sustava temeljenog na sadržaju

Prilikom implementacije sustava temeljenog na sadržaju, koristile su se informacije iz podataka koji opisuju filmove. Ovi podaci omogućili su izgradnju modela koji koristi sadržaj i daje preporuke na temelju sličnosti između različitih filmova. Kao što je već spomenuto, riječ „sadržaj“ u ovom kontekstu odnosi se na karakteristike odnosno atribute filmova.

Sustav temeljen na sadržaju obično se sastoji od dva glavna dijela, izrade korisničkog profila i izračuna sličnosti između sadržaja. Izrada korisničkog profila sastoji se od analize povijesti korisnika kako bi se razumjele njegove sklonosti. Podaci o stavkama koje je korisnik ocijenio visokom ocjenom se prikupljaju te se koriste za izradu profila interesa. Takav profil sadrži ključne informacije o ukusu tog korisnika.

U slučaju ove implementacije takvog sustava, za ciljanog korisnika identificiraju se visoko ocijenjeni filmovi. Zanimljivi su filmovi kojima je korisnik dodijelio ocjenu 4 ili više. Kako bi se dodatno poboljšala relevantnost preporuka, daje se prednost filmovima koji su nedavno gledani i ocijenjeni. Ova vremenska komponenta ostvaruje se uz pomoć parametra koji koristi *timestampove* u korisnikovoj povijesti ocjenjivanja kako bi odredio koliko je svaki film nedavno ocijenjen. Ta ocjena novosti, odnosno *Recency score*, normalizira se na skalu od 0 do 5, gdje 0 označava najstarije recenzije, a 5 najnovije. *Recency score* se zatim množi faktorom 0.5 te se kombinira s originalnom ocjenom korisnika, stvarajući ponderiranu ocjenu *Weighted rating*. Dobivanje *Weighted ratinga* prikazano je formulom 3.4

$$\text{Weighted\_rating} = \text{Rating} + 0.5 \times \text{Recency\_score} \tag{3.4}$$

Nakon što se izradi korisnički profil, potrebno je izračunati sličnosti između sadržaja u korisničkom profilu i drugih dostupnih sadržaja. U skupu podataka dostupni su žanrovi, te je odlučeno koristiti ih kao ključnu karakteristiku za usporedbu filmova. Žanrovi su prepoznatljive značajke koje karakteriziraju sadržaj svakog filma pa se usporedbom skupova žanrova može utvrditi koliko su teme pojedinih filmova slične. Dakle, uspoređuju se cijeli skupovi žanrova koji opisuju filmove. Primjerice, ako je film A opisan žanrovima akcija, triler, drama, a film B žanrovima akcija i triler, tada će se ti filmovi smatrati relativno sličnima. S druge strane, film C koji sadrži žanrove komedija i romantika bit će procijenjen kao manje sličan filmovima A i B. Filmovi čiji su skupovi žanrova slični mogu se smatrati relevantnijima za korisnike koji su već pokazali interes za filmove s takvom kombinacijom žanrova.

Dakle, u ovoj implementaciji žanrovi se koriste kako bi prikazali karakteristike odnosno sadržaj određenog filma. Iz tog razloga mjeri se upravo sličnost između skupova žanrova kao bi se pokazala sličnost između filmova. Proces mjerenja sličnosti započinje korištenjem funkcije koja generira kombinacije žanrova filmova te se zatim svaki film prikazuje odgovarajućim vektorom.

Prilikom procesa vektorizacije, važno je uzeti u obzir popularnost žanrova. Ideja je da rijetki žanrovi mogu bolje reflektirat specifične ukusa korisnika. S druge strane popularni žanrovi su široko zastupljeni u korisničkim profilima te stoga i manje učinkoviti za definiranje preferencija. U ovoj implementaciji, manje generički žanrovi imaju veću težinu u procesu preporučivanja. To se postiže uporabom TF-IDF vektorizacije. Izraz za dobivanje TF-IDF vekora prikazan je formulom 3.5

$$\text{tf-idf}_{i,j} = \text{tf}_{i,j} \cdot \log \frac{N}{\text{df}_i} \tag{3.5}$$

$\text{tf}_{i,j}$ = frekvencija pojave i u unosu j
df_i = broj unosa koji sadrže i

N = ukupan broj unosa

Formula prikazuje učestalost pojavljivanja pojmova, odnosno broj puta koliko se određeni pojam pojavljuje u jednom unosu, skaliran u odnosu na ukupnu zastupljenost tog pojma u svim unosima. Što je određeni pojam manje zastupljen u cjelokupnom skupu filmova, to mu je dodijeljena veća težina. Korištenje logaritma omogućuje zaglađivanje rezultata dijeljenja.

U implementaciji se funkcija TfidfVectorizer koristi za pretvaranje tekstualnih podataka o filmovima, konkretno žanrova, u numeričke vektore. U slučaju ove implementacije računa se koliko se često određeni žanr pojavljuje u pojedinom filmu te koliko je rijedak taj žanr u cijelom skupu filmova. Na taj način, TF-IDF omogućuje da rjeđi žanrovi dobiju veću težinu i tako budu bitniji. Na kraju tog procesa dobiveni su vektori koji predstavljaju svaki film. Ti vektori su zapravo redci matrice prkazane na slici 3.7. i služe kao osnova za račun sličnosti između filmova.

	(Action,)	(Action, Adventure)	(Action, Adventure, Animation)	(Action, Adventure, Children's)	(Action, Adventure, Comedy)	(Action, Adventure, Crime)	(Action, Adventure, Drama)	(Action, Adventure, Fantasy)	(Action, Adventure, Horror)	(Action, Adventure, Mystery)	...
title											
Heat (1995)	0.24	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
Sudden Death (1995)	1.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
GoldenEye (1995)	0.26	0.37	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
Cutthroat Island (1995)	0.24	0.34	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
Money Train (1995)	1.00	0.00	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...
...	...	...	...	...	...	...	...	...	...	...	...

Slika 3.7. Matrica filmova i žanrova

Za izračunavanje sličnosti koristi se više metrika, kosinusna, euklidska i pearsonova. Način rada ovih metrika detaljno je opisan u prethodnom poglavlju, dok će razlike i sličnosti u konkretnoj primjeni biti analizirane u sljedećem poglavlju. Rezultat koraka izračuna sličnosti jest matrica sličnosti iz koje se izravno generiraju preporuke.

Preporuke se generiraju uz pomoć već spomenutog korisničkog profila na način da se za filmove koje je korisnik visoko ocijenio, pronalaze slični filmovi koje korisnik još nije pogledao. U slučaju ove implementacije za najbolja tri filma generira se pet preporuka za prvi, četiri za drugi i tri za treći film iz korisničkog profila.

Rezultat takvog sustava prikazan je na slici 3.8. Ispis prikazuje tri najrelevantnija odnosno najbolje ocijenjena filma za određenog korisnika te dane preporuke na temelju tih filmova. Osim toga ispis sadrži informaciju o filmu na kojem se preporuka temelji kao i mjeru sličnosti između tog i preporučenog filma.

Top 3 filma

	title	movie_id	genres	rating	recency_score	weighted_rating
663476	Thin Red Line, The (1998)	2427	Action|Drama|War	5.0	4.904249	7.452125
485996	Big Lebowski, The (1998)	1732	Comedy|Crime|Mystery|Thriller	5.0	4.256794	7.128397
13796	City of Lost Children, The (1995)	29	Adventure|Sci-Fi	5.0	3.641255	6.820627

	movie_title	genres	similarity_score	from_movie_id
0	Glory (1989)	Action|Drama|War	1.000000	2427
1112	Heaven & Earth (1993)	Action|Drama|War	1.000000	2427
1217	Flying Tigers (1942)	Action|Drama|War	1.000000	2427
1318	Braveheart (1995)	Action|Drama|War	1.000000	2427
3761	Fighting Seabees, The (1944)	Action|Drama|War	1.000000	2427
3862	Arsenic and Old Lace (1944)	Comedy|Mystery|Thriller	0.636177	1732
4534	Goodbye, Lover (1999)	Comedy|Crime|Thriller	0.599745	1732
4559	Dead Men Don't Wear Plaid (1982)	Comedy|Crime|Thriller	0.599745	1732
4926	Lock, Stock & Two Smoking Barrels (1998)	Comedy|Crime|Thriller	0.599745	1732
5551	Fantastic Voyage (1966)	Adventure|Sci-Fi	1.000000	29
5899	Supernova (2000)	Adventure|Sci-Fi	1.000000	29
6060	Mighty Peking Man (Hsing hsing wang) (1977)	Adventure|Sci-Fi	1.000000	29

Slika 3.8. Preporuke sustava temeljenog na sadržaju

3.4. Implementacija sustava temeljenog na suradnji

Kao što je već spomenuto, glavna podjela sustava za davanje preporuka je na sustave temeljene na sadržaju i sustave temeljene na suradnji. U prošlom poglavlju riječ je bila o implementaciji sustava temeljenog na sadržaju, gdje su preporuke temeljene na karakteristikama odnosno sadržaju filmova, poput žanrova i ključnih riječi. U ovom poglavlju naglasak je stavljen na sustave temeljene na suradnji.

Sustavi temeljeni na suradnji, poznati i kao kolaborativno filtriranje temelje se na informacijama o interakcijama korisnika i stavaka. Iz tih informacija identificiraju se slični korisnici i obrasci ponašanja te oni služe kao temelj za davanje preporuka. U slučaju ove implementacije takvog sustava, kao mjera interakcije korisnika i proizvoda koriste se ocjene koje je korisnik dodijelio filmovima.

Rad sustava temeljenog na suradnji započinje kreiranjem matrice gdje je svaki korisnik predstavljen retkom, dok su filmovi koje je ocijenio raspoređeni po stupcima. U usporedbi s osnovnom matricom iz prošlog sustava, onog temeljenog na sadržaju, gdje su bili predstavljeni filmovi i žanrovi, ovdje je naglasak na korisnicima i njihovim ocje-

nama. U prvom pristupu sličnost se računa na temelju karakteristika filmova, dok se ovdje određuje sličnost između korisnika.

U ovoj implementaciji, korisnici su predstavljeni kroz vektore koji sadrže njihove ocjene. Matrica takvih vektora prikazana je slikom 3.9.

movie_id	1	2	3	4	5	6	7	8	9	10	...	3943	3944
user_id													
1	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
6	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
8	4.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
9	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
10	5.0	5.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	...	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...
6030	0.0	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
6032	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0
6035	4.0	0.0	1.0	2.0	1.0	0.0	3.0	0.0	0.0	0.0	...	0.0	0.0

Slika 3.9. Matrica korisnika i ocjena filmova

Sličnost se izračunava između tih vektora, čime se identificiraju korisnici sa sličnim ukusom. Sličnost se mjerila kosinusnom, euklidskom i pearsonovom metrikom.

Sustav zatim identificira ciljanog korisnika, onog kojem se daju preporuke te sortira ostale korisnike na temelju prethodno izračunate matice sličnosti. Odabire se 10 najsličnijih korisnika.

Za svakog od sličnih korisnika, identificiraju se filmovi koje su pogledali i računa se prosječna ocjena. Filmovi s najboljom prosječnom ocjenom bit će preporučeni ciljanom korisniku, uz uvjet da ih on nije pogledao.

Bitno je napomenuti kako nule imaju veliku ulogu u ovakvom računanju prosječne ocijene za davanje preporuka. Naime ako se nule ne ubroje, često će biti preporučeni filmovi koji imaju visoku recenziju od samo jednog korisnika dok ih drugi uopće nisu pogledali. S druge strane, prosječna vrijednost koja ubraja nule može stvoriti lažni dojam loše procijenjene ocijene. Iz tog razloga, ispisuje se rezultat koji sadrži informaciju o ocjenama koje su slični korisnici dali za svaki film.

Rezultat dobiven sustavom temeljenim na suradnji prikazan je na slici 3.10. Slika sadrži preporuke filmova te odgovarajuće žanrove kako bi se rezultat mogao usporediti s onim koji je generirao sustavom temeljenim na suradnji. Osim toga, ispisuje se srednja

vrijednost ocjena koje su dali slični korisnici, broj nula u ocjenama te popis tih ocjena, kako bi se moglo bolje usporediti djelovanje različitih metrika.

	title	genres	predicted_rating	zero_ratings_count	non_zero_ratings
0	Boogie Nights (1997)	Drama	4.4	0	[5, 5, 4, 5, 4, 4, 3, 4, 5, 5]
1	Leaving Las Vegas (1995)	Drama|Romance	4.1	0	[3, 4, 5, 5, 5, 4, 4, 4, 4, 3]
2	Simple Plan, A (1998)	Crime|Thriller	4.0	0	[3, 4, 5, 4, 4, 5, 5, 3, 3, 4]
3	Fugitive, The (1993)	Action|Thriller	4.0	0	[4, 4, 4, 5, 3, 3, 4, 4, 4, 5]
4	What's Eating Gilbert Grape (1993)	Drama	3.9	1	[4, 4, 5, 4, 5, 5, 4, 3, 5]
5	Shakespeare in Love (1998)	Comedy|Romance	3.8	1	[4, 3, 5, 4, 4, 5, 5, 3, 5]
6	Crying Game, The (1992)	Drama|Romance|War	3.8	1	[3, 4, 5, 4, 4, 4, 5, 4, 5]
7	Groundhog Day (1993)	Comedy|Romance	3.7	1	[5, 4, 5, 3, 4, 5, 4, 4, 3]
8	Grifters, The (1990)	Crime|Drama|Film-Noir	3.5	2	[4, 5, 5, 5, 4, 4, 3, 5]
9	Schindler's List (1993)	Drama|War	3.4	2	[4, 5, 4, 4, 3, 4, 5, 5]

Slika 3.10. Preporuke sustava temeljenog na suradnji

3.5. Usporedba rezultata

Ko što je već spomenuto u ovom radu, za implementaciju osnovnih sustava za davanje preporuka koriste se kosinusna, euklidska i pearsonova metrika. Te metrike implementirane su prilikom izračuna sličnosti u sustavu temeljenom na sadržaju i sustavu temeljenom na suradnji. U ovom poglavlju usporedit će se rezultati dobiveni primjenom različitih metrika i dviju različitih implementacija sustava, jedne temeljene na suradnji, a druge na sadržaju.

