SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Sustavi za davanje preporuka

Dina Petrak
Voditelj: *Marko Đurasević*

Zagreb, svibanj 2023.

SADRŽAJ

1. Uvod **1**

2. Vrste sustava za davanje preporuka **2**

2.1. Sustavi temeljeni na suradnji 3

2.2. Sustavi temeljeni na sadržaju 5

3. Analiza implementacije *content based recommendera* **7**

4. Zaključak **11**

5. Literatura **12**

6. Sažetak **14**

1. Uvod

Napredak u tehnologiji uzrokovao je eksplozivni rast količine podataka koju svakodnevno generiramo i pohranjujemo. Taj trend doveo je do potrebe za što boljim upravljanjem informacija. Naime, često je teško odrediti koje su informacije relevantne i vrijedne pažnje, a koje nisu. Ovaj problem filtriranja izražen je i u svakodnevnom životu.

U svijetu u kojem sve više ovisimo o internetskim platformama za kupovinu i aplikacijama za slušanje glazbe ili gledanje filmova, suočavamo se s neprestanim izborom. Naravno, na prvu bi se moglo pomisliti da je velika količina opcija dobra stvar, ali previše opcija može dovesti do pojave koja se naziva paraliza odlučivanja.

Sustavi za davanje preporuka ublažavaju preopterećenost informacijama filtriranjem, prioritiziranjem i učinkovitim dostavljanjem relevantnih informacija. Uspješno rješavaju problem velikog volumena proizvoda te korisnicima pružaju personalizirani sadržaj i usluge. Upravo zbog toga, takvi sustavi sastavni su dio mnogih platformi.

Sustavi za preporuke su podvrsta strojnog učenja te se temelje na davanju 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.

Naravno, takvi sustavi pomažu korisniku da se snađe u velikoj količini informacija, ali uvelike koriste i pružateljima usluga. Primjerice, kvalitetne preporuke gledateljima *Netflixa* pridonose pozitivnom iskustvu korisnika te ih zadržavaju upravo na toj platformi.

U ovom radu proučeni su sustavi za davanje preporuka. Dana je njihova osnovna podjela te je opisan način rada glavnih vrsta sustava za davanje preporuka. Osim toga provedena je analiza jednostavne implementacije sustava temeljenog na sadržaju objekta.

2. Vrste sustava za davanje preporuka

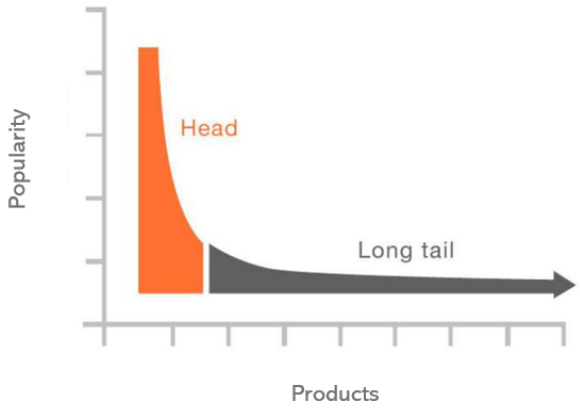
Sustavi za davanje preporuka sužavaju izbor te pomažu predvidjeti, odnosno pronaći ono što korisnik traži u velikom broju opcija. Takvi sustavi neizbježan su dio mnogih platformi te mogu uključivati gotovo bilo kakav proizvod.

No, postavlja se pitanje kako takvi sustavi zapravo rade?

Postoji puno tehnika i načina za implementaciju takvih sustava no najosnovnija podjela je na personalizirane i nepersonalizirane sustave.

Nepersonalizirani sustavi su jednostavniji jer ne uzimaju u obzir karakteristike svakog korisnika već daju preporuke temeljene na generalnom iskustvu. Primjer takvog sustava su *Popularity Recommenders*. Takvi sustavi preporučivat će artikle koji su najpopularniji ili najbolje ocijenjeni u tom trenutku. Prednost takvih sustava je jednostavnost jer se podatci ne moraju analizirati za svakog korisnika posebno. No, nepersonalizirane preporuke su manje precizne odnosno generalno lošije za korisnika te manje produktivne za nuditelje usluga.

