SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

SEMINAR

Pregled pristupa izrade sustava preporuke

Daniel Guja
Voditelj: *prof. dr. sc. Siniša Srbljić*

Zagreb, svibanj 2016.

SADRŽAJ

1. Uvod

Količina sadržaja dostupnih na internetu svakim danom raste što rezultira problemom preopterećenja informacijama te je vrlo teško pronaći relevantne sadržaje. Primjerice, proizvodi koji se mogu kupiti na internet trgovinama broje se u stotinama tisuća, novinskih članaka je jednako toliko. Korisnici te resurse mogu pretraživati, ali kako proizvodi mogu biti vrlo slični i može ih biti vrlo mnogo, korisnici nerijetko imaju problema s pronalaženjem onih resursa koji su njima vrijedni. Sustavi preporuka skupa sadržaja koji su za ciljanog korisnika potencijalno korisni predstavljaju rješenje tog problema.

Postoji više vrsta sustava preporuka sadržaja. Najjednostavniji sustav preporuka je onaj u kojem korisnik na temelju svojih interesa stvara skup sadržaja te takav skup preporučuje drugim korisnicima. Drugi način preporuke sadržaja je stvaranje skupa sadržaja prebrojavanjem. Tako nastaju liste sadržaja kao što su "najboljih 10", "najpopularniji" i slično. Treća vrsta sustava preporuka su sustavi koji su "skrojeni" pojedinim korisnicima. Primjer takvog sustava preporuka je sustav preporuka filmova koji na temelju preferenci pojedinog korisnika stvaraju skup filmova koji su potencijalno zanimljivi tom korisniku.

Primjer koliko sustavi preporuke mogu biti uspješni najbolje pokazuje knjiga *Touching the Void* od Joe Simpson. Ova knjiga o planinarenju nije bila popularna kada je izašla na tržište. Nekoliko godina kasnije Jon Krakauer napisao je knjigu *Into Thin Air* na istu temu. Amazonov sustav preporuke registrirao je nekoliko korisnika koji su kupili obje knjige i počeo je preporučivati knjigu *Touching the Void* korisnicima koji su kupili ili su pregledavali knjigu *Into Thin Air*. U konačnici, knjiga *Touching the Void* postala je popularnija od *Into Thin Air*. Da nije bilo Amazona i njegovog sustava preporuke, korisnici vjerojatno nikad ne bi pregledali ovu knjigu. Ovaj primjer pokazuje koliko su sustavi preporuke bitni te koliko podižu kvalitetu usluge.

1.1. Matrica korisnosti

Prije nego li predstavimo probleme s kojima se suočavamo prilikom oblikovanja sustava preporuka, potrebno je definirati matricu korisnosti (engl. *utility matrix*). Postoje dva subjekta u sustavima preporuka, to su korisnici (engl. *users*) i resursi (engl. *items*). Korisnici preferiraju određene proizvode i potrebno je te preference zapisati kao podatke. Podaci su prikazani u obliku matrice korisnosti koja za svaki par korisnik-resurs opisuje koliko je određeni resurs koristan odnosno cijenjen korisniku.

Problem generiranja preporuka može se formalno zapisati na sljedeći način: neka je $C = (c_1, c_2, ..., c_m)$ skup korisnika sustava, $S = (s_1, s_2, ..., s_n)$ skup svih resursa koji se mogu preporučiti, a *R* potpuno uređeni skup. Neka je $u : C \times S \rightarrow R$ funkcija korisnosti. Vrijednost funkcije $u(c_i, s_j)$ predstavlja korisnost resursa s_j korisniku c_i. Cilj sustava je za svakog korisnika $c_i \in C$ odrediti resurs s_j za koji je vrijednost funkcije u maksimalna:

$$\forall c_i \in C, j = \arg\max_{j_k \in S} u(c_i, s_j) \tag{1.1}$$

Tablica 1.1: Matrica korisnosti

Korisnik / Film	Love at last	Romance forever	Fast and furious
Ana	5	?	1
Branko	?	3	?
Ivana	1	?	5
Matija	?	?	4

Matrica korisnosti je rijetko popunjena matrica, tj. nije poznato kako svaki korisnik cijeni svaki resurs što prikazuje tablica 1.1 u kojoj je prikazano kako su korisnici ocijenili filmove. Oznaka ? predstavlja nepoznatu vrijednost, odnosno korisnik nije ocijenio film gdje se nalazi oznaka ?.