3.5.1. Usporedba rezultata sustava temeljenih na sadržaju i suradnji

Generalna razlika između rezultata dobivenih sustavom temeljenim na sadržaju i sustavom temeljenim na suradnji jest specifičnost preporuka.

Top 3 filma

	title	movie_id	genres	rating	recency_score	weighted_rating
144042	Searching for Bobby Fischer (1993)	529	Drama	5.0	4.686641	7.343321
162472	Dances with Wolves (1990)	590	Adventure|Drama|Western	5.0	4.329302	7.164651
776394	Three Kings (1999)	2890	Drama|War	5.0	3.936229	6.968114

	movie_title	genres	similarity_score	from_movie_id
0	12 Angry Men (1957)	Drama	1.000000	529
616	Two Jakes, The (1990)	Drama	1.000000	529
678	Onegin (1999)	Drama	1.000000	529
697	Whole Wide World, The (1996)	Drama	1.000000	529
747	Emperor and the Assassin, The (Jing ke ci qin ...	Drama	1.000000	529
791	True Grit (1969)	Adventure|Western	0.621807	590
1111	Tashunga (1995)	Adventure|Western	0.621807	590
1112	Shane (1953)	Drama|Western	0.538572	590
1417	Geronimo: An American Legend (1993)	Drama|Western	0.538572	590
1504	Truce, The (1996)	Drama|War	1.000000	2890
1506	Apocalypse Now (1979)	Drama|War	1.000000	2890
2682	Paths of Glory (1957)	Drama|War	1.000000	2890

Slika 3.11. Preporuke sustava temeljenog na sadržaju

	title	genres	predicted_rating	zero_ratings_count	non_zero_ratings
0	Pulp Fiction (1994)	Crime|Drama	4.1	1	[5, 5, 5, 4, 5, 4, 4, 5, 4]
1	Braveheart (1995)	Action|Drama|War	3.2	3	[4, 4, 5, 4, 5, 5, 5]
2	Terminator 2: Judgment Day (1991)	Action|Sci-Fi|Thriller	3.0	3	[4, 3, 4, 5, 5, 4, 5]
3	Game, The (1997)	Mystery|Thriller	2.9	3	[4, 3, 4, 4, 5, 5, 4]
4	Forrest Gump (1994)	Comedy|Romance|War	2.7	4	[5, 4, 4, 4, 5, 5]
5	Being John Malkovich (1999)	Comedy	2.3	4	[4, 3, 5, 4, 5, 2]
6	Patriot, The (2000)	Action|Drama|War	2.2	4	[5, 2, 4, 4, 3, 4]
7	Hunt for Red October, The (1990)	Action|Thriller	2.1	5	[5, 5, 3, 4, 4]
8	American History X (1998)	Drama	2.1	5	[4, 4, 4, 5, 4]
9	Gladiator (2000)	Action|Drama	2.1	5	[5, 4, 4, 5, 3]

Slika 3.12. Preporuke sustava temeljenog na suradnji

Rezultati sustava temeljenog na sadržaju i suradnji za nasumičnog korisnika prikazani su na slici 3.11. i 3.12.

Sustav temeljen na sadržaju pruža usmjerene i precizne preporuke koje se oslanjaju na karakteristike filmova koje je korisnik već pogledao, što ga čini vrlo specifičnim. U rezultatima sustava temeljenog na sadržaju nije se pojavio novi žanr koji nije jednak onome što je korisnik već gledao. To može biti prednost jer osigurava da preporuke budu relevantne, ali istovremeno može postati ograničenje jer korisnik možda neće dobiti priliku otkriti nešto izvan svojih uobičajenih interesa.

S druge strane, sustav temeljen na suradnji pruža širi spektar preporuka. Sustav je preporučio puno različitih žanrova, uključujući komediju, triler, kriminalističke filmove kojih nisu tipični za korisnikove uobičajene preferencije. Međutim, sustav ne prepoznaje specifične karakteristike koje su očito vrlo bitne, primjerice to da korisnik voli *Westerne*. Ovaj pristup omogućava veću raznolikost, ali ponekad može rezultirati manje relevantnim preporukama.

Nakon što su razmotrene generalne razlike između sustava temeljenog na sadržaju i sustava temeljenog na suradnji, u nastavku će se usporediti rezultati dobiveni primjenom različitih metrika.

3.5.2. Usporedba rezultata različitih metrika

Bitno je napomenuti da se u sustavima za davanje preporuka koristi rangiranje sličnosti, što znači da, iako se primjenjuju različite mjere sličnosti, one mogu dovesti do istih rezultata kada je riječ o preporukama. To je uočeno u ovoj implementaciji sustava temeljenog na sadržaju. Različite mjere daju isto rangiranje jer za svaki film postoji mogućnost da

se nađu drugi filmovi s gotovo identičnom raspodjelom žanrova. Ako su brojčane vrijednosti dovoljno slične i obrasci u vektorima slični, rangiranje će biti isto neovisno o metodi mjerenja sličnosti.

Iako i ostale mjere daju slične rezultate rangiranja u sustavu temeljenom na sadržaju, kosinusna mjera se pokazuje kao najprimjerenija zbog otpornosti na nule. Naime, u implementaciji sustava temeljenog na sadržaju, filmovi su predstavljeni vektorima koji odražavaju njihove žanrove, pri čemu su vrijednosti različite od nule rijetke. Određeni film opisuje vrlo mali broj žanrova u usporedbi s ukupnim brojem mogućim žanrovima te su zbog toga vektori većinom ispunjeni nulama. S obzirom na to da se kosinusna metrika dobro nosi s problemom nula, prigodna je za ovaj slučaj.

U slučaju sustava temeljenog na suradnji, svaka metrika ima svoje pogodnosti i mane. Kosinusna vrijednost primjerice ne računa apsolutnu razliku između vektora već gleda samo proporcionalnosti. Iz tog razloga, ako gledamo primjer situacije u kojoj je jedan korisnik filmovima dao ocijene [1,1,1,1] a drugi korisnik ocijene [5,5,5,5], kosinusa sličnost dat će visoku vrijednost. Ta sličnost ne odražava razliku u apsolutnim vrijednostima tih ocjena, već samo bilježi obrazac, što je problematično jer takvi korisnici mogu imati zaista različite preferencije. U tom slučaju, euklidska udaljenost je pogodnija metrika.

S druge strane, euklidska udaljenost ne radi dobro kada su podatci rijetki. Naime, ta metrika ne uzima u obzir udaljenosti među vrijednostima koje su nula, što je problem kada korisnici imaju puno neocijenjenih filmova. Korištenje euklidske udaljenosti može dovesti do lažne sličnosti u slučaju kada dva korisnika imaju vrlo različite obrasce ocjenjivanja ali imaju puno neocijenjenih filmova.

Pearsonova korelacija mjeri linearnu korelaciju između dviju varijabli. U sustavima preporuka, ova korelacija može identificirati slične korisnike jer prepoznaje linearne obrasce u njihovom ocjenjivanju. No, takav pristup loše prepoznaje nelinearne povezanosti. Dakle čak i kada postoji jasna veza, ako ona nije linearna, pearsonov koeficijent bit će nizak.

Prilikom implementacije oba sustava primijećene su prednosti i nedostaci svake metrike, ali i samih sustava. Sustav temeljen na sadržaju i sustav temeljen na suradnji djeluju iz potpuno različitih perspektiva kada se primjenjuju na ovaj skup podataka o filmo-

vima. Snage i problemi ovih pristupa bit će korišteni u daljnjem razvoju sustava preporuka, koji će nastojati ujediniti najbolje karakteristike oba pristupa kako bi se poboljšala točnost i relevantnost preporuka.

4. Problem hladnog starta

U stvarnom svijetu je čest slučaj da sustavi za davanje preporuka ne rade u idealnim uvjetima. Upravo u takvim situacijama pokreću se pitanja o tome koliko su sustavi sposobni prilagoditi se i izdržati izazove u različitim uvjetima. Daljnja analiza usredotočit će se upravo na takve scenarije.

Hladni start je pojam za situaciju u kojoj sustavi nemaju dovoljno podataka za efikasno funkcioniranje. U kontekstu sustava za davanje preporuka, to znači da sustav ne raspolaže dovoljnom količinom informacija kako bi mogao davati relevantne preporuke.

Problem hladnog starta je posebno čest i izražen kada se pokreće potpuno novi sustav za davanje preporuka. U početnim fazama, rada sustav posjeduje iznimno malo podataka o korisnicima, stavkama ili međusobnim interakcijama. Vrlo često, takav sustav zapravo treba sakupiti informacije od nule, što može biti vrlo izazovno. U tim situacijama koriste se strateški pristupi poput prikupljanja osnovnih informacija od korisnika prilikom početka korištenja. Primjerice, u slučaju sustava za davanje preporuka filmova, moguće je korisniku ponuditi da odabere omiljene žanrove ili omogućiti mu da ocijeni već neke pogledane filmove. Osim toga, sustav za početak može preporučiti generalno popularne ili dobro ocijenjene filmove. Za tako nešto ipak su potrebni podatci iz nekog oblika vanjskog izvora. To su samo neke od mogućnosti nošenja s problemom hladnog starta u potpuno novim sustavima.

Osim hladnog starta cijelog sustava, kada se gotovo uopće ne posjeduju potrebni podatci, bitno je spomenuti još dvije vrste hladnog starta. Hladni start korisnika i hladni start proizvoda odnose se na situacije kada sustav ima određene podatke, ali nema konkretno informacije o specifičnom korisniku ili predmetu.

Hladni start korisnika obično se dešava kada novi korisnik počne koristiti sustav, a

sustav nema informacije o njegovim obrascima ponašanja i interesima. Primjerice, u slučaju implementiranog sustava za davanje preporuka to bi bio korisnik koji nije dao ocjenu niti jednom filmu.

S druge strane hladni start proizvoda nastaje kada se u sustav dodaje nova stavka te za nju ne postoje nikakve povratne informacije. Za naš sustav to bi bilo dodavanje filma koji je potpuno nov te stoga nema nikakvu ocjenu.

U nastavku će se govoriti o načinima nošenja s određenim problemima hladnog starta. Osim toga, bit će riječi o tome kako sustavi funkcioniraju nakon početne faze hladnog starta, kada postoje neki podaci, no oni i dalje nisu dovoljni za potpuno funkcionalne preporuke.

Podatci iz *MovieLense* seta podataka vrlo su opsežni. Prosječan broj ocjena po filmu iznosi 270, dok prosječan broj ocjena koje je jedan korisnik dodijelio iznosi 166. Bitno je napomenuti kako najmanji broj filmova ocijenjenih od strane jednog korisnika iznosi 20, što je dovoljno za efikasan rad. S druge strane, postoje filmovi koji imaju vrlo mali broj ocjena pa čak i oni koji nisu dobili niti jednu ocjenu. Iz tog razloga, prvo će biti riječ o tome kako optimizirati sustav za davanje preporuka tako da radi sa filmovima koji nemaju ocijene, odnosno kako riješiti problem hladnog starta proizvoda.

4.1. Hladni start proizvoda

Hladni start proizvoda odnosi se na problem kada sustav nema dovoljno podataka o određenom proizvodu te iz tog razloga ne može generirati preporuke. Kod sustava za preporuke filmova, hladni start smatra se dodavanje novog filma u sustav. Prvi korak dodavanja filma je ispravno označavanje značajki i karakteristika, u slučaju ove implementacije to su žanrovi. Nakon što je film dodan u bazu podataka, moguće je i za njega generirati preporuke.

U sustavima temeljenim na sadržaju

Sustavi temeljeni na sadržaju generalno imaju vrlo malo problema s hladnim startom. Implementirani sustav kreira korisnički profil na temelju omiljenih filmova korisnika te zatim preporučuje filmove koji imaju slične žanrove. U kontekstu hladnog starta, takav

sustav neće uključiti novi film u korisnički profil jer takav film nema recenzija. Međutim, potpuno novi film može biti preporučen ako je žanrovski sličan filmovima koji su već dio korisničkog profila. U ovom slučaju, hladni start zapravo ne predstavlja značajan problem. Ipak, vjerojatnost da novi film bude preporučen vrlo je mala, jer u bazi postoji puno filmova koji su žanrovski slični. Primjerice, ako se u korisničkom profilu nalazi žanr komedija, a u bazi postoji petstotinjak filmova upravo tog žanra, vjerojatnost da sustav preporuči novi film postaje izuzetno mala zbog velikog broja sličnih opcija. Iz tog razloga u implementaciju je dodana mala prednost za nove filmove.

Prilikom generiranja preporuka, nasumično se odabire 10% negledanih filmova koji dobivaju prednost. Ako postoji više filmova istog žanra koji se mogu preporučiti, a među njima je i novi film, on će imati veću šansu za preporuku zahvaljujući ovoj prednosti. Na slici 4.1. prikazane su preporuke za nasumično odabranog korisnika, gdje se jasno vidi prednost koja je dana negledanim filmovima. Film *Hi-Yo Silver* nema niti jednu recenziju jer je novi film u sustavu. Međutim, budući da je nasumično odabran, dobio je prednost u procesu generiranja preporuka. Zbog te prednosti, film se prvi preporučuje korisniku.

Top 3 filma

	title	movie_id	genres	rating	recency_score	weighted_rating
356632	Unforgiven (1992)	1266	Western	5.0	4.999999	7.499999
77532	Pulp Fiction (1994)	296	Crime|Drama	5.0	4.999992	7.499996
612716	Indecent Proposal (1993)	2269	Drama	5.0	4.999938	7.499969

	movie_title	genres	similarity_score	from_movie_id
0	Hi-Yo Silver (1940)	Western	1.1	1266
1	Cimarron (1931)	Western	1.0	1266
30	Duel in the Sun (1946)	Western	1.0	1266
102	Alvarez Kelly (1966)	Western	1.0	1266
130	Wild Bill (1995)	Western	1.0	1266
171	Blue Collar (1978)	Crime|Drama	1.0	296
227	Grand Canyon (1991)	Crime|Drama	1.0	296
573	Twilight (1998)	Crime|Drama	1.0	296
626	Turn It Up (2000)	Crime|Drama	1.0	296
633	Crows and Sparrows (1949)	Drama	1.1	2269
634	Empty Mirror, The (1999)	Drama	1.1	2269
635	Phantom Love (Ai No Borei) (1978)	Drama	1.1	2269

Slika 4.1. Preporuke sustava temeljenog na sadržaju

U sustavima temeljenim na suradnji

Za razliku od sustava temeljenog na sadržaju, sustavi temeljeni na suradnji imaju problema s hladnim startom proizvoda. Implementirani sustav temeljen na suradnji oslanja se na analizu povijesnih podataka o interakcijama korisnika s proizvodima što su u tom slučaju ocijene dodijeljene filmovima. Kad se pojavi novi film, on još nije ocijenjen te

dolazi do problema. Nedostatak podataka o interakcijama s filmom onemogućava uključivanje tog novog filma u proces generiranja preporuka.