Osim toga, nepersonalizirani sustavi često preporučuju popularne proizvode koji tako postaju sve popularniji. Taj problem naziva se problem dugog repa ili *long tail* problem.



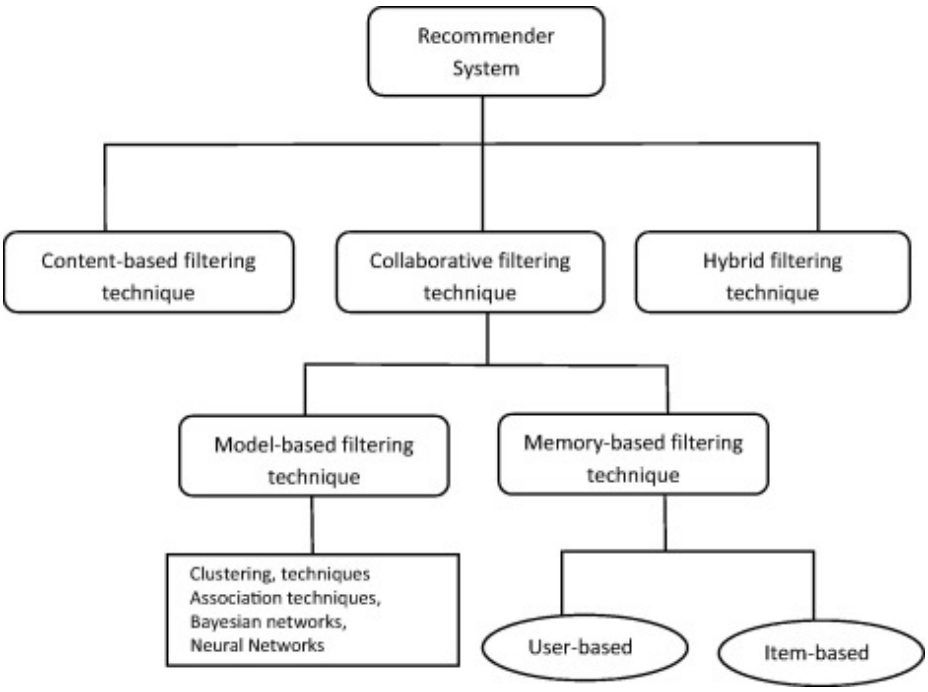
Slika 2.1: Problem dugog repa [9]

Kao što je prikazano na slici 2.1 većina preporuka usmjerena je na popularne stavke koje čine "glavu" distribucijske krivulje, dok manje popularne stavke čine "dugi rep" distribucije. To predstavlja problem jer se nepersonalizirani sustavi preporuka fokusiraju samo na popularne stavke, dok stavke u dugom repu ostaju nevidljive korisnicima.

Jedan od načina za rješavanje problema dugog repa u sustavima za davanje preporuka je korištenje algoritama koji uzimaju u obzir i manje popularne stavke te koji se fokusiraju na proizvode relevantne upravo tom korisniku.

Takvi sustavi nazivaju se personalizirani sustavi.

Personalizirani sustavi dijele se 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*. Osnovne podjele prikazane su na slici 2.2.