1.2. Ključni problemi

Prilikom oblikovanja sustava rješavanje sljedećih problema je ključno:

1. prikupljanje poznatih podataka o korisnicima i proizvodima
2. predviđanje nepoznatih vrijednosti na temelju poznatih podataka
3. ocjenjivanje metoda preporuka

Prilikom prikupljanja podataka o korisnicima i proizvodima potrebno je definirati način na koji će korisnici unositi koliko cijene pojedini resurs. Postoje dva načina:

- eksplicitno prikupljanje podataka
- implicitno prikupljanje podataka

Prilikom eksplicitnog prikupljanja podataka, jednostavno rješenje je tražiti korisnike da ocijene resurse. Prednost ovog pristupa je ta što su informacije dobivene izravno od korisnika i vrlo je jednostavno za provesti ovaj postupak. Problem je što će samo mali dio korisnika dodijeliti ocjenu nekom resursu, bilo zbog toga što korisnici ne žele ocjenjivati resurse ili zbog toga što je resursa vrlo mnogo odnosno ovakav sustav nije skalabilan.

Ideja implicitnog prikupljanja podataka je naučiti ocjenjivati proizvode na temelju drugih korisničkih akcija. Prednost je ta što ovim načinom korisnici ne ocjenjuju izravno resurse, odnosno ovakav sustav je skalabilan. Ovakvim sustavom relativno je jednostavno dodijeliti nekom resursu visoku ocjenu, ali je vrlo teško dodijeliti nisku ocjenu.

U praksi je uobičajeno koristiti istovremeno eksplicitno i implicitno prikupljanje podataka.

Predviđanje nepoznatih vrijednosti matrice korisnosti temeljni je problem sustava preporuka. Prilikom dodavanja novih resursa u sustav preporuka potrebno je omogućiti da ti resursi budu preporučeni određenim korisnicima. Isto vrijedi i za nove korisnike koji nemaju povijest unutar sustava preporuke. Prethodna dva zahtjeva nazivaju se hladni početak (engl. *cold start*). Problem predviđanja nepoznatih vrijednosti matrice korisnosti ćemo obraditi detaljnije u sljedećim poglavljima.

1.3. Klasifikacija sustava preporuke

Koristeći različite tehnike i algoritme prilikom procjene vrijednosti funkcije korisnosti resursa, sustavi preporuke mogu se podijeliti u tri osnovne grupe:

- sustavi preporuka zasnovani na sadržaju (engl. *content-based recommendations*) - usredotočuje se na značajke resursa i njihova međusobna sličnost određena je mjerenjem sličnosti njihovih značajki
- kolaborativno filtriranje (engl. *collaborative filtering*) - naglasak je na odnosu između korisnika i resursa, sličnost resursa određena je sličnošću korisnosti koje je dodijeli korisnik tim resursima
- mješoviti (hibridni) sustavi preporuke

2. Preporuke zasnovane na sadržaju

Tehnika preporuke zasnovane na sadržaju temelji se na pretpostavci da će se korisniku svidjeti resursi s karakteristikama sličnima onim resursima koji su mu se svidjeli u prošlosti. Sustav na temelju prethodno ocijenjenih ili odabranih sadržaja nastoji naučiti što korisnik preferira.