Proces pronalaženja sličnih korisnika temelji se na analizi obrazaca ocjena. Prvo se identificiraju korisnici koji imaju sličan obrazac ocjenjivanja kao i ciljani korisnik za kojeg se generiraju preporuke. Zatim se analiziraju filmovi koje slični korisnici ocjenjuju, ali koje ciljani korisnik još nije pogledao, i na temelju toga sustav daje preporuke. Zbog nedostatka ocjena, novi filmovi nikako ne mogu ući u proces analiziranja filmova koje su slični korisnici ocijenili. On može biti predložen jedino kada se nađe slični korisnik koji ga je pogledao.

Iz tog razloga, rješavanje problema hladnog starta u sustavima temeljenim na suradnji zahtijeva prikupljanje ocjena za filmove na neki način izvan samog algoritma. Postoji mnogo metoda kojima se mogu preporučiti novi filmovi, primjerice koristeći postupak automatske preporuke novih filmova. Međutim, takav pristup ne uzima u obzir individualne preferencije korisnika te se udaljava od ideje personalizacije. Ideja za rješavanje ovog problema temelji se na integraciji sustava temeljenog na suradnji sa sustavom temeljenim na sadržaju koji nema problema s hladnim startom. Ta tema bit će detaljnije obrađena u idućem poglavlju gdje će biti riječ o hibridnom sustavu.

4.2. Hladni start korisnika

Situacija u kojoj sustav za preporuke nema dovoljno podataka o korisniku kako bi dao učinkovite preporuke naziva se hladni start korisnika. Ovaj problem najčešće se javlja kada novi korisnik počne koristiti platformu. Tada korisnik još nema interakcija s proizvodima što dovodi do toga da algoritam ne može generirati relevantne preporuke. U slučaju implementacija sustava za preporuke filmova, problem hladnog starta korisnika javio bi se kada bi se našao korisnik koji nije ocijenio niti jedan film.

Kao što je već spomenuto u analizi podataka, skup podataka *MovieLense* vrlo je pogodan jer svi korisnici imaju barem dvadeset recenzija. Iz tog razloga niti jedna implementacija sustava za davanje preporuka neće imati problema kada se koristi *MovieLense* skup podataka. No ipak, u stvarnom svijetu skupovi podataka nisu toliko idealni te pojava novih korisnika je ključan dio rasta platforme. Stoga je problem hladnog starta zasigurno

aspekt o kojem treba razmišljati pri svakoj implementaciji.

Za potrebe poboljšavanja implementacija sustava za preporuke, modificiran je *MovieLense* skupa podataka te su iz njega obrisane neke interakcije, tako da neki korisnici nemaju nikakvih recenzija. Na taj način generiran je skup podataka koji prikazuje situaciju u kojoj bi se sustav našao u problemu hladnog starta korisnika.

Korisnicima koji imaju identifikator u rasponu od jedan do pet obrisane su recenzije. Bitno je napomenuti kako oni nisu potpuno isključeni iz skupa podataka. Naime, pretpostavka je da su podatci o korisniku iz datoteke *users* poznati. Ti podaci ne sadrže informacije o interakcijama korisnika s filmovima, već samo njihove karakteristike kao što su spol, godine i zanimanje. Takvi podaci mogli su biti prikupljeni, primjerice, tijekom registracije korisnika, odnosno prije samog korištenja sustava. Postavlja se pitanje kako implementirani sustavi rade s ovakvim podatcima.

Implementirani sustavi temeljeni na sadržaju i suradnji suočavaju se s izazovima kada je riječ o novim korisnicima koji nemaju nikakve podatke o recenzijama. Sustav temeljen na sadržaju započinje generiranje preporuka tako što kreira korisnički profil. Takav profil kreira se na temelju filmova koje je korisnik ocijenio ocjenom višom od 4. Međutim, ako korisnik nema niti jednu recenziju, sustav nije u mogućnosti generirati profil te ne može ni započeti proces davanja preporuke. U takvoj situaciji sustav ne generira nikakve preporuke, već prikazuje poruku da korisnik nema pregledanih filmova.

Sustav temeljen na suradnji radi tako da pokušava pronaći korisnike koji imaju sličan obrazac ocjenjivanja kao ciljani korisnik. No, kod korisnika bez recenzija, ovaj pristup ne funkcionira ispravno. Takav sustav sve korisnike bez recenzija tretira na isti način te im pruža generičke, odnosno netočne preporuke. Greška se javlja u računu sličnosti između praznog vektora ciljanog korisnika i vektora drugih korisnika. Različite metrike sličnosti daju različite rezultate jer se temelje na različitim formulama, no svi su ti rezultati pogrešni zbog nedostatka podataka. Prema takvom izračunu, svi korisnici ispadaju jednako slični ciljanom korisniku, što rezultira neutemeljenim odnosno nasumičnim preporukama.

U oba sustava, onom temeljenom na sadržaju te onom temeljenom na suradnji, problem hladnog starta korisnika je izuzetno velik izazov. Potpuni nedostatak podataka

ne samo da otežava, već u potpunosti onemogućuje generiranje relevantnih preporuka. Kada korisnik prikupi barem nekoliko ocjena, problem se može ublažiti. No, dok potpuno nedostaju recenzije, jedino rješenje je da korisnik počne pružati povratne informacije. Kao što je ranije spomenuto, prikupljanje nekih povratnih informacija o interesima može se ostvariti na mnogo različitih načina. Primjerice, jedan od njih uključuje inicijalni upitnik. Međutim, ako je pretpostavka da te informacije nisu dostupne, postoje određena tehnička rješenja koja nisu idealna, ali mogu ublažiti problem hladnog starta korisnika.

4.2.1. Preporuke temeljene na popularnosti

Jedna od ideja generiranja preporuka za potpuno nove korisnike su preporuke temeljene na popularnosti. Naravno takve preporuke su nepersonalizirane i generičke ali mogu pomoći u smanjenju problema hladnog starta. Takve preporuke mogu biti temeljene samo na broju ocjena koje neki film ima, što nije uvijek idealno, jer broj pregleda ne odražava nužno kvalitetu. Osim toga, preporuke mogu biti temeljene na prosječnoj ocjeni, što opet može biti problematično jer film s jednom ocjenom 5 može biti rangiran vrlo visoko. Rješenje je kombinirati oba pristupa. Tako se korisnicima pružaju preporuke za filmove koje su često gledani i dobro ocijenjeni. U ovom radu implementiran je jednostavan algoritam koji stvara takvu kombiniranu ocjenu filmova. Ova ocjena sastoji se od kombinacije 90% korisničkih ocjena i 10% broja pregleda, normaliziranih na skalu od 1 do 5.

Osim takve vrste sustava temeljenih na popularnosti, u ovom radu istražena je i opcija koja analizira koji su filmovi najpopularniji baš u kontekstu početnih odabira korisnika. Korištenjem *timestamp* podataka, za svakog korisnika određeno je koji je bio prvi film koji je pogledao. Zatim je izračunat broj puta kada je svaki film bio prvi put pogledan. Bitno je utvrditi koji su filmovi stvarno prvi izbor novih korisnika, jer oni predstavljaju relevantne podatke za analizu početnih preferencija korisnika.

Ovaj proces proveden je pomoću metode lakta. Na grafu se prikazuje broj korisnika koji su određene filmove prvi put gledali. Na x osi su filmovi, a na y osi je broj korisnika kojima je taj film bio prvi izbor. Filmovi su rangirani prema broju pregleda. Zatim se traži točka na grafu gdje se broj pregleda naglo smanjuje. Ta točka označava prijelaz

između filmova koji su popularni kao prvi izbori i onih koji to nisu.

Nakon takvog filtriranja, dobiveno je samo 26 filmova, što nije previše i smatra se dobrim rezultatom jer se time izdvajaju najrelevantniji naslovi. Ovi filmovi bili su prvi odabrani filmovi za gotovo tri tisuće korisnika, što čini gotovo 50% svih korisnika.

Ove filmove moguće je povezati s informacijama o njihovoj općoj popularnosti, odnosno s njihovom kombiniranom ocjenom koja uzima u obzir i broj pregleda i prosječnu ocjenu filma. Na taj način, nije riječ samo o filmovima koji su prvi odabrani od strane korisnika, već i o filmovima koji su generalno dobri i popularni.

Slika 4.2. prikazuje rezultate dobivene takvim sustavom preporuka. Prikazani filmovi često su bili prvi izbor korisnika, što je prikazano u koloni *num_users_first_view*, gdje je za svaki film vidljivo koliko je korisnika odabralo taj film kao svoj prvi izbor. Ovi filmovi dodatno su filtrirani pomoću kriterija *mix_grade*, koji uzima u obzir ukupan broj pregleda *views* i prosječnu ocjenu *average_rating*.

movie_id	title	num_users_first_view	views	average_rating	mix_grade	mix_grade_position
1210.0	Star Wars: Episode VI - Return of the Jedi (1983)	290	2880	4.023264	4.564259	32
260.0	Star Wars: Episode IV - A New Hope (1977)	179	2988	4.453481	5.000000	1
1193.0	One Flew Over the Cuckoo's Nest (1975)	160	1723	4.390017	4.608870	28
110.0	Braveheart (1995)	139	2442	4.234644	4.649854	25
1198.0	Raiders of the Lost Ark (1981)	134	2511	4.477499	4.897929	3
527.0	Schindler's List (1993)	127	2303	4.510204	4.874477	5
1196.0	Star Wars: Episode V - The Empire Strikes Back...	119	2987	4.293606	4.848313	7
858.0	Godfather, The (1972)	113	2223	4.524966	4.867525	6
593.0	Silence of the Lambs, The (1991)	112	2575	4.352233	4.796031	10
1197.0	Princess Bride, The (1987)	87	2316	4.303972	4.682547	24
2028.0	Saving Private Ryan (1998)	79	2649	4.337863	4.801785	9
912.0	Casablanca (1942)	79	1669	4.412822	4.616339	27
1617.0	L.A. Confidential (1997)	66	2287	4.219939	4.595366	29
1221.0	Godfather: Part II, The (1974)	63	1692	4.357565	4.570021	31
908.0	North by Northwest (1959)	52	1314	4.384323	4.496451	41

Slika 4.2. Preporuke temeljene na popularnosti

Kao što je već spomenuto, u situaciji hladnog starta, tradicionalni pristupi poput kolaborativnog filtriranja i preporuka temeljenih na sadržaju mogu biti vrlo ograničeni. Međutim, osim već spomenutih klasičnih metoda, postoje i druge vrste preporuka koje mogu biti korisne u uvjetima hladnog starta.

4.2.2. Preporuke temeljene na kontekstu

Jedan od takvih pristupa su sustavi za preporuke temeljene na kontekstu. Takvi sustavi koriste informacije o korisnikovom okruženju i situaciji u kojoj se nalazi, poput lokacije, prigode ili doba dana te tako pružaju još personaliziranije prijedloge. Primjerice, kada je korisnik na putovanju te se nalazi na drugačijoj lokaciji, mogu se preporučiti restorani koji su popularni za to područje.

Generalno kada se govori o kontekstualnom filtriranju postoji nekoliko različitih pristupa kako se primjenjuju. Tri glavne metode su kontekstualno predfiltriranje, kontekstualno postfiltriranje te kontekstualno modeliranje. U predfiltriranju se kontekst koristi kao filtar na početku procesa, dok se u postfiltriranju prvo primjenjuju klasični pristupi za preporuke, a zatim se koristi kontekst za filtriranje rezultata. U kontekstualnom modeliranju se kontekst integrira odmah u izradu modela.

U ovom radu, cilj je istražiti kako kontekstualno filtriranje može biti primijenjeno u situaciji hladnog starta. Čak i u slučajevima kada sustav nema nikakvih podataka o korisničkoj povijesti, kontekstualne informacije mogu pomoći u predviđanju što bi korisniku moglo biti relevantno u određenom trenutku.

Od podataka koji opisuju kontekst preporuka, na raspolaganju je vrijeme kada je film pogledan, odnosno kada je ocjena dana. Ideja je iskoristiti te informacije za grupiranje prema nekim vremenskim intervalima, primjerice prema određenim dijelovima dana, danima u tjednu ili vikendima. Takvo grupiranje omogućilo bi prepoznavanje u kojim vremenskim razdobljima korisnici najčešće konzumiraju određeni sadržaj.

Naravno, takav pristup nije personaliziran, ali u kontekstu problema hladnog starta može biti koristan. Na primjer, ako podaci pokazuju da se komedije češće gledaju tijekom vikenda, sustav može preporučiti komedije novim korisnicima koji počinju koristiti platformu za vrijeme vikenda.

Također je važno napomenuti da je *timestamp* zapravo vrijeme kada je korisnik ocijenio film, a ne nužno vrijeme kada je film pogledan. Zapravo, za ovakvu analizu je potrebno pretpostaviti da su korisnici ocijenili filmove odmah nakon gledanja. Međutim, vrlo je moguće da ta pretpostavka nije točna, što može utjecati na preciznost zaključka.

U ovom radu provedena je analiza temeljena na grupiranju filmova prema dijelovima dana, danima u tjednu te prema tome radi li se o vikendu ili radnom danu. Primjerice, dijelovi dana podijeljeni su na kategorije noć, večer, popodne i jutro te je za svaku od tih kategorija analizirano koliko je filmova dobilo recenzije upravo u tom vremenskom intervalu. Zatim je bilo bitno uočiti i prepoznati filmove koji se ističu po broju ocjena unutar pojedinih vremenskih kategorija. Na taj način, moguće je identificirati filmove koji su često gledani u određeno vrijeme. Takav pristup primijenjen je i za grupiranje po danima u tjednu te prema tome radi li se o vikendu ili radnom danu.