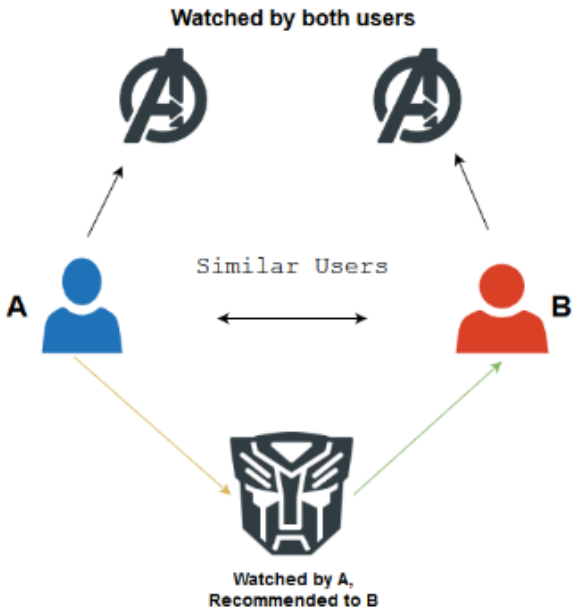


Slika 2.2: Podjela sustava za davanje preporuka [5]

2.1. Sustavi temeljeni na suradnji

Sustavi temeljeni na suradnji koriste metode koje se temelje isključivo na zabilježenim interakcijama između korisnika i stavki te uz pomoć njih stvaraju nove preporuke. Te interakcije pohranjene su u matrici interakcija.

Glavna ideja tih metoda je da su prošle interakcije između korisnika i stavki dovoljne za detektiranje sličnih korisnika ili sličnih stavki te da se na temelju tih procijenjenih sličnosti efikasno mogu napraviti predviđanja. Primjerice ako korisnici A i B imaju sličan ukus za određeni proizvod, tada je vjerojatno da će A i B imati sličan ukus i za druge proizvode. Rad takvih sustava prikazan je na slici 2.3.



Slika 2.3: Rad sustava temeljenog na suradnji [2]

Sustavi temeljeni na suradnji podijeljeni su na dvije potkategorije koje se općenito nazivaju pristupi temeljeni na memoriji i pristupi temeljeni na modelima.

Pristupi temeljeni na memoriji rade izravno s vrijednostima zabilježenih interakcija, ne pretpostavljajući model. U osnovi, daju predikcije temelje na pretraživanju najbližih susjeda. Filtriranje temeljeno na korisnicima izračunava koju bi ocjenu neki korisnik dao određenom proizvodu tako što u obzir uzima kakve ocijene su toj stavci dali njemu slični korisnici. Filtriranje temeljeno na stavkama ocijene proizvoda predviđa ovisno o tome kako su slični proizvodi ocijenjeni od strane tog korisnika.

Pristupi temeljeni na modelima koriste strojno učenje tako što izrađuju model te preporuke temelje na predikcijama tog modela. Značajke povezane s *datasetom* parametriziraju se kao ulazi modela kako bi se pokušao riješiti problem optimizacije. Pristupi temeljeni na modelima uključuju korištenje stvari poput stabala odlučivanja, neuronskih mreža te bayesovske statistike.

Glavna prednost kolaborativnih pristupa je što ne zahtijevaju nikakve informacije o korisnicima ili stavkama, tako da ih se može koristiti u mnogim situacijama. Osim

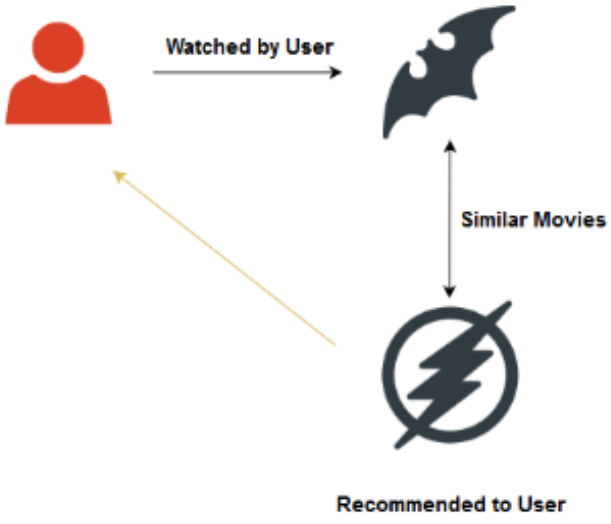
toga, što više korisnici stupaju u interakciju sa stavkama, to će nove preporuke postajati točnije. Nove interakcije zabilježene tijekom vremena donose nove informacije i čine sustav učinkovitijim.