Prilikom generiranja preporuka sustav uspoređuje karakteristike svih resursa s onim karakteristikama koje je korisnik odabrao. Iz skupa resursa izdvajaju se resursi koji su karakteristikama bliski onima koje je korisnik odabrao kao cijenjene. Na taj se način značajno smanjuje broj resursa koji su potencijalno zanimljivi korisniku. Primjerice, ako je korisnik ocijenio visokom ocjenom filmove određenog žanra, sustav pretpostavlja kako korisnik preferira određeni žanr filmova te će u budućnosti predlagati filmove tog istog žanra.

Za ostvarenje ove tehnike potreban je opis resursa (engl. *item profiles*) i opis preferenci korisnika (engl. *user profiles*) koji su obrađeni u sljedećim potpoglavljima.

2.1. Opis resursa

Tehnika izračunavanja preporuka zasnovanih na sadržaju zahtjeva izgradnju opisa resursa. Opis resursa je skup bitnih značajki koji ga opisuju. U jednostavnom slučaju, opis resursa se sastoji od značajki koje su lako prepoznate. Uobičajena reprezentacija je vektor s vrijednostima *istina* ili *laž* te realnih vrijednosti. Svaka komponenta vektora izražava mjeru važnosti neke karakteristike u opisu.

Tablica 2.1: Opis filmova

Film / Značajke	romantika	akcija
Love at last	0.9	0
Romance forever	1.0	0.01
Fast and furious	0.2	1

Tablica 2.1 prikazuje moguće značajke filmova. U ovom primjeru kao značajke odabrani su *romantika* i *akcija*. Značajke mogu poprimiti vrijednost unutar intervala $[0, 1]$ i one pokazuju u kojoj mjeri je zastupljena romantika i akcija u određenom filmu.

2.2. Opis korisnika

Analogno opisu resursa, potreban je i opis korisnika. Opis korisnika je vektor s istim značajkama kao i u opisu resursa te one opisuju sklonosti, interese i potrebe korisnika. Matrica preferenci je poveznica između resursa i korisnika. Na temelju poznatih vrijednosti matrice gradi se opis svakog korisnika.

Opis korisnika moguće je izračunati na više načina. Jedan od mogućih načina je koristeći linearnu regresiju. Neka je $r(i, j) = 1$ ako je korisnik j ocijenio resurs i, inače je vrijednost $r(i, j) = 0$. Ocjena koju je dodijelio korisnik j resursu i označimo s $y^{(i,j)}$. Neka je θ^j vektor koji opisuje korisnika j, a x^i opis resursa i. Za korisnika j i resurs i predviđenu vrijednost računamo kao $(\theta^j)^T(x^i)$. Tada učimo parametar θ^j na sljedeći način:

$$\min_{\theta^j} \frac{1}{2} \sum_{i:r(i,j)=1} ((\theta^j)^T(x^i) - y^{(i,j)})^2 \tag{2.1}$$

2.3. Generiranje preporuka

Tehnike određivanja vrijednosti funkcije korisnosti klasificiraju se u heurističke i tehnike zasnovane na modelu. Kod heurističkih tehnika predviđanje se određuje pomoću heurističkih formula temeljenih na metodama za pretraživanje informacija. Primjeri takvih formula su euklidska udaljenost i kosinusova sličnost. Kod tehnika zasnovanih na modelu predviđanje se temelji na modelu naučenom na podacima koristeći strojno učenje i statičke metode poput algoritama za klasteriranje, stabala odluke, Bayesovih i neuronskih mreža.

2.4. Prednosti i nedostaci

Kao i svaka tehnika tako i tehnika preporuka zasnovanih na sadržaju sa sobom donosi prednosti i nedostatke.

Prilikom generiranja preporuka nisu potrebni podaci svih ostalih korisnika da bi se napravila predviđanja za nekog određenog korisnika. Također, omogućeno je vrlo

dobro predviđanje koji resursi bi mogli biti cijenjeni kod korisnika s jedinstvenim preferencama. Omogućeno je preporučivanje novih i nepopularnih proizvoda jer opis resursa ovisi samo o resursu, a ne o korisnicima što djelomično rješava problem *cold start*. Predviđanje zasnovano na sadržaju lako je interpretirati. To su prednosti preporuka zasnovanih na sadržaju.