Na temelju dobivenih rezultata, nije moguće izvući korisne zaključke vezane uz dijelove dana, dane u tjednu ili razlikovanje između vikenda i radnih dana. Filmovi koji su najgledaniji u pojedinim kategorijama općenito su najgledaniji filmovi. Iz tog razloga, ne može se utvrditi značajna povezanost između vremenskih intervala i gledanja određenih filmova.

Kao dodatni korak analize, izdvojeno je pedeset najgledanijih filmova unutar svake vremenske kategorije (dijelovi dana, dani u tjednu, vikend/radni dan), s time da ti filmovi nisu među pedeset generalno najgledanijih. Rezultati tog koraka za podjelu po dijelovima dana prikazani su na slici 4.3. Kao što se može primijetiti, takvih filmova ima vrlo malo.

```
Analiza za Jutro:
Top filmovi specifičnih za Jutro (nisu generalni top 50):
                                        title  count  general_count
2865                              Speed (1994)    246           1649
1559  Indiana Jones and the Last Crusade (1989)    242           1627
697                              Contact (1997)    241           1382
947                 Edward Scissorhands (1990)    241           1472
3380                               X-Men (2000)    239           1510

Analiza za Noć:
Top filmovi specifičnih za Noć (nisu generalni top 50):
                                  title  count  general_count
4346                 Die Hard (1988)    808           1665
5046  Hunt for Red October, The (1990)    799           1650

Analiza za Popodne:
Top filmovi specifičnih za Popodne (nisu generalni top 50):
                                 title  count  general_count
8449               GoodFellas (1990)    281           1655
10483  When Harry Met Sally... (1989)    272           1568
7644       Breakfast Club, The (1985)    269           1537
8938             Lethal Weapon (1987)    262           1627

Analiza za Večer:
Top filmovi specifičnih za Večer (nisu generalni top 50):
                                       title  count  general_count
11161                    Casablanca (1942)    430           1669
12693  Monty Python and the Holy Grail (1974)    394           1598
13832                   Toy Story 2 (1999)    386           1583
```

Slika 4.3. Preporuke temeljene na kontekstu

Preporuke temeljene na kontekstu ne smatraju se najprikladnijima za ovaj slučaj,

odnosno ove podatke. Prvi problem je početna pretpostavka da su filmovi gledani kada su i ocijenjeni. Međutim, čak i uz tu pretpostavku, ne postoji jasna poveznica između određenih vremenskih intervala i gledanja specifičnih filmova.

4.2.3. Preporuke temeljene na demografiji

Kao što je već spomenuto, problem hladnog starta korisnika javlja se zbog nedostatka zabilježenih interakcija. Takav nedostatak podataka sprječava formiranje korisničkog profila u sustavima temeljenim na sadržaju te određivanje sličnih korisnika u sustavima temeljenim na suradnji.

Prije samog korištenja sustava, za svakog korisnika, iako još nema recenzija, dostupni su podatci koji opisuju njegov identitet. Ti podatci uključuju demografske karakteristike poput spola i godina. Te informacije mogu poslužiti kao osnova za određivanje sličnosti među korisnicima.

Primjerice, ako sustav prepoznaje da je korisnik muškarac u ranim tridesetima i da ta demografska skupina voli određene žanrove ili čak konkretne filmove, može preporučiti takav sadržaj. Ova metoda slična je sustavima temeljenim na suradnji, gdje se povratne informacije koriste za utvrđivanje sličnosti među korisnicima, ali se oslanja na dostupne podatke. Takav sustav može generirati preporuke i prije nego što korisnik da vlastite recenzije, što je iznimno korisno za stvaranje inicijalnih preporuka.

Sustavi koji koriste demografske podatke korisnika, poput spola, dobi, ili interesa, nazivaju se sustavi temeljeni na demografiji. U sklopu ovog rada implementiran je jednostavan sustav temeljen na demografiji koji bi se koristio za rješavanje hladnog starta korisnika.

Za implementaciju takvog sustava korišteni su dostupni podaci iz skupa podataka *MovieLens*. Informacije iz datoteke *Users* smatraju se dostupnima i one uključuju dob, spol i zanimanje. Ove informacije dostupne su i za nove korisnike koji još nisu ostavili recenzije. Osim nabrojanih podataka, dostupna je i lokacija, no ona se ne smatra toliko relevantnom za ovu analizu.

Prvi korak implementacije je mapiranje tekstualnih demografskih podataka u numeričke vrijednosti Dobne skupine mapirane su u vrijednosti od nula do šest, primjerice,

dobna skupina 18-24 mapirana je u vrijednost 1. Spolovi M i F mapirani su u vrijednosti 0 i 1, dok su zanimanja, poput *college grad student* ili *programmer* mapirana u brojeve od nula do dvadeset. Ovaj proces omogućava stvaranje numeričke matrice potrebne za matematičke izračune.

Nakon toga, izrađuje se demografska matrica korisnika, gdje su redci korisnici, a stupci predstavljaju njihove demografske značajke, dob, spol te zanimanje. Ova matrica omogućava usporedbu korisnika na temelju sličnosti njihovih demografskih karakteristika.

Postupak dobivanja sličnih korisnika te zatim filmova isti je kao i kod implementiranog sustava temeljenog na suradnji. Za računanje sličnosti između korisnika koristi se kosinusna sličnost a rezultat tog koraka je matrica sličnosti. Na temelju ove matrice biraju se slični korisnici, a zatim se preporučuju filmovi koje su ti korisnici ocijenili.

	user_id	age_num	gender_num	occ_num
0	2858	0.0	1.0	10.0
1	2657	0.0	1.0	10.0
2	1508	0.0	1.0	10.0
3	468	0.0	1.0	10.0
4	4086	0.0	1.0	10.0
5	5302	0.0	1.0	10.0
6	1088	0.0	1.0	10.0
7	119	0.0	1.0	10.0
8	5459	0.0	1.0	10.0
9	2155	0.0	1.0	10.0

	movie_id	predicted_rating	zero_ratings_count	non_zero_ratings	title	genres	user_age	user_gender	user_occ
0	39.0	2.7	3	[3, 4, 5, 3, 3, 4, 5]	Clueless (1995)	Comedy|Romance	0.0	1.0	10.0
1	2485.0	2.6	4	[5, 5, 4, 3, 4, 5]	She's All That (1999)	Comedy|Romance	0.0	1.0	10.0
2	1197.0	2.6	4	[4, 4, 5, 5, 4, 4]	Princess Bride, The (1987)	Action|Adventure|Comedy|Romance	0.0	1.0	10.0
3	2581.0	2.6	4	[5, 5, 4, 3, 4, 5]	Never Been Kissed (1999)	Comedy|Romance	0.0	1.0	10.0
4	527.0	2.5	5	[5, 5, 5, 5, 5]	Schindler's List (1993)	Drama|War	0.0	1.0	10.0
5	3114.0	2.5	4	[4, 4, 5, 3, 4, 5]	Toy Story 2 (1999)	Animation|Children's|Comedy	0.0	1.0	10.0
6	1569.0	2.4	5	[5, 5, 5, 4, 5]	My Best Friend's Wedding (1997)	Comedy|Romance	0.0	1.0	10.0
7	2762.0	2.4	5	[5, 5, 4, 5, 5]	Sixth Sense, The (1999)	Thriller	0.0	1.0	10.0
8	3081.0	2.4	5	[5, 5, 4, 5, 5]	Sleepy Hollow (1999)	Horror|Romance	0.0	1.0	10.0
9	708.0	2.4	4	[4, 5, 4, 3, 4, 4]	Truth About Cats & Dogs, The (1996)	Comedy|Romance	0.0	1.0	10.0

Slika 4.4. Preporuke temeljene na demografiji

Rezultat takvog sustava prikazan je na slici 4.4. Prva tablica prikazuje slične korisnike, dok druga prikazuje dane preporuke i karakteristike ciljanog korisnika. U ovom primjeru, izabran je korisnik kojemu su uklonjene sve recenzije, odnosno stvorena je situacija hladnog starta za tog korisnika. Korisnik je ženska osoba, dobne skupine od 18-24, te je zanimanjem učenik. Vrlo vjerojatno je riječ o djevojci koja završava srednju školu. Njoj slični korisnici imaju iste demografske karakteristike, a filmovi koji su

izabrani vrlo dobro opisuju što bi neka djevojka tih godina mogla gledati.

4.3. Problemi hladnog starta nakon početne faze

Sustavi za davanje preporuka suočavaju se s različitim izazovima tijekom svog životnog ciklusa. U prethodnim poglavljima bila je riječ o problemima hladnog starta, koji su se odnosili na situacije kada sustavi za preporuke nisu imali nikakvih povijesnih podataka. To uključuje slučajeve potpuno novih sustava, kao i situacije s novim korisnicima ili novim stavkama koje još nemaju nikakve recenzije ili nisu dali povratne informacije. Iako su problemi dobro poznati u takvim situacijama, izazovi se često nastavljaju i nakon što sustav prebrodi takve početne faze. Ovi problemi nastaju jer, iako sustav ima neke recenzije, on još uvijek ne može davati potpuno učinkovite preporuke.

Slično kao što je učinjeno za problem hladnog starta, i za ovaj problem prilagođen je skup podataka. U ovom slučaju simulirala se situacija u kojoj sustav već ima određeni broj recenzija, ali nedovoljno za optimalno funkcioniranje. Ova faza predstavlja korak nakon hladnog starta, gdje sustav ima ograničene informacije o korisnicima i stavkama.

Odabrani su određeni korisnici i filmovi kojima su recenzije djelomično uklonjene. Primjerice, odabranim korisnicima je broj recenzija smanjen tako da imaju između jedne i deset recenzija, dok je za pojedine filmove smanjen tako da imaju između jedne i pet recenzija. Ova prilagodba podataka omogućava analizu učinkovitosti sustava u uvjetima ograničenih podataka. Analiza obuhvaća sustave temeljene na sadržaju i sustave temeljene na suradnji, kako bi se procijenilo te usporedilo kako se svaki od njih nosi s ograničenim informacijama.

Jedan od uočenih problema je problem prevelike specifičnosti. Taj problem odnosi se na situaciju kada sustavi pružaju preporuke koje su preusko usmjerene na dosadašnje interese korisnika. Ovaj problem najviše dolazi do izražaja u implementaciji sustava temeljenih na sadržaju u slučaju kada korisnik ima vrlo malo recenzija. Način na koji ova implementacija radi je da odabire najnovije i dobro ocijenjene filmove koji su žanrovski slični ili isti kao filmovi koje je korisnik već ocijenio. U slučaju da korisnik ima samo jednu dobru recenziju, svi preporučeni filmovi bit će upravo tog žanra. Ako korisnik odabire samo iz predloženih filmova, to može dovesti do situacije u kojoj su sve

preporuke unutar istog žanra kao prvi pogledani film. Takav ciklus može se prekinuti jedino odabirom filma iz novog žanra.

Osim toga, uočen je problem noviteta. On se javlja kod stavaka koje nemaju mnogo recenzija te stoga ostaju zanemarene u odnosu na starije i više recenzirane stavke. Sustavi temeljeni na sadržaju nemaju nužno problema s ovim, jer oni gledaju samo sličnosti žanrova, pa tako svi filmovi istog žanra imaju jednake šanse za preporuku. Međutim, može se primijeniti već spomenuti postupak u kojem se novim filmovima daje veća težina kako bi se potpomogla njihova vidljivost i šanse za preporuku. S druge strane, sustavi temeljeni na suradnji zbog načina rada favoriziraju filmove s puno recenzija. Kada pronađu slične korisnike, oni odabiru filmove s najvišim prosječnim ocjenama među tim korisnicima. U slučaju nove stavke koju je, primjerice, pogledao samo jedan korisnik, vjerojatnost da će ona imati najvišu prosječnu ocjenu je vrlo mala.

Još jedan problem sa sustavima temeljenim na suradnji nastaje kada korisnik nema dovoljno recenzija, što dovodi do generičkih preporuka. U ekstremnim slučajevima, kada korisnik ima samo jednu recenziju, sustav često preporučuje popularne i generičke filmove koji su gotovo identični za svakog korisnika. U toj fazi, preporuke zapravo nisu personalizirane. Međutim, s povećanjem broja recenzija koje korisnik daje, sustav postaje sve precizniji u prepoznavanju obrazaca te tako daje i bolje preporuke.

5. Hibridni sustavi za davanje preporuka

Kao što je već spomenuto, sustavi preporuka temeljeni na suradnji i sustavi temeljeni na sadržaju imaju svoje prednosti, ali i nedostatke.

Sustavi temeljeni na suradnji koriste povratne informacije od sličnih korisnika kako bi prepoznali obrasce ponašanja i tako pružili preporuke. To omogućuje veliku raznolikost preporuka jer se uzimaju u obzir preferencije drugih korisnika sličnog ukusa. No, prevelika raznolikost nije uvijek poželjna. Osim toga, ti sustavi mogu biti neprecizni u situacijama hladnog starta. Kada se pojavi novi korisnik koji još nije dao recenzije ili kada je dodana nova stavka, sustav može loše funkcionirati. Primjerice, može zanemariti nove stavke ili pružati generičke preporuke.

S druge strane, sustavi temeljeni na sadržaju su manje raznoliki jer pružaju preporuke koje su direktno povezane s onim što je korisnik već pogledao. To je podosta izraženo u implementiranom sustavu gdje se kao karakteristika za mjerenje sličnosti uzimaju žanrovi filmova. Takvi sustavi mogu imati problema s prevelikom specifičnosti preporuka, gdje korisnici dobivaju usko ciljani sadržaj, što ograničava njihovu priliku za otkrivanje novih vrsta sadržaja. Međutim, prednost ovih sustava je da dobro funkcioniraju u situaciji hladnog starta, točnije, ne pate toliko od problema koji nastaju kao posljedica premale količine povijesnih interakcija.

Kako bi se iskoristile prednosti oba sustava te izbjegle njihove slabosti, pojavila se ideja kombiniranja tih tehnika. Sustav koji kombinira karakteristike različitih pristupa naziva se hibridni sustav. Generalno, kada se spominje hibridni sustav, govori se o sustavu koji objedinjuje sustave temeljene na sadržaju i one temeljene na suradnji. Cilj takvog sustava je postići ravnotežu između raznovrsnih i specifičnih preporuka. Osim toga,

kombiniranjem karakteristika pokušava s riješiti problem hladnog starta, što bi omogućilo novim korisnicima i stavkama bolju priliku za dobivanje preciznih preporuka od samog početka.