Međutim, budući da se za davanje predikcija koriste samo prošle interakcije, sustavi temeljeni na suradnji pate od problema hladnog pokretanja. Takvi sustavi teško preporučuju nove stavke jer za njih ne postoje zabilježene interakcije.

2.2. Sustavi temeljeni na sadržaju

Za razliku od sustava temeljenih na suradnju koji se oslanjaju samo na interakcije između korisnika i stavki, sustavi temeljeni na sadržaju koriste dodatne informacije o karakteristikama korisnika ali i proizvoda. Riječ sadržaj u samom nazivu ovih sustava odnosi se na atribute koji opisuju stavke i korisnike.

Dakle, osnovna ideja sustava temeljenih na sadržaju je da se preporuke generiraju na osnovu korisničkog profila. Takav sustav uzima podatke koje korisnik pruža na platformi, primjerice ocjenjivanjem, te na temelju tih podataka generira profil koji kasnije koristi za izradu predikcija. Rad takvog sustava prikazan je slikom 2.4.



Slika 2.4: Rad sustava temeljenog na sadržaju [2]

Sustavi temeljeni na suradnji zahtijevaju izvore podataka za definiranje i pohranjivanje atributa proizvoda. Primjerice prilikom davanja preporuka za knjige, bilo bi potrebno poznavati žanr knjige, godinu izdavanja ili cijenu. Osim toga, takvi sustavi

uključuju i izvor podataka za korisnike, odnosno zabilježene povratne informacije o proizvodima.

Sustavima temeljenim na sadržaju objekta moguće je izbjeći problem hladnog pokretanja. Iako ti sustavi zahtijevaju neke početne ulazne podatke kako bi mogli davati preporuke, kvaliteta ranijih predikcija općenito je bolja od sustava temeljenih na suradnji koji zahtijevaju korelacije iznimno velikog broja podataka. Osim toga, velika prednost tih sustava leži i u činjenici da za davanje uspješnih preporuka nisu potrebni nikakvi podatci od drugih korisnika. Preporuke su obično vrlo prilagođene interesima korisnika te su iz tog razloga visoko relevantne.

Naravno, negativna strana toliko dobro prilagođenih preporuka je činjenica da sam korisnik može pretpostaviti da će mu se stavka svidjeti. Primjerice ako ima omiljenog autora nije mu potrebna preporuka sustava da pročita još jednu njegovu knjigu. Dakle, u sustavima temeljenim na sadržaju može se pojaviti nedostatak novosti i raznolikosti. Nadalje, velik nedostatak tih sustava je potrebna za konstantnim definiranjem i dodavanjem atributa. Svaki put kada se pojavi novi korisnik ili doda nova stavka potrebno je detektirati njihove karakteristike i efikasno ih pohraniti.

U ovom radu obrađena je jednostavna implementacija sustava temeljenog na sadržaju objekta.

3. Analiza implementacije *content based recommendera*

Jedan od najčešće korištenih skupova podataka za testiranje sustava preporuka je *MovieLens* skup podataka koji sadrži podatke o ocjenama filmova s *MovieLens* web stranice. U ovoj analizi implementacije korišten je skup podataka koji sadrži 1 milijun anonimnih ocjena od strane 6000 korisnika *MovieLensa* za otprilike 4000 filmova.