Predviđanje zasnovano na sadržaju dovodi do prejakog profiliranja, odnosno nikada se neće preporučiti određeni resurs korisniku koji nema opis sličan opisu resursa. Ponekad je teško izgraditi opis resursa jer može biti teško pronaći bitne značajke koje ga opisuju. Također, novi korisnici nemaju izgrađen korisnički opis te preporuke za te korisnike ne moraju odgovarati njihovim stvarnim preferencama.

3. Kolaborativno filtriranje

Sustavi za kolaborativno filtriranje umjesto da koriste značajke resursa kako bi odredili sličnosti, koriste sličnost korisničkih ocjena za neka dva resursa. Odnosno, predviđanje korisnosti pojedinog resursa za nekog korisnika generira se na temelju vrednovanja resursa od strane ostalih korisnika.

Kod sustava za kolaborativno filtriranje ocjene koje su korisnici dodijelili resursima koriste se kao aproksimativna reprezentacija njihovih interesa i potreba. Dakle, umjesto vektora opisa korisnika koristi se redak u matrici korisnosti. Za razliku od preporuka zasnovanih na sadržaju, modeli ne sadrže podatke o resursima već se ocjene dodijeljene od strane određenog korisnika uspoređuju s ocjenama koje su dodijelili ostali korisnici. Odnosno, umjesto vektora značajki resursa koristi se stupac matrice korisnosti za film koji je njime predstavljen.

Korisnici su slični ako je udaljenost između njihovih vektora mala za pogodno definiranu mjeru udaljenosti. Preporuke za korisnika c_j generirane su tako da se pronađu njemu najsličniji korisnici i tada se preporučuju resursi koji su cijenjeni tim najsličnijim korisnicima.

Matricu korisnosti možemo promatrati na dva načina, tako da unutar nje tražimo slične korisnike, odnosno slične resurse. Na temelju toga, tehnike kolaborativnog filtriranja klasificiraju se na sljedeća dva načina:

1. Preporuke generirane na temelju sličnosti korisnika
2. Preporuke generirane na temelju sličnosti resursa

3.1. Mjere sličnosti

Bitna komponenta prilikom implementiranja sustava kolaborativnog filtriranja je odabir funkcije sličnosti. Funkcija sličnosti upotrebljava se prilikom računanja sličnosti korisnika te sličnosti resursa. U ovom poglavlju predloženo je nekoliko mjera sličnosti.

3.1.1. Jaccardova udaljenost

Jaccardova udaljenost vrijednosti matrice gleda kao skupove. Ova mjera udaljenosti pogodna je kada matrica korisnosti sadrži vrijednosti *istina* i *laž*. Na primjer, Jaccardova udaljenost se može koristiti ako matrica korisnosti sadrži informaciju o tome je li korisnik kupio određeni resurs.

$$Jaccard\_udaljenost(S_1, S_2) = 1 - \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \tag{3.1}$$

Tablica 3.1: Matrica korisnosti

Korisnik / Resurs	R1	R2	R3	R4
C1	1	1		1
C2		1	1	1

Na temelju tablice 3.1 Jaccardova udaljenost jednaka je: $Jaccard\_udaljenost(C_1, C_2) = \frac{1}{2}$

3.1.2. Kosinusova udaljenost

Prilikom računanja kosinusove udaljenosti nepostojeće vrijednosti možemo tretirati kao da su jednaki 0. Kosinusovu udaljenost računa se kao:

$$s(u, v) = \frac{r_u \cdot r_v}{||r_u|| \cdot ||r_v||} \tag{3.2}$$

Kosinusova udaljenost između *Ivana* i *Matija* na temelju primjera 1.1 jednaka je $s(Ivana, Matija) = \frac{5 \cdot 4}{\sqrt{1^2 + 5^2} \cdot \sqrt{4^2}} = 0.981$

3.2. Preporuke generirane na temelju sličnosti korisnika

Preporuke generirane na temelju sličnosti korisnika imaju osnovnu pretpostavku da korisnici koji su u prošlosti ocijenili sličnim ocjenama resurse imaju jednake preference. Dakle, prilikom izračunavanja preporuka za korisnika *x*, preporučuju se resursi koje su slični korisnici ocijenili najvišim ocjenama. Na taj način se postiže da se korisniku *x* preporuče sadržaji koje preferiraju njemu slični korisnici.

