UNIVERZA V LJUBLJANI
FAKULTETA ZA RAČUNALNIŠTVO IN INFORMATIKO

Blaž Bahar

Primerjava različnih tipov priporočilnih sistemov

DIPLOMSKO DELO

VISOKOŠOLSKI STROKOVNI ŠTUDIJSKI PROGRAM PRVE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

Mentor: viš. pred. dr. Aleksander Sadikov

Ljubljana, 2012

Univerza *v Ljubljani*
Fakulteta *za računalništvo in informatiko*

Tržaška 25
1001 Ljubljana, Slovenija
telefon: 01 476 84 11
01 476 83 87
faks: 01 426 46 47
01 476 87 11
www.fri.uni-lj.si
e-mail: dekanat@fri.uni-lj.si

Št. naloge: 00304/2012
Datum: 13.04.2012

Univerza v Ljubljani, Fakulteta za računalništvo in informatiko izdaja naslednjo nalogo:

Kandidat: **BLAŽ BAHAR**

Naslov: **PRIMERJAVA RAZLIČNIH TIPOV PRIPOROČILNIH SISTEMOV**

A COMPARISON OF DIFFERENT TYPES OF RECOMMENDER SYSTEMS

Vrsta naloge: Diplomsko delo visokošolskega strokovnega študija prve stopnje

Tematika naloge:

Kandidat naj implementira oz. prilagodi tri različne metode priporočilnih sistemov; enega na podlagi sodelovanja uporabnikov, drugega na podlagi vsebine in tretjega, za osnovno primerjavo, s preprosto uporabo prilagojenih povprečij. Nadalje naj kandidat vse tri sisteme med seboj primerja na realni podatkovni bazi z ocenami uporabnikov. Podatke o vsebini produktov naj pridobi preko prostodostopne storitve "freebase". Kandidat naj poskusi določiti v kakšnih primerih kateri izmed sistemov deluje bolje od ostalih.

Mentor:

viš. pred. dr. Aleksander Sadikov

Dekan:

prof. dr. Nikolaj Zimic

Fakulteta za računalništvo in informatiko • UNIVERZA V LJUBLJANI • 2

IZJAVA O AVTORSTVU
diplomskega dela

Spodaj podpisani Blaž Bahar

z vpisno številko 63080018

sem avtor diplomskega dela z naslovom:

Primerjava različnih tipov priporočilnih sistemov

S svojim podpisom zagotavljam, da:

- sem diplomsko delo izdelal samostojno pod mentorstvom: viš. pred. dr. Aleksander Sadikov
- so elektronska oblika diplomskega dela, naslov (slov., angl.), povzetek (slov., angl.) ter ključne besede (slov., ang.) identični s tiskano obliko diplomskega dela
- soglašam z javno objavo elektronske oblike diplomskega dela v zbirki »Dela FRI«

V Ljubljani, dne 24. 9. 2012 Podpis avtorja:

Zahvala

Zahvaljujem se mentorju na fakulteti za računalništvo in informatiko v Ljubljani viš. pred. dr. Aleksandru Sadikovu za vodenje, svetovanje, nudenje pomoči, posvečanje svojega časa pri pisanju te diplomske naloge. Hvala tudi za skripto colrec, s katero sem si veliko pomagal pri analizi metode izbiranja s sodelovanjem na osnovi uporabnikov.

Zahvala gre tudi asistentu na fakulteti za računalništvo in informatiko v Ljubljani, dr. Martinu Možini, ki mi je pomagal predelati skripto colrec.

Zahvaljujem se staršema, ki sta me podpirali pri študiju in pisanju te diplomske naloge, ter mi pomagata v vsakdanjem življenju. Hvala tudi vsem ostalim sorodnikom in prijateljem.

Kazalo

Kazalo slik

Kazalo tabel

Povzetek

V diplomski nalogi se je primerjalo tri različne tipe priporočilnih sistemov: osnovna metoda, metoda izbiranja s sodelovanjem na osnovi uporabnikov in vsebinsko osnovana metoda.
Spoznavalo se je, kaj priporočilni sistemu sploh so, čemu so namenjeni. Predstavljene so bile vse tri obravnavane metode in še nekatere druge. Spoznavalo se je prednosti in slabosti metode izbiranja s sodelovanjem in vsebinsko osnovane metode. Obravnavano je bilo ocenjevanje priporočilnih sistemov.
Podrobneje se je spoznavalo metodo izbiranja s sodelovanjem. Predstavljeno je bilo priporočanje na osnovi uporabnikov z metodo najbližjih sosedov. Predstavljen je bil pojem soseščine. Predstavljena sta bila algoritma za računanje podobnosti med uporabniki in napovedovanje ocen.
Podrobneje se je spoznavalo vsebinsko osnovano metodo. Predstavljena je bila arhitektura tovrstnih sistemov. Predstavljeni sta bili funkciji za računanje podobnosti med produkti
Podrobneje se