Skup podataka sastoji se od datoteka s ocjenama, korisnicima i filmovima, odnosno *ratings*, *users* i *movies* datoteka. Podatci su učitani uz pomoć *Pandas* biblioteke te je zbog uvida strukture i sadržaja skupova podataka ispisano prvih 5 primjeraka svake datoteke. Ispisi za *ratings*, *users* i *movies* su redom prikazani na slikama 3.1, 3.2 i 3.3.

	user_id	movie_id	rating
0	1	1193	5
1	1	661	3
2	1	914	3
3	1	3408	4
4	1	2355	5

Slika 3.1: Datoteka *ratings*

	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*

	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 3.3: Datoteka *movies*

Kao što je već napomenuto, sustavi temeljeni na sadržaju generiraju preporuke temeljem preferencija korisnika odnosno njegova profila. Pokušavaju pronaći stavke koji odgovaraju korisnikovim ranijim preferencijama. Razina sličnosti između predmeta obično se utvrđuje na temelju atributa stavki koje su se svidjele korisniku.

Promatranjem i analizom danih podataka, uviđeno je da se žanrovi mogu koristiti kao atributi koji bi predstavljali filmove. Drugim riječima, žanrovi sami po sebi mogu biti dovoljni za stvaranje učinkovitih i relevantnih preporuka.

Relevantan aspekt koji treba uzeti u obzir prilikom izgradnje takve vrste preporučitelja jest popularnost žanrova. Naime, cilj je razumjeti koji su žanrovi stvarno relevantni kada se radi o definiranju ukusa korisnika. Razumna pretpostavka je da će upravo nepopularni žanrovi biti važniji u karakteriziranju korisnikovih preferencija.

Primjerice, ako se korisniku svidio film *Interstellar* koji je kombinacija znanstvene fantastike, drame, misterije i avanture, cilj je stvoriti preporuku za film slične kombinacije žanrova. U tom slučaju, manje generički žanr (znanstvena fantastika) treba imati veću težinu pri preporučivanju.

Za izradu jednostavnog sustava za davanje preporuka, korišten je *tf-idf* vektor, čiji je izraz dan formulom 3.1.

$$\text{tf-idf}_{i,j} = \text{tf}_{i,j} \cdot \log \frac{N}{\text{df}_i} \tag{3.1}$$

$\text{tf}_{i,j}$ = frekvencija pojave i u dokumentu j
df_i = broj dokumenata koji sadrže i
N = ukupan broj dokumenata

Formula prikazuje frekvenciju termina, odnosno broj puta koliko se određeni žanr pojavljuje u dokumentu, koji se u osnovi skalira ovisno o tome koliko se puta određeni žanr pojavljuje u svim dokumentima. Što je manji broj filmova koji sadrže određeni žanr, to je veća rezultirajuća težina. Logaritam služi za glađenje rezultata dijeljenja.

Dakle, *tf-idf* pomaže u prepoznavanju relevantnosti žanrova tako da veću težinu daje manje popularnim žanrovima.

Središnja pretpostavka sustava temeljenog na sadržaju je da će se korisniku svidjeti slične stavke. Dakle, nakon što su žanrovi, a samim time i filmovi kodirani *tf-idf* reprezentacijom, potrebno je pronaći slične *tf-idf* vektore.

Mjera sličnosti koja se koristi naziva se kosinusna udaljenost te je izražena formulom 3.2.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{3.2}$$

Ova mjera sličnosti dobila je svoje ime zbog činjenice da je jednaka kosinusu kuta između dva uspoređivana vektora. Što je manji kut između dva vektora, to će kosinus biti veći, što rezultira većim faktorom sličnosti.

Da bismo izračunali kosinusne sličnosti između svih *tf-idf* vektora možemo koristiti *scikit-learn*. Ta biblioteka sadrži mnoge metrike za računanje udaljenosti, među kojima je i kosinusna sličnost. U ovom slučaju potrebno je izračunati udaljenost između svih ulaznih redaka, odnosno *tf-idf* vektora. Te udaljenosti pohranjene su u matricu koja je prikazana slikom 3.4.

title	Spiral Staircase, The (1946)	Insider, The (1999)	Father of the Bride (1950)	Children of the Damned (1963)	Atlantic City (1980)
title					
Toy Story (1995)	0.00	0.00	0.17	0.00	0.00
Jumanji (1995)	0.00	0.00	0.00	0.00	0.00
Grumpier Old Men (1995)	0.00	0.00	0.40	0.00	0.14
Waiting to Exhale (1995)	0.00	0.39	0.45	0.00	0.06
Father of the Bride Part II (1995)	0.00	0.00	1.00	0.00	0.00
...	...	...	...	...	...
Meet the Parents (2000)	0.00	0.00	1.00	0.00	0.00
Requiem for a Dream (2000)	0.00	1.00	0.00	0.00	0.14
Tigerland (2000)	0.00	1.00	0.00	0.00	0.14
Two Family House (2000)	0.00	1.00	0.00	0.00	0.14
Contender, The (2000)	0.53	0.32	0.00	0.13	0.05

Slika 3.4: Matrica sličnosti

Iz dane matrice vidljivo je da je film *Father of the Bride* vrlo sličan *Father of the Bride Part II*, što je očekivano jer su nastavci filmova obično istih karakteristika te samim time i istih žanrova. Osim toga može se uočiti da su filmovi *The Contender* i *The Spiral Staircase* također sliči ali u manjoj mjeri nego prijašnji primjer. Naime oba filma su trileri, no to je dosta čest žanr u cijelom skupu podataka. S druge strane, animirani obiteljski film *Toy Story* i horor film *Children of the Damned* nisu nimalo slični.

Za zadani film, sustav prolazi po matrici sličnosti te određuje 10 stavki s najvećom vrijednosti i njih daje kao preporuke. Primjerice za film *A Space Odyssey* prikazan slikom 3.5, implementirani sustav daje predikcije prikazane slikom 3.6.

	movie_id	title	genres
912	924	2001: A Space Odyssey (1968)	Drama|Mystery|Sci-Fi|Thriller

Slika 3.5: Primjer relevatnog filma

Kao što je i očekivano, najviše sličnosti imaju filmovi koji dijele najviše žanrova. Zanimljivo je primijetiti da su većina njih znanstveno fantastični filmovi. Razlog tomu,

	title	genres
0	X-Files: Fight the Future, The (1998)	Mystery|Sci-Fi|Thriller
1	Client, The (1994)	Drama|Mystery|Thriller
2	Talented Mr. Ripley, The (1999)	Drama|Mystery|Thriller
3	Communion (1989)	Drama|Sci-Fi|Thriller
4	Gattaca (1997)	Drama|Sci-Fi|Thriller
5	Thirteenth Floor, The (1999)	Drama|Sci-Fi|Thriller
6	Event Horizon (1997)	Action|Mystery|Sci-Fi|Thriller
7	2010 (1984)	Mystery|Sci-Fi
8	Stalker (1979)	Mystery|Sci-Fi
9	Deep Impact (1998)	Action|Drama|Sci-Fi|Thriller

Slika 3.6: Preporuke dobivene temeljem zadanog filma

vrlo vjerojatno, jest činjenica da je znanstvena fantastika jedan od najmanje uobičajenih žanrova u cijelom skupu podatak pa zato ima i veću težinu.

4. Zaključak

Današnje digitalno doba omogućava pristup iznimno velikoj količini informacija što, iako se na prvo čini pozitivno, često dovodi do preopterećenosti. Sustavi za davanje preporuka omogućavaju novu vrstu personaliziranog pretraživanja te tako rješavaju problem preopterećenosti informacijama. Oni korisnicima omogućuju pristup proizvodima i uslugama koje bi inače bili izgubljeni u prevelikom bazenu opcija.

Upravo iz tog razloga, uključivanje sustava za preporuke na platforme, vrijedna je investicija. Sustavi za preporuke ne poboljšavaju samo korisničko iskustvo i angažman, već i generiraju više prihoda za poslovanja.

U ovom radu dana je osnovna podjela personaliziranih sustava za davanje preporuka na one temeljene na sadržaju i one temeljene na suradnji. Za oba načina davanja preporuka objašnjen je osnovan princip rada.

Osim toga, prikazana je jednostavna implementacija sustava na temelju sadržaja te je testirana na *MovieLens* skupu podataka. Iako je takva implementacija zaista jednostavna, daje logične i prihvatljive rezultate.

Velika prednost preporuka temeljenih na sadržaju je da ne pate od problema hladnog starta, jer nam je potrebna samo osnovna informacija o korisniku. U ovom slučaju, temeljem samo jednog filma, sustav je dao prilično dobre rezultate.

Međutim, velik nedostatak takvih sustava je njihova tendencija preporučivanja iste vrste sadržaja. Kako bi mogli preporučiti novu vrstu sadržaja, korisnik bi sam morao pronaći novu stavku i pokazati interes za nju.

Iako i sustavi temeljeni na suradnji i sustavi temeljeni na sadržaju imaju svoje prednosti i nedostatke, oni su neizostavan dio mnogih platformi. S pojavom dubokog učenja i drugih naprednih algoritama, područje sustava za davanje preporuka se neprestano razvija. Stoga će biti uzbudljivo vidjeti kako se ovi sustavi nastavljaju transformirati i poboljšavati u budućnosti.

5. Literatura

[1] Shubham Kumar Agrawal. Recommendation system - understanding the basic concepts, 2021. URL https://www.analyticsvidhya.com/blog/2021/07/recommendation-system-understanding-thebasic-concepts/.

[2] Shivaadith Anbarasu. Content-based recommender system, 2023. URL https://pianalytix.com/content-based-recommender-system/.

[3] Shuvayan Das. Beginners guide to learn about content based recommender engines, 2015. URL https://www.analyticsvidhya.com/blog/2015/08/beginners-guide-learn-content-based-recommender-systems/.

[4] Zuzanna Deutschman. Recommender systems: Machine learning metrics and business metrics, 2023. URL https://neptune.ai/blog/recommender-systems-metrics.

[5] B.A. Ojokoh F.O. Isinkaye, Y.O. Folajimi. Recommendation systems: Principles, methods and evaluation, 2015. URL https://www.sciencedirect.com/science/article/pii/S1110866515000341.

[6] Alexandre Escolà Nixon. Building a movie content based recommender using tf-idf, 2020. URL https://towardsdatascience.com/content-based-recommender-systems-28a1dbd858f5.

[7] Baptiste Rocca. Introduction to recommender systems, 2019. URL https://towardsdatascience.com/introduction-to-recommender-systems-6c66cf15ada.

[8] Natalie Severt. An introduction to recommender systems, 2023. URL https://www.iteratorshq.com/blog/an-introduction-recommender-systems-9-easy-examples/.

[9] Gokhan Simsek. Recommendation systems, 2018. URL https://softwareengineeringdaily.com/2018/10/24/recommendation-systems-by-gokhan-simsek/.

[10] The Upwork Team. What content-based filtering is and why you should use it, 2021. URL https://www.upwork.com/resources/what-is-content-based-filtering.

[11] Maruti Techlabs. Types of recommendation systems and their use cases, 2021. URL https://medium.com/mlearning-ai/what-are-the-types-of-recommendationsystems-3487cbafa7c9.

[12] Vatsal. Recommendation systems explained, 2021. URL https://towardsdatascience.com/recommendation-systems-explained-a42fc60591ed.

6. Sažetak

Sustavi za davanje preporuka neizostavan su dio iskustva na internetu te su odlično rješenje za problem preopterećenosti informacijama. U ovom radu navedena je osnovna podjela sustava za davanje preporuka za one temeljene na suradnji te one temeljene na sadržaju. Za svaku od podvrsta opisan je osnovni princip rada. Osim toga, analizirana je jednostavna implementacija sustava temeljenog na sadržaju. Implementacija je testirana na MovieLens skupu podataka koji se sastoji od korisnika, ocjena i filmova te je pokazano kako se novi filmovi mogu preporučiti na temelju njihovih atributa. Iako je implementacija zaista jednostavna, ima potencijala dati prilično dobre rezultate.