3.3. Preporuke generirane na temelju sličnosti resursa

Kod preporuka temeljenih na filtriranju po sličnosti resursa, za resurse koje je ciljni korisnik pregledavao ili ocijenio nalaze se i preporučuju slični resursi. Prilikom određivanja sličnosti među sadržajima uspoređuje se vrednovanje ciljnog korisnika s vrednovanjem ostalih korisnika. Dakle, sličnost resursa se određuje na temelju ocjena koje su im korisnici dodijelili.

Preporuke temeljene na sličnosti resursa u praksi obično daju bolje rezultate od preporuka generiranih na temelju sličnosti korisnika. Odnose između resursa je uobičajeno lakše pronaći nego između korisnika pa je time i resurse lakše klasificirati. Na primjer, pjesme obično pripadaju jednom žanru. Nemoguće je da jedna pjesma istovremeno pripada rock glazbi 70-ih godina i baroknoj glazbi 1700-ih godina, ali jedan korisnik može kupiti album oba žanra. Dakle, lakše je prepoznati sličnosti između pjesama jer one pripadaju jednom žanru nego prepoznati da su dva korisnika slična jer preferiraju isti žanr, a ujedno mogu potpuno ignorirati drugi žanr.

3.4. Prednosti i nedostaci

Kolaborativno preporučivanje sadržaja ograničeno je na one resurse koje su ocijenili ostali korisnici. Preporuke su neovisne o sadržaju jer se prilikom generiranja preporuka koriste ocjene ostalih korisnika, što nije slučaj kod preporuka generiranih na temelju sadržaja. Takav način rada omogućuje da se korisniku preporuče resursi koji su različiti od onoga što je korisnik pregledavao u prošlosti.

Nedostaci kolaborativnog filtriranja vidljivi su kada je mali broj resursa ocijenjen (engl. *sparse rating problem*). Ovaj problem je pogotovo prisutan kod sustava s velikim brojem resursa koje je moguće preporučiti. Resursi koji su vrednovani od strane malog broja korisnika gotovo nikad neće biti preporučeni, neovisno o visini ocjena. Također, ukoliko se skup sadržaja za preporuke često mijenja, ranije dodijeljene ocjene neće koristiti novim korisnicima. Kao i kod generiranja preporuka temeljenih na sadržaju, tako i kod kolaborativnog filtriranja javlja se problem *cold start*. Kako na početku rada sustava nema ocijenjenih sadržaja od strane korisnika, nije moguće generirati odgovarajuće preporuke. Isti se problem javlja i u slučaju novog sadržaja u sustavu. Sve dok novi sadržaj ne ocijeni dovoljan broj korisnika, sustav ga neće preporučivati.

4. Mješoviti sustavi preporuke

Mješoviti sustavi preporuke istovremeno koriste preporuke zasnovane na sadržaju i kolaborativno filtriranje. Empirijski se pokazalo da je ponekad spajanjem ova dva pristupa predikcije moguće uspješnije generirati nego li korištenjem samo jednog od njih. Pristupi implementaciji mogu se klasificirati ovisno o načinu kombiniranja različitih tehnika za generiranje preporuka. Moguće implementacije su: implementiranje zasebno sustava preporuka temeljenih na sadržaju i kolaborativnog filtriranja te ih nakon toga spojiti u jedan sustav ili dodavanjem preporuka temeljenih na sadržaju u kolaborativno filtriranje (i obrnuto). Također, ovaj pristup može biti pogodan za rješavanje nekih problema kao što su *cold start* i problem rijetkosti matrice korisnosti (engl. *sparsity problem*).