Hibridne sustave moguće je klasificirati u sedam pristupa [1]. Ti pristupi obuhvaćaju različite načine kombiniranja modela:

- Težinski sustav
- *Switching* sustav
- Miješani sustavi
- Sustavi kombinacije značajki
- Sustavi obogaćivanja značajki
- Kaskadni sustav
- *Meta* sustav

U težinskom hibridnom sustavu rezultati različitih modela kombiniraju se uz pomoć unaprijed definiranih težina. Svaki model generira svoje preporuke, a rezultat se zatim težinski kombinira. Primjerice, oba modela daju preporuke filmova i predviđenu ocjenu od jedan do pet. Težinski sustav za svaku preporuku kalkulira novu ocjenu na način da uzima zasebne ocjene, množi ih s težinom, primjerice 0.3 za sadržajni model i 0.7 za suradnički model. Filmovi se zatim rangiraju prema toj težinski izračunatoj ocjeni.

Switching sustavi biraju odgovarajući sustav za generiranje preporuka ovisno o specifičnoj situaciji. Umjesto da se uvijek koristi isti model ili da se kombiniraju rezultati svih modela, ovaj sustav odabire najbolji model za danu situaciju. Sustav prati određene kriterije, poput povijesti interakcija ili korisničkog profila, i na temelju tih kriterija odlučuje koji će model biti korišten u tom trenutku.

Miješani hibridni sustavi kombiniraju preporuke iz različitih modela kako bi korisnicima pružili preciznije rezultate. Ovi sustavi izdvajaju različite podskupove sadržaja

[1] Prema članku dostupnom na `https://medium.com/analytics-vidhya/7-types-of-hybrid-recommendation-system-3e4f78266ad8`.

koji bi mogli biti preporučeni korisniku. Ti se podskupovi generiraju na temelju različitih kriterija primjerice korisničkog profila, povijesti interakcija ili karakteristika stavki. Zatim se skupovi kandidata koriste kao ulazi za različite modele preporuka, poput sadržajnih i suradničkih modela, a predikcije tih modela kombiniraju se kako bi se proizvele konačne preporuke.

U hibridnom sustavu kombinacije značajki dodaje se virtualni model preporuka koji doprinosi sustavu. Taj model služi za procesiranje i generiranje novih značajki koje se zatim koriste u glavnom sustavu za davanje preporuka. Na primjer, značajke suradničkog modela preporuka moguće je umetnuti u značajke koje se koriste u sadržajnom model preporuka. Umjesto da se sustav oslanja isključivo na značajke sadržaja, poput žanra, model može uzeti u obzir i preferencije korisnika koji su slični. Takav hibridni model uzima suradničke informacije iz pod modela te ih uključuje u sustav temeljen na sadržaju. Time se omogućuje obogaćivanje preporuka koje glavni sustav generira.

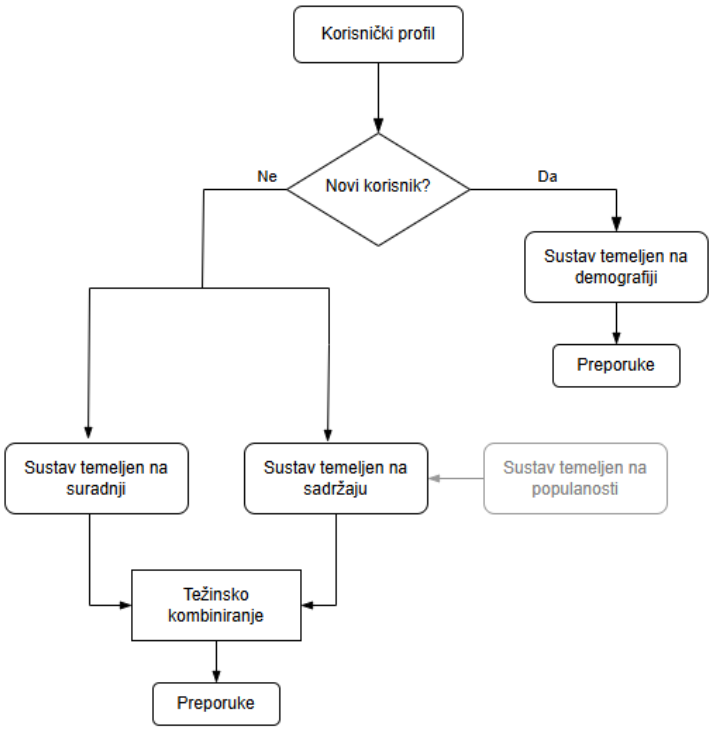
Hibridni sustav obogaćivanja značajki isto tako korisni dodatni model. Taj dodatni model doprinosi glavnom tako što generira ocijene ili klasificira korisnički profil. Na taj način pružaju se dodatne informacije glavnom modelu. Primjerice, pravilom pridruživanja mogu se otkriti obrasci u odabirima. U kontekstu sustava za preporuku filmova, pravilo pridruživanja može otkriti da korisnici koji gledaju film A često gledaju i film B. Ova se informacija zatim može dodati korisničkom profilu. Tu informaciju tada je moguće dodati korisnički profil .

Kaskadni hibridni sustav radi na način da koristi primarni model za generiranje preporuka te zatim koristi sekundarni model za dodatnu obradu ili rješavanje sitnijih problema u preporukama. Takav sustav ima strogu hijerarhijsku strukturu.

Meta hibridni sustav baš kao i hibridni sustav kombinacije i obogaćivanja značajki koristi model koji doprinosi kako bi poboljšao skup podataka za glavni model preporuka. Takav sustav zamjenjuje originalni skup podataka pomoćnim modelom odnosno modelom koji doprinosi. Taj naučeni model koristi se kao ulaz za glavni model preporuka.

5.1. Implementacija hibridnog sustava

U ovom radu podosta se govorilo o korisnosti koje podatci nose. Prilikom analize skupa podataka koji je korišten u ovom radu, donesen je zaključak kako je taj skup vrlo kvalitetan. No, postavilo se pitanje kako bi sustavi radili u uvjetima koji nisu toliko idealni. Sustav temeljen na suradnji imao je iznimno velikih poteškoća pri radu s ograničenom količinom podataka, dok je sustav temeljen na sadržaju bolje podnosio problem hladnog starta. S druge strane, preporuke sustava temeljenog na sadržaju podosta su vezane uz filmove koje je korisnik već pogledao, dok su preporuke sustava temeljenog na suradnji ponekad previše raznolike. Iz tog razloga, nameće se ideja o kombiniranju sustava na način koji bi istaknuo što više pozitivnih karakteristika. Struktura hibridnog sustava implementiranog u ovom radu prikazana je slikom 5.1.



Slika 5.1. Struktura implementiranog sustava preporuka

Proces implementacije hibridnog sustava zapravo je započet evaluacijom pozitivnih i negativnih strana svakog sustava te pronalaženjem situacija u kojima sustavi jednostavno ne bi mogli efikasno raditi. Konkretno, kada se govori o problemima vezanim uz hladni start, posebice uz hladni start korisnika, trenutni sustavi ne ostvaruju ciljane rezultate. No, u radu su obrađeni i implementirani sustavi koji se vrlo dobro nose s problemima

hladnog starta. Sustavi temeljeni na popularnosti ne zahtijevaju podatke o ciljanom korisniku pa iz tog razloga efikasno prevladavaju problem hladnog starta. Međutim, takvi sustavi nisu personalizirani. U situaciji gdje se može pretpostaviti da su osobni podatci korisnika dostupni, sustavi temeljeni na demografiji pokazali su se kao izvrsno rješenje za problem hladnog starta.

Ta karakteristika demografskih sustava iskorištena je za situacije kada ne postoji dovoljno podataka o interakcijama pojedinog korisniku. Pogodan pristup kombinacije demografskog sustava s ostalim sustavima jest *Switching*. Kao što je već spomenuto, takav postupak izrade hibridnog modela temelji se na biranju odgovarajućeg sustav za generiranje preporuka ovisno o specifičnoj situaciji. U ovoj implementaciji hibridnog sustava, ako korisnik ima pet ili manje pozitivnih recenzija, generirat će se preporuke temeljene na demografiji. Implementacija samog sustava temeljenog na demografiji opisana je u prijašnjem poglavlju.

Drugi problem hladnog starta koji se pokazao manje značajnim jest problem hladnog starta proizvoda. U sustavu temeljenom na sadržaju implementirana je ideja da se ponekim nasumično odabranim novim filmovima daje malena prednost. S druge strane, sustav temeljen na suradnji ne daju željene rezultate kada je ovaj problem u pitanju. Ta činjenica služila je kao početna motivacija za spajanje rezultata sustava temeljenog na sadržaju i onog na suradnji. Naravno, mješavina rezultata tih sustava iskoristila bi puno prednosti i u drugim aspektima učinkovitosti. No, prije spajanja tih sustava, napravljene su određene promjene u implementaciji svakog od odvojenih sustava.

Generalna ideja, način rada i algoritam implementiranih sustava ostala je ista. Promjene pojedinih sustava provedene su s ciljem da povećaju efikasnost hibridnog sustava. Glavni problem koji je potaknuo izmjene bio je odabir preporuka i rangiranje preporuka. Naime, kada se sustav temeljen na sadržaju koristi samostalno, on generira određen broj preporuka koje su zatim prikazane ciljanom korisniku. U slučaju implementiranog sustava generiraju se redom pet, četiri, tri preporuke za svaki film iz korisničkog profila. U korisničkom profilu se nalaze tri filma koja su posljednje dobro ocijenjeni. Taj broj generiranih preporuka pogodan je za prikaz korisniku, no nije primjeren za kombiniranje i korištenje u hibridnom sustavu.

Sustav temeljen na sadržaju koji će se koristiti kao komponenta hibridnog sustava treba generirati više preporuka pri čemu je redoslijed tih preporuka također bitan. Taj sustav implementiran je na način da se u korisničinom profilu nalazi sedam filmova, te se za svaki od tih film generira od četiri do deset preporuka. Osim toga, redoslijed preporuka je ciklički strukturiran. Preporuke su generirane na način da se naizmjenično odabire preporuka temeljena na filma u korisničkom profilu. Cilj je bio osigurati da se preporuke temeljene na svih sedam filmova uzmu u obzir.

Kao što je već spomenuto, sustavi temeljeni na sadržaju imaju podosta mana. Jedan od problema s kojim se suočava implementirani sustav je situacija u kojoj postoji više filmova s jednakom sličnosti. Tada sustav nije u mogućnosti odabrati najpogodniji film. Iz tog razloga, u hibridni sustav dodana je komponenta koja rješava taj problem.

Sustav temeljen na sadržaju kombiniran je sa sustavom temeljenim na popularnosti, koji se onda koristi kao odlučujući faktor pri odabiru između jednako sličnih filmova. Takav način kombiniranju sustava predstavlja kaskadni pristup. Kao što je već spomenuto, kaskadni hibridni sustav prvo koristi osnovni model za generiranje preporuka te zatim koristi sekundarni model za dodatnu obradu ili rješavanje sitnijih problema.

Sustav temeljen na suradnji nije bilo potrebno previše izmijeniti. Povećan je broj korisnika koji se smatraju sličnima ciljanom korisniku te broj ukupnih preporuka koje sustav daje. U ovom sustavu, sličnost se određuje među korisnicima na temelju ocjena koje su dali, te se rijetko dešava situacija da različiti korisnici imaju istu razinu sličnosti.

Nakon što su implementacije sustava malo izmijenjene, njihovi rezultati su kombinirani. Takvo spajanje rezultata postiglo se težinskim pristupom. Ocjene preporuka oba modela prilagođene su tako da budu u intervalu od nula do pet. U sustavu temeljenom na sadržaju ocjena preporuke zapravo predstavlja mjesto u poretku na listi svih ostalih preporuka. Preporuka koja je prva na listi ima ocjenu 5.0, druga ima 4.9 i tako redom. U sustavu temeljenom na sadržaju ocjene preporuka predstavljaju prosječnu ocjenu sličnih korisnika, skaliranu na raspon od 0 do 5. Unaprijed definirana težina za rezultate dobivene sadržajnim modelom iznosi 0.4, dok je za suradnički model postavljena na 0.6.

Rezultati implementiranog sustava za nasumičnog korisnika prikazani su na slici 5.2. U ovom slučaju korisnik ima prethodnih recenzija pa se ne koristi demografsko filtrira-

nje. Gornji dio slike prikazuje 7 filmova iz korisničkog profila koji se koriste kao temelj za preporuke temeljene na sadržaju. Na ostatku slike prikazane su preporuke hibridnog sustava. Osim samih naslova filmova, odnosno preporuka, u tablici su prikazani i žanrovi filmova te hibridna kombinira ocjena, suradnička ocjena i sadržajna ocjena. Filmovi koji imaju suradničku ili sadržajnu ocjenu 0 nisu se našli u top 50 filmova tog sustava. Iz rezultata se može uočiti kao je postignuta ravnoteža između sustava temeljenog na sadržaju i onog na suradnji.

	title	movie_id	genres	rating	recency_score	weighted_rating
542276	Lethal Weapon 3 (1992)	2002	Action|Comedy|Crime|Drama	5.0	4.819512	7.409756
396941	Die Hard 2 (1990)	1370	Action|Thriller	5.0	4.304878	7.152439
129024	Jurassic Park (1993)	480	Action|Adventure|Sci-Fi	5.0	4.109756	7.054878
453251	Hunt for Red October, The (1990)	1610	Action|Thriller	5.0	4.109756	7.054878
108846	True Lies (1994)	380	Action|Adventure|Comedy|Romance	5.0	4.109756	7.054878
96995	Forrest Gump (1994)	356	Comedy|Romance|War	5.0	3.809756	6.904878
279281	On Golden Pond (1981)	1124	Drama	5.0	3.156098	6.578049

	movie_title	genres	hybrid_score	collab_score	content_score
0	Lethal Weapon (1987)	Action|Comedy|Crime|Drama	4.124	3.54	5.0
1	Die Hard (1988)	Action|Thriller	3.568	2.68	4.9
2	Air Force One (1997)	Action|Thriller	3.090	4.15	1.5
3	In the Line of Fire (1993)	Action|Thriller	3.080	2.80	3.5
4	Schindler's List (1993)	Drama|War	3.000	5.00	0.0
5	Rock, The (1996)	Action|Adventure|Thriller	2.490	4.15	0.0
6	Thomas Crown Affair, The (1999)	Action|Thriller	2.376	2.56	2.1
7	Fargo (1996)	Crime|Drama|Thriller	2.340	3.90	0.0
8	Men in Black (1997)	Action|Adventure|Comedy|Sci-Fi	2.196	3.66	0.0
9	Good Will Hunting (1997)	Drama	2.196	3.66	0.0
10	Star Trek: The Wrath of Khan (1982)	Action|Adventure|Sci-Fi	1.992	0.12	4.8
11	Sixth Sense, The (1999)	Thriller	1.974	3.29	0.0
12	Die Hard: With a Vengeance (1995)	Action|Thriller	1.880	0.00	4.7
13	Princess Bride, The (1987)	Action|Adventure|Comedy|Romance	1.840	0.00	4.6
14	M*A*S*H (1970)	Comedy|War	1.800	0.00	4.5
15	Hoogste tijd (1995)	Drama	1.760	0.00	4.4
16	Lethal Weapon 2 (1989)	Action|Comedy|Crime|Drama	1.720	0.00	4.3
17	Killer, The (Die xue shuang xiong) (1989)	Action|Thriller	1.680	0.00	4.2
18	Star Trek: First Contact (1996)	Action|Adventure|Sci-Fi	1.640	0.00	4.1
19	E.T. the Extra-Terrestrial (1982)	Children's|Drama|Fantasy|Sci-Fi	1.608	2.68	0.0

Slika 5.2. Preporuke hibridnog sustava

5.2. Evaluacija hibridnog sustava

Evaluacija sustava je korak koji procjenjuje koliko je dobro implementacija određenog sustava napravljena. Generalna ideja jest usporediti rezultate koje sustav daje s ciljanim rezultatima. Primjerice, u pojedinim modelima moguće je podijeliti podatke u skup za treniranje i skup za testiranje. Model se uči na podatcima za treniranje te se zatim testira na neviđenim podatcima. Predikcije koje model daje uspoređuju se s ciljanim podatcima. Takav postupak samo je jedan od mogućih metoda koje se koriste za evaluaciju sustava.

U sustavima za davanje preporuka, problem evaluacije malo je drugačiji. Naime, sus-

tavi za davanje preporuka ne generiraju predikcije filmova koje će korisnik gledati, već daju preporuke. Iz tog razloga idealna evaluacija sustava za preporuke mjerila bi kvalitetu preporuka a ne točnost predikcije. Kada bi se testirao sustav koji generira preporuke koje savršeno odgovaraju već pogledanim filmovima, takav sustav ne bi bio naročito učinkovit. Idealno, sustav za preporuke bi trebao korisnicima pružati sadržaje koji su usko povezani s njihovim interesima ali im istovremeno i proširivati opcije.

No, problem s takvom evaluacijom jest nedostatak podataka o tom koliko je preporuka kvalitetna. Ocjena filma izražava mjeru o tome koliko je film dobar, a ne koliko je preporuka dobra. Generalno, kao mjera kvalitete u skupovima podataka dostupne su samo dane recenzije. Isto tako, u *MovieLens* skupu podataka, dostupne su samo ocjene filmova, što nije idealno za evaluaciju sustava za preporuke.

Objašnjena je razlika između preporuka filmova i samog filma, odnosno kvalitete preporuke i ocijene filma. Međutim, čak i u situacijama kada se ocjene koriste kao pokazatelj kvalitete, postoje određeni izazovi. Primjerice, može se pretpostaviti da korisnici ocjenjuju samo dio sadržaja koji su pogledali. Osim toga, neki korisnici mogu ostavljati ocijene samo kada imaju jako pozitivno iskustvo dok neki drugi to čine samo kad su imali negativna iskustva. Ovakav pristup evaluaciji ima brojne nedostatke, no ta opcija je opravdana uzevši u obzir da se raspolaže isključivo ocjenama filmova, koje su jedini dostupni podaci u kontekstu vrednovanja.

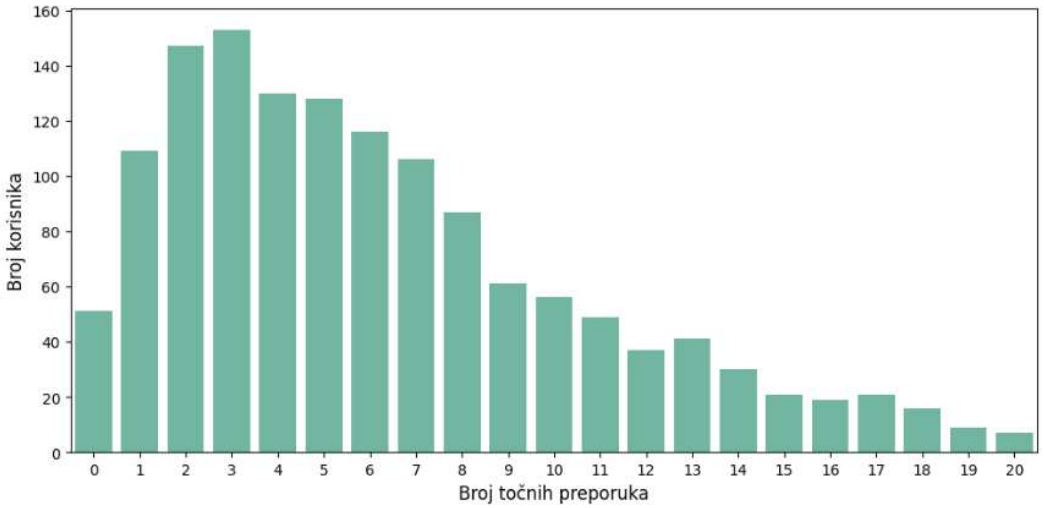
U sklopu ovog rada napravljena je evaluacija koja uspoređuje generirane preporuke i filmove koje je korisnik zaista pogledao. Ideja je da se odredi vremenski trenutak i da se postupa kao da se situacija odvija upravo u tom trenutku. Definira se određeni *timestamp* te se za svaku ocjenu nakon tog trenutka pretpostavlja da nije zabilježena, odnosno da se nije dogodila. Preporuke za filmove generiraju se koristeći isključivo podatke koji su zabilježeni prije tog *timestampa*. Dobiveni rezultati se zatim uspoređuju sa stvarnim podatcima o pogledani i ocjenjenim filmovima.

Taj proces započeo je određivanjem vremenskog trenutka u koji se sustav pozicionirao. Problem je bio što se takvim pristupom gubi mogućnost evaluacije određenih korisnika. Naime, ako je vremenski trenutak postavljen prerano ne postoji dovoljno povijesnih podataka o interakcijama, a ako je trenutak prekasno ne postoji dovoljno filmova koji

bi se mogli usporediti preporukama. Iz tog razloga, određen je trenutak u kojem postoji najviše korisnika koji imaju 5 ili više povijesnih recenzija i 20 ili više filmova koji nisu gledani, odnosno za koje se pretpostavlja da ih korisnik nije pogledao. Taj vremenski trenutak nalazi se otprlike u osamdesetom percentilu.

Kada je pronađen optimalni vremenski trenutak, sve recenzije nakon njega postavljene su na *Nan* vrijednosti. Na taj način je simuliramo da ti filmovi nisu bili pogledani. Primjerice, jedan korisnik imao je pedeset recenzija, trideset prije određenog trenutka i dvadeset nakon njega. Preporuke za tog korisnika bile su generirane na temelju trideset recenzija koje je korisnik dao do tog trenutka, kao i svih drugih dostupnih podataka prije tog trenutka. Preporučeni filmovi uspoređeni su sa dvadeset filmova čije su recenzije bile postavljene na *Nan*. Za svakog korisnika izbrojan je broj filmova koje je stvarno pogledao i koji se podudaraju s preporukama.

Bitno je napomenuti da iako je optimizacija rađena s uvjetom da korisnik ima 5 ili više recenzija, u postupku evaluacije obuhvaćeni su i korisnici koji imaju manje od navedenog praga. Naime, ti korisnici dobit će preporuke temeljene na demografiji. Broj korisnika koji je mogao ući u proces evaluacije jest 1394. Od tih korisnika, njih 1343 pogledalo je barem jedan film koji bi im sustav za preporuke predložio. Na slici5.3. prikazan je broj dobrih preporuka po korisnicima.



Slika 5.3. Distribucija broja točnih preporuka po korisnicima

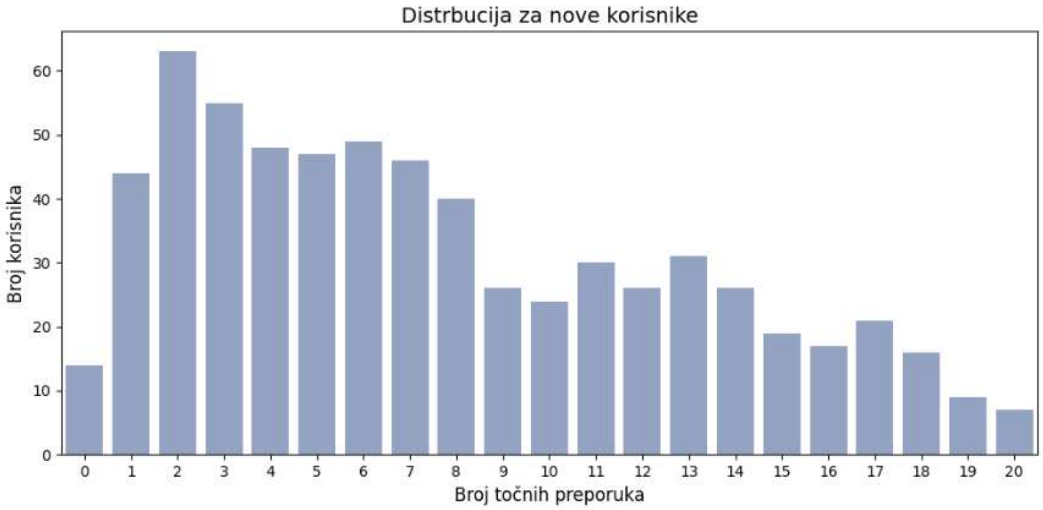
Na osi *x* prikazan je broj preporuka koje su bile ispravne za svakog korisnika, odnosno broj filmova koje je određeni korisnik zaista pogledao. *Y* os prikazuje broj korisnika koji

su imali taj broj ispravnih preporuka.

Iz grafa se može uočiti da se za većinu korisnika ostvaruje između jedne i osam dobrih preporuka. Ovakva distribucija smatra se poželjnom. Naime, kao što je već spomenuto, sustav za davanje preporuka ne teži preciznom predviđanju svih filmova koje će korisnik gledati. Cilj takvog sustava je generirati preporuke koje su relevantne za korisnika, dok istovremeno pružaju raznovrsnost sadržaja. Graf ukazuje upravo na to, sustav pruža određene preporuke koje su usko vezane uz interese korisnika, ali nudi i nešto šire preporuke.

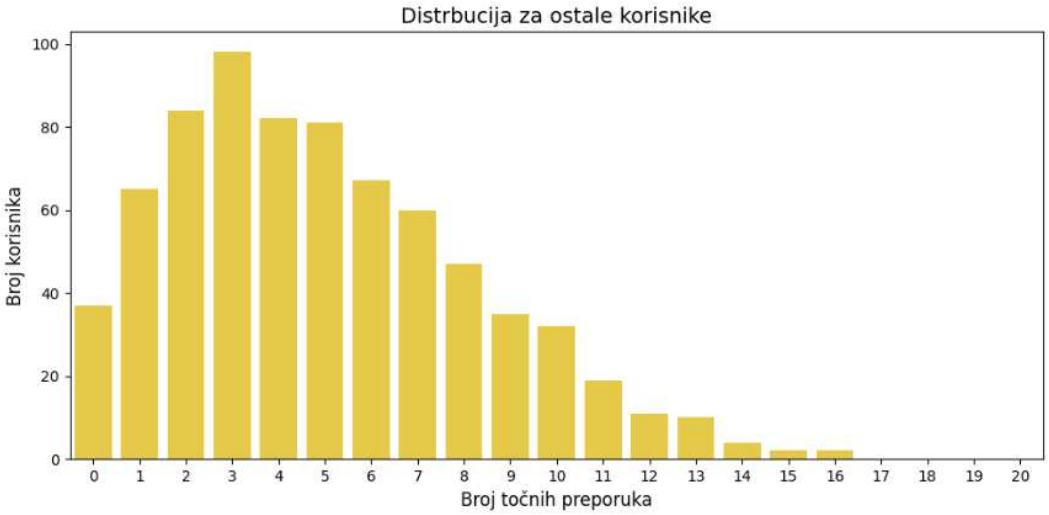
Osim toga, bilo je vrlo bitno analizirati kako funkcionira standardna verzija hibridnog sustava i ona prilagođena novim korisnicima. Na taj način provjerava se radi li jedan sustav izuzetno dobro, dok drugi ne ostvaruje željene rezultate.

Kao što je već spomenuto, u ovu evaluaciju uključeno je 1394 korisnika. Od njih, 658 nije imalo dovoljno povijesnih interakcija za standardni sustav davanja preporuka te su klasificirani kao novi korisnici. Za njih su generirane preporuke temeljene na demografskim podacima. Graf koji prikazuje distribuciju broja dobrih preporuka po novim korisnicima prikazan je na slici 5.4.



Slika 5.4. Distribucija broja točnih preporuka po novim korisnicima

Ostali korisnici, njih 736, imaju dovoljno podataka, te se za njih koristi hibridni sustav davanja preporuka. Graf koji prikazuje distribuciju broja dobrih preporuka po tim korisnicima prikazan je na slici 5.5.



Slika 5.5. Distribucija broja točnih preporuka po ostalim korisnicima

Oba grafa, onaj koji prikazuje distribuciju broja dobrih preporuka po novim korisnicima te onaj koji prikazuje distribuciju broja dobrih preporuka po ostalim korisnicima, prikazuju zadovoljavajuće rezultate. U oba grafa može se primijetiti kako metoda davanja preporuke daje relevantne rezultate. U situaciji kada su korisnici novi može se uočiti veći broj korisnika koji imaju velik broj dobrih preporuka. Razlog za to leži u činjenici da novi korisnici imaju veći broj filmova koji se mogu podudarati s preporukama.

Naime, cijela ideja evaluacije bila je odrediti vremenski trenutak i pretpostaviti da se nalazimo u njemu. Korisnici koji su klasificirani kao novi smatraju se novima u kontekstu tog specifičnog trenutka. Iz tog razloga, ti korisnici imaju puno filmova koji su postavljeni na *Nan* vrijednosti, odnosno filmova koje će tek pogledati u budućnosti, a koji se mogu podudarati s preporukama.

Ovakva evaluacija pokazala je zadovoljavajuće rezultate. Hibridni sustav daje preporuke koje su dovoljno specifične i dovoljno široke za svakog korisnika. Isto tako, obje verzije sustava, ona za nove korisnike i ona za ostale korisnike, funkcioniraju učinkovito.

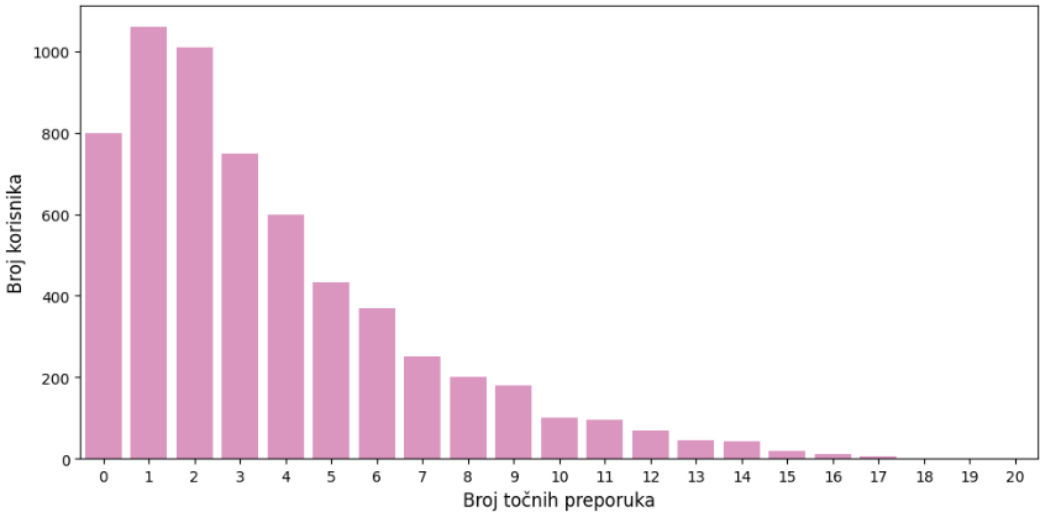
No, ovakvom metodom evaluacije, izgubljeno je dosta korisnika. Naime, kao što je već spomenuto, u ovu evaluaciju uključeno je 1394 korisnika, što čini oko dvadeset posto ukupnog broja korisnika. Bilo je poželjno napraviti i neka vrstu dodatne provjere koja obuhvaća sve korisnike.

Iz tog razloga, napravljena je još jedan korak u evaluaciji. Ovog puta, podaci su modi-

ficirani na način da je za svakog korisnika 30% posljednjih recenzija postavljeno na *NaN* vrijednosti. Tako se simulira situacija u kojoj ti filmovi još nisu pogledani. Preporuke se generiraju na temelju preostalih podataka i uspoređuju s filmovima koji su označenima kao ne pogledani. Zapravo je metoda vrlo slična prethodnoj, razlika je samo u načinu na koji se za svakog korisnika određuje koji filmovi su već pogledani te koji još nisu.

U ovom slučaju, podatci za svakog korisnika puno su stabilniji te je broj korisnika s malo povijesnih podataka vrlo nizak. To znači da se preporuke za gotovo sve korisnike generiraju standardnim putem, a ne korištenjem demografskih preporuka. Isto tako, svi korisnici imaju određeni broj filmova koji su označeni kao ne pogledani, no taj broj je manji nego u prethodnom postupku. Iz tog razloga očekivano je i da će broj filmova koji se podudaraju s preporukama biti manji.

Rezultati te evaluacije prikazani su na slici 5.6. Rezultati su zadovoljavajući jer se i dalje za većinu korisnika može uočiti jasna povezanost između preporuka i njihovih interesa. To pokazuje da sustav uspijeva prepoznati obrasce ponašanja i preferencije korisnika. Isto tako, nije slučaj da je velik broj preporuka previše specifičan za korisnike.



Slika 5.6. Distribucija broja točnih preporuka po svim korisnicima

5.3. Analiza povratnih informacija korisnika o kvaliteti preporuka

U prošlom poglavlju bila je riječ o evaluaciji koja se temeljila na podatcima o gledanim filmovima. Preporuke koje je generirao sustav bile su uspoređene sa stvarnim filmovima

koje su korisnici pogledali. Takva vrsta evaluacije dala je uvid u to da generirane preporuke dobro reflektiraju stvarne ukuse korisnika, no da one ipak nisu suviše specifične. Doduše, navedena metoda evaluacije nije nužno obuhvatila suštinu kvaliteta preporuka.

Kvaliteta preporuke treba odražavati koliko je preporučeni film privlačan korisnicima, odnosno vjerojatnost da će taj film rado odabrati i pogledati. Takva vrsta evaluacije preporuka može se napraviti tek kada je preporuka dana te kada postoji neka vrsta povratne informacije o interakciji s njom. Primjerice, jednostavna i korisna informacija u stvarnom svijetu bila bi odabir određene preporuke, odnosno gledanje tog filma. U ovom radu provedena je slična analiza na ostvarivoj razini.

Cilj je bio simulirati stvarno korištenje ovakvog sustava te od stvarnih korisnika prikupiti povratne informacije o kvalitetama preporuka. Svakoj osobi koja je željela sudjelovati u ovoj analizi poslana je lista filmova koje je trebalo ocijeniti od jedan do pet. Ocjenjeni filmovi zapravo su odražavali filmsku povijest tih osoba, odnosno njihov korisnički profil. Naravno, filmovi koje ispitanici nisu ocijenili smatrani su nepogledanima i nisu se koristili kao osnova za davanje preporuka. U ovakvoj situaciji nije bilo moguće pristupiti svim filmovima koje su osobe pogledale te je bilo potrebno na neki način prilagoditi pristup dobivanja povijesti gledanih filmova. U tu svrhu izdvojeno je 150 najgledanijih filmova iz skupa podataka te je za svakog korisnika nasumično odabrano 50 filmova iz te grupe. Iako takav pristup nije idealan, povećava vjerojatnost da su osobe pogledale film, odnosno daje više podataka o preferencijama.

Ti podatci su zatim korišteni za generiranje preporuka. Preporuke su generirane hibridnim sustavom tako da je pokriven i slučaj kada osoba nije gledala dovoljno ponuđenih filmova. Za svakog ispitanika, generirano je dvadeset preporuka koje su zatim dane na ocjenjivanje.

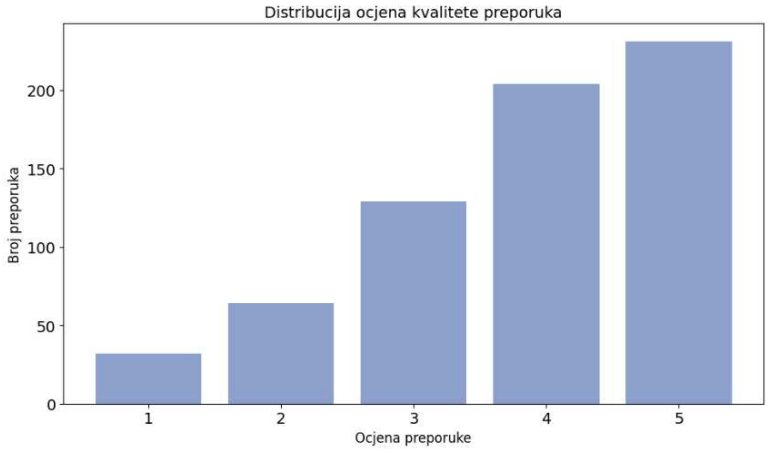
Uputa ispitanicima bila je ocijeniti kvalitetu preporuke od jedan do pet, odnosno ocijeniti koliko je preporučeni film privlačan. Niže ocjene označavale bi da film vjerojatno ne bi pogledali, dok bi više ocjene ukazivale na to da bi film rado pogledali. Ako ispitaniku nije poznat preporučen film, informacije o filmu trebao je prikupiti čitanjem opisa ili gledanjem najave filma. To je trebalo poslužiti kao osnova za donošenje odluke o tome bi li ispitanik zapravo pogledao preporučeni film, odnosno koliko mu je taj film

zanimljiv.

U ovoj analizi sudjelovale su 33 osobe. Nakon što su prikupljene povratne informacije ispitanika o kvalitetama preporuka, provedena je analiza kako bi se utvrdilo koliko su korisnici zadovoljni. Primaran cilj ovakve evaluacije bio je utvrditi imaju li svi ispitanici barem jedan film koji bi rado pogledali. Pretpostavka je da bi takav film tada postao dio povijesti interakcija te bi utjecao na generiranje budućih preporuka i poboljšao njihovu kvalitetu.

Iz povratnih informacija može se vidjeti da je zaista svakom ispitaniku preporučen barem jedan film koji bi rado pogledao. Ta činjenica sama po sebi je dobar pokazatelj određene kvalitete sustava. Isto tako, bitno je napomenuti kako su ispitanicima bili preporučeni i neki filmovi koji im nisu zanimljivi. Naravno takvih filmova bio je značajno manje.

Na slici 5.7. prikazana je distribucija ocjena kvalitete preporuka. Na osi x su prikazane ocjene a na y osi broj preporuka koje su dobile tu ocjenu. Iz grafa se može vidjeti da su preporuke većinom ocjenjene pozitivno, odnosno da su ispitanici uglavnom bili zadovoljni ovim preporukama.

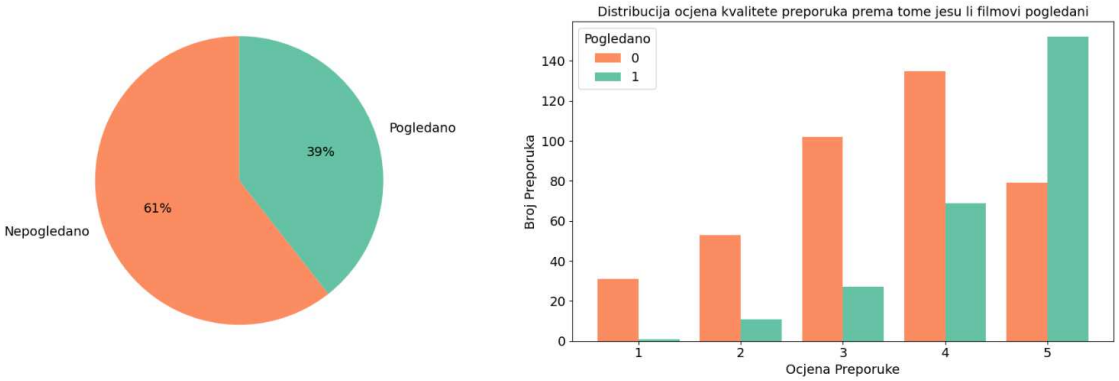


Slika 5.7. Distribucija ocjena kvalitete preporuka

Osim što su ocjenjivali koliko su zadovoljni preporukama, ispitanicima je rečeno da označe filmove koje su već pogledali. Ta informacija, s jedne strane, daje uvid u to koliko sustav uspješno pogađa ukus korisnika jer su određen broj preporučenih filmova korisnici zaista već pogledali. S druge strane, ipak je korisno vidjeti i kako se ispitanici

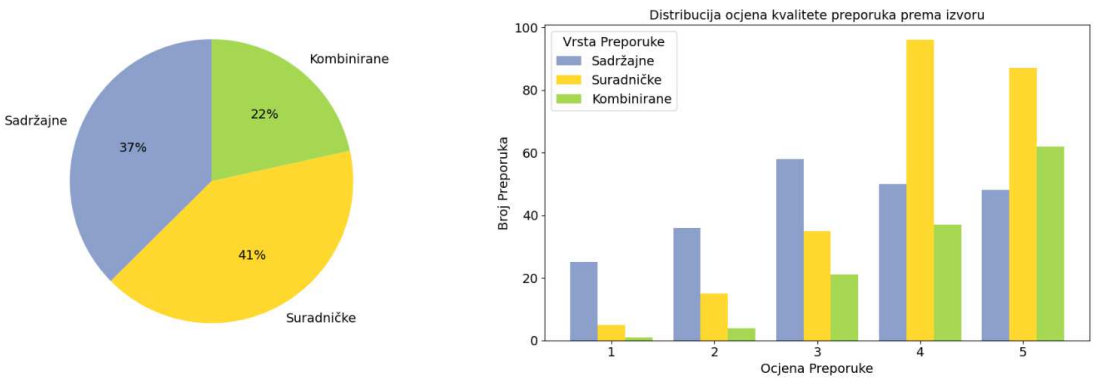
osjećaju prema potpuno novim filmovima te koliko su im oni zanimljivi.

Na slici 5.8. prikazana su dva grafa. Prvi pokazuje koliko su preporuka ispitanici već pogledali dok je na drugom prikazana distribucija ocjena kvalitete preporuka, podijeljena prema tome je li film pogledan ili nije.



Slika 5.8. Distribucija ocjena kvalitete preporuka prema tome jesu li filmovi pogledani

Kao što je već spomenuto, preporuke za ovu analizu generirane su na temelju hibridnog sustava. Takav sustav funkcionira tako da, ako korisnik nije nov, izračunava težinski ocjenu između sadržajnih i suradničkih preporuka te preporučuje filmove koji imaju najboljih 20 ocjena. Iz tog razloga, može se dogoditi da su generirane preporuke isključivo rezultat sadržajnog sustava te da suradničke preporuke nisu imale utjecaj na tu preporuku. To se događa kada filmovi imaju visoke ocjene u sadržajnom sustavu, ali nisu uopće preporučeni od strane suradničkog filtriranja. Također, moguća je i obrnuta situacija, kada preporuke dolaze isključivo od suradničkog sustava, a sadržajni sustav nije imao nikakav utjecaj. Kombinirane preporuke nastaju kada oba sustava izravno utječu na rezultat, odnosno kada su oba sustava preporučila isti film.



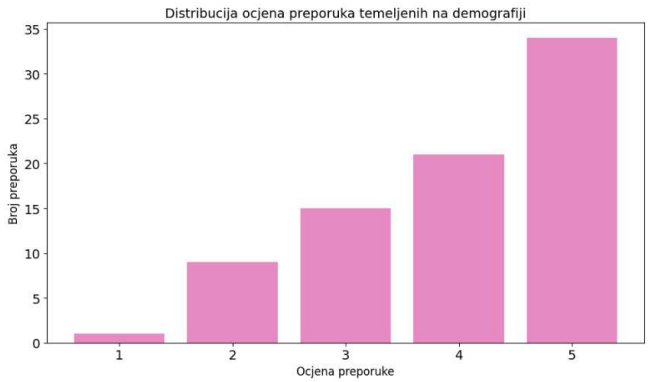
Slika 5.9. Distribucija ocjena kvalitete preporuka prema izvoru

Na slici 5.9. prikazana su dva grafa. Prvi graf prikazuje postotak preporuka prema određenom izvoru, odnosno broj preporuka koje su nastale kao sadržajne, suradničke ili kombinirane. Na drugom grafu prikazana je distribucija ocjena kvalitete preporuka prema izvoru.

Iz tog grafa može se uočiti kako su samostalne sadržajne preporuke najslabije ocijenjene od strane ispitanika. Takav rezultat mogao bi biti posljedica nedovoljne količine dostupnih sadržajnih informacija koje opisuju film. Naime, u ovoj implementaciji sustava za davanje preporuka temeljenih na sadržaju, kao osnova za mjeru sličnosti korišteni su žanrovi. No, te preporuke bi potencijalno bile kvalitetnije kada bi se u obzir uzele još neke karakteristike filmova, poput redatelja ili glumca, koji nisu bili dostupne u ovom skupu podataka.

Međutim, vrlo je bitno uočiti da kombinirane preporuke, iako ih je brojčano najmanje, daju najbolje rezultate. Preporuke koje su preporučene od strane oba sustava, sadržajnog i suradničkog, ispitanicima su se pokazale kao najzanimljivije. Iz tog razloga, pogrešno je zaključiti kako sadržajne preporuke nemaju značajnu ulogu. Iako samostalno ne pokazuju određene rezultate, u kombinaciji sa suradničkim sustavom daju vrlo kvalitetne preporuke.

Kao, što je već spomenuto, u sklopu hibridnog sustava implementiran je i sustav temeljen na demografiji koji se koristi kada korisnici nemaju dovoljno povijesnih informacija. U provedenoj analizi, prilikom prikupljanja recenzija filmova, četiri ispitanika nisu gledala gotovo nijedan film te ih zbog toga nisu mogli ocijeniti. Iz tog razloga, preporuke za njih nisu mogle biti generirane standardnim načinom, već je primijenjen demografski pristup.



Slika 5.10. Distribucija ocjena kvalitete preporuka temeljenih na demografiji

Na slici 5.10. prikazana je distribucija ocjena kvalitete preporuka temeljenih na demografiji. Iz slike se može uočiti da su ispitanici demografskim preporukama većinom dali visoke ocjene. Ta informacija je vrlo bitna jer pokazuje da čak i za korisnike koji nemaju povijesnih informacija, odnosno za nove korisnike, sustav daje relevantne preporuke.

Provedena analiza pokazuje da sustav generira preporuke koje korisnici generalno smatraju vrlo zanimljivima. Naravno, ova analiza predstavlja samo simulaciju stvarnog korištenja sustava. Iako je opseg mali i uvjeti nisu idealni, ova analiza pokazuje kako korisnici smatraju ove preporuke vrijedim izvorom pri odabiru filmova za gledanje.

6. Zaključak

U ovom radu riječ je bila o sustavima za davanje preporuka. Prikazana je njihova podjela te su objašnjeni principi funkcioniranja tih sustava. Analizirani su različiti pristupi preporučivanju, uključujući sustave temeljene na sadržaju i suradničke sustave. Implementirani sustavi temeljeni na sadržaju i na suradnji imali su određene prednosti i mane.

Sustav temeljen na suradnji generira preporuke koje su izravno povezane s onim što je korisnik već pogledao. Dobra strana takvog pristupa je što te preporuke odražavaju preferencije korisnika, čime bi se trebala postići relevantnost. U implementiranom sustavu se kao karakteristika za mjerenje sličnosti uzimaju žanrovi filmova. Prilikom analize preporuka samostalnog sustava, može se uočiti da se te preporuke zaista žanrovski ne odmiču od gledanih filmova. Međutim, analiza korisničkih povratnih informacija pokazuje da, unatoč uskoj povezanosti, preporučeni filmovi nisu uvijek iznimno relevantni zbog široke definicije pojedinih žanrova. Iz tog razloga, bilo bi zanimljivo vidjeti kako bi takav sustav funkcionirao koristeći neki skup podataka koji uključuje dodatne karakteristike filmova. Pri tome treba paziti jer takvi sustavi mogu imati problema s prevelikom specifičnosti preporuka.

Sustavi temeljeni na suradnji funkcioniraju na način da pronalaze slične korisnike i na temelju njihovih preferencija generiraju preporuke. Pretpostavka je da će se korisnicima s prethodno sličim obrascima ponašanja svidjeti isti filmovi. Ovakve preporuke su generalno manje specifične za ciljanog korisnika. Takva raznolikost može predstavljati problem ako je prevelika. U implementiranom sustavu, korisnici su izrazili zadovoljstvo suradničkim preporukama. Iz tog razloga može se zaključiti da te preporuke nemaju pretjerano izražen spomenuti problem.

U slučaju oba sustava, onom temeljenom na sadržaju i onom temeljen na suradnji

uočeno je kako imaju određenih problema s hladnim startom. Sustav temeljen na sadržaju je prihvatljivo funkcionirao s hladnim startom proizvoda, ali pri hladnom startu korisnika nastali su problemi. Suradnički sustav je imao probleme u oba slučaja. Iz tog razloga ispitani su alternativni pristupi koji nisu imali problema s hladnim startom. Najefikasniji su s pokazali sustavi temeljeni na popularnosti i oni na demografiji. Ipak, pristup temeljen na demografiji omogućuje personalizirani pristup.

Za svaki sustav utvrđene su određene prednosti i nedostaci, što je potaknulo ideju o kombinaciji različitih pristupa na način koji bi iskoristio prednosti svakog sustava. Takav hibridni sustav implementiran je tako da težinski kombinira ocjene sadržajnog i suradničkog sustava, pri ćemu se u sadržajnom sustavu koristi *popularity* sustav kao mehanizam za razrješavanje izjednačenja.

Taj sustav evaluiran je vrstom *out of time* evaluacije koja je pokazala da su generirane preporuke vezane uz ukus i preferencije korisnika. Osim toga, napravljena je analiza s povratnim informacijama stvarnih korisnika koji su ocjenjivali kvalitetu preporuke. Iz te analize uočeno je da sadržajni sustav samostalno postiže nešto slabije rezultate. S druge strane, samostalni suradnički pristup pokazao se učinkovitijim. No, iznimno je bitno to što su korisnici kao najbolje, ocijenili preporuke koje su nastale kombiniranjem sadržajnog i suradničkog sustava.

Isto tako, iz povratnih informacija ispitanika, može se vidjeti da su demografske preporuke također postigle izvrstan rezultat. S obzirom na to da su korisnici vrlo pozitivno ocijenili demografske preporuke, bilo bi zanimljivo napraviti daljnju analizu tih preporuka. Ukoliko bi se rezultati također pokazali ovako pozitivnim, mogla bi se ispitati mogućnost kombiniranja demografskih preporuka i u standardnom pristupu, a ne samo kod novih korisnika.

U konačnici, implementacija te zatim analiza raznih sustava za davanje preporuka, pokazala je kako svaki sustav ima svoje prednosti i nedostatke. Međutim, pravilna kombinacija sustava ističe njihove kvalitete. Taj rezultat najbolje se vidi u zadovoljstvu korisnika.

Literatura

[1] Vatsal. (2021) Recommendation systems explained. [Mrežno]. Adresa: https://towardsdatascience.com/recommendation-systems-explained-a42fc60591ed

[2] M. Techlabs. (2021) Types of recommendation systems and their use cases. [Mrežno]. Adresa: https://medium.com/mlearning-ai/what-are-the-types-of-recommendationsystems-3487cbafa7c9

[3] B. Rocca. (2019) Introduction to recommender systems. [Mrežno]. Adresa: https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada

[4] B. O. F.O. Isinkaye, Y.O. Folajimi. (2015) Recommendation systems: Principles, methods and evaluation. [Mrežno]. Adresa: https://www.sciencedirect.com/science/article/pii/S1110866515000341

[5] F. Casalegno. (2022) Recommender systems – a complete guide to machine learning models. [Mrežno]. Adresa: https://towardsdatascience.com/recommender-systems-a-complete-guide-to-machine-learning-models-96d3f94ea748

[6] G. Simsek. (2018) Recommendation systems. [Mrežno]. Adresa: https://softwareengineeringdaily.com/2018/10/24/recommendation-systems-by-gokhan-simsek/

[7] A. E. Nixon. (2020) Building a movie content based recommender using tf-idf. [Mrežno]. Adresa: https://towardsdatascience.com/content-based-recommender-systems-28a1dbd858f5

[8] T. U. Team. (2024) What content-based filtering is and why you should use it. [Mrežno]. Adresa: https://www.upwork.com/resources/what-is-content-based-

filtering

[9] A. Ajitsaria. (2019) Build a recommendation engine with collaborative filtering. [Mrežno]. Adresa: https://realpython.com/build-recommendation-engine-collaborative-filtering/

[10] J. Wilk. (2023) How to build a movie recommendation system based on collaborative filtering. [Mrežno]. Adresa: https://www.freecodecamp.org/news/how-to-build-a-movie-recommendation-system-based-on-collaborative-filtering/

[11] T. Khan. (2020) Relationship between cosine similarity and euclidean distance. [Mrežno]. Adresa: https://medium.com/ai-for-real/relationship-between-cosine-similarity-and-euclidean-distance-7e283a277dff

[12] G. D. Luca. (2024) Euclidean distance vs cosine similarity. [Mrežno]. Adresa: https://www.baeldung.com/cs/euclidean-distance-vs-cosine-similarity

[13] K. Bondarenko. (2019) Similarity metrics in recommender systems. [Mrežno]. Adresa: https://bond-kirill-alexandrovich.medium.com/similarity-metrics-in-recommender-systems-aed9d3b2315f

[14] Prasad. (2023) Understanding pearson and cosine similarity. [Mrežno]. Adresa: https://prrasad.medium.com/understanding-pearson-and-cosine-similarity-f7a9afc22e1a

[15] Z. Deutschman. (2023) Recommender systems: Machine learning metrics and business metrics. [Mrežno]. Adresa: https://neptune.ai/blog/recommender-systems-metrics

[16] M. Milankovich. (2015) The cold start problem for recommender systems. [Mrežno]. Adresa: https://medium.com/@markmilankovich/the-cold-start-problem-for-recommender-systems-89a76505a7

[17] A. Bojic. (2024) How to solve the cold start problem in recommender systems. [Mrežno]. Adresa: https://thingsolver.com/blog/the-cold-start-problem/

[18] A. Klacanova. (2022) Context-aware and hybrid recommendations. [Mrežno]. Adresa: https://graphaware.com/blog/content-aware-and-hybrid-recommendations/

[19] S. Gowthaman. (2024) Demographic based recommendation system. [Mrežno]. Adresa: https://sajithgowthaman.medium.com/demographic-based-recommendation-system-b89b490eb220

[20] J. chiang. (2021) 7 types of hybrid recommendation system. [Mrežno]. Adresa: https://medium.com/analytics-vidhya/7-types-of-hybrid-recommendation-system-3e4f78266ad8

Sažetak

Usporedba kontekstno utemeljenih i kolaborativnih sustava za davanje preporuka te implementacija hibridnog sustava

Dina Petrak

U ovom radu opisan je rad sustava za davanje preporuka te je dana osnovna podjela sustava. Opisano je funkcioniranje sadržajnog sustava i sustava temeljenog na suradnji. Ti sustavi implementirani su korištenjem različitih metrika te su uspoređene performanse. Analizirane su prednosti i nedostatci svakog pristupa. Posebni interes je usmjeren na problem hladnog starta te funkcioniranje sustava u situacijama kada nedostaje podataka. Ispitane su i implementirane alternativne verzije sustava koje se dobro nose s manjkom podataka. Implementiran je hibridni sustav koji težinski kombinira sadržajne i suradničke preporuke, te koristi demografske preporuke u situaciji novih korisnika. Hibridni sustav testiran je evaluacijom izvan vremenskog okvira. Osim toga, provedena je analiza zadovoljstva korisnika s preporukama, iz koje su izvedeni zaključci o kvalitetama i nedostatcima sustava koji su kombinirani te hibridnog sustava.

Ključne riječi: Sustav za filmske preporuke, preporuke temeljene na sadržaju, kolaborativno filtriranje, problem hladnog starta, demografske preporuke, hibridni sustav preporuka

Abstract

Comparison of content-based and collaborative recommendation systems and implementation of a hybrid system

Dina Petrak

This paper describes how recommendation systems work and depicts the main classification. The operation of content-based recommendation and collaborative filtering is explained. Both systems were implemented using different metrics and their performances were compared. The advantages and disadvantages of each approach were analyzed. The main focus was directed to the cold start problem and the functioning of the systems in situations of data shortage. Alternative versions of the recommendation system that cope well with lack of data have been tested and implemented. A hybrid system has been implemented that combines content and collaborative recommendations by a weighted approach and uses demographic recommendations in the situation of new users. The hybrid system was tested with out-of-time evaluation. In addition, an analysis of user satisfaction with recommendations was carried out, from which conclusions were drawn about the qualities and shortcomings of the combined systems and the hybrid system.

Keywords: Movie recommendation system, content-based recommendation, collaborative filtering, cold start problem, demographic recommendation, hybrid recommendation system