Netflix je jedan od primjera gdje se koristi hibridni sustav preporuka. Preporuke su generirane uspoređivanjem što korisnici gledaju i njihovim navikama prilikom pretraživanja sličnih korisnika (kolaborativno filtriranje) i nuđenjem filmova koji imaju slične karakteristike s filmovima koje su korisnici ocijenili visokim ocjenama (preporuke generirane na temelju sadržaja).

5. Zaključak

Sustavi preporuka su posrednici između korisnika i resursa. Njihov odnos zapisan je u matrici korisnosti. To je struktura podataka koja sadrži informacije koliko korisnici cijene određene resurse. Većina vrijednosti matrice je nepoznata i osnovni problem sustava preporuka je predviđanje tih nepoznatih vrijednosti na temelju poznatih.

Sustavi preporuka mogu se klasificirati u tri osnovne skupine. Oni pokušavaju predvidjeti koliko korisnik cijeni određeni resurs pronalaženjem sličnih resursa koje je korisnik već ocijenio. Sustavi preporuka zasnovanih na sadržaju mjere sličnost resursa uspoređivanjem značajki resursa te oni predstavljaju prvu skupinu sustava preporuka. Prilikom usporedbe koriste se funkcije sličnosti koje govore koliko su dva korisnika, odnosno dva resursa bliska. Drugu skupinu sustava preporuka čine sustavi za kolaborativno filtriranje. Takvi sustavi mjere sličnosti korisnika na temelju njihovih preferenci. Sličnosti resursa određuju se na temelju ocjena koje su im korisnici dodijelili. Treća grupa sustava preporuka su mješoviti sustavi preporuke i oni istovremeno koristi prve dvije skupine.

Sustavi preporuka koriste se vrlo često zbog velikog broja sadržaja dostupnog korisnicima. Iako imaju relativno mnogo mana, prilikom predviđanja sadržaja potencijalno zanimljivih korisniku ostvaruju dobre rezultate. Upravo zbog toga, sustavi preporuka povećavaju kvalitetu usluga dostupnih na internetu.

6. Literatura

Martina Holenko Dlab. Primjena sustava za preporuke kod hipermedijske programske potpore za učenje. *Odjel za informatiku Sveučilišta u Rijeci*. URL `https://www.fer.unizg.hr/_download/repository/Holenko_Dlab_kvalifikacijski.pdf`.

J. Leskovec, A. Rajaraman, i J.D. Ullman. *Mining of Massive Datasets*. Cambridge University Press, 2014. ISBN 9781107077232. URL `https://books.google.hr/books?id=1l-WoAEACAAJ`.

Prem Melville i Vikas Sindhwani. Recommender systems.

Michael J. Pazzani i Daniel Billsus. Content-based recommendation systems. U *THE ADAPTIVE WEB: METHODS AND STRATEGIES OF WEB PERSONALIZATION. VOLUME 4321 OF LECTURE NOTES IN COMPUTER SCIENCE*, stranice 325–341. Springer-Verlag, 2007.

Pregled pristupa izrade sustava preporuke

Sažetak

Količina sadržaja dostupnih na internetu svakim danom raste što rezultira problemom preopterećenja informacijama. Korisnici zbog preopterećenja informacijama imaju poteškoća pri pronalaženju relevantnih sadržaja. Sustavi preporuka pokušavaju riješiti taj problem vršeći automatizirano izdvajanje podskupa sadržaja koji je potencijalno zanimljiv ciljnim korisnicima. U radu je navedena klasifikacija sustava za preporuke, svaka skupina je razrađena te su prikazani problemi koji se javljaju prilikom primjene pojedinih tehnika za generiranje preporuka.

Ključne riječi: sustavi preporuke, matrica korisnosti, sličnosti korisnika, sličnosti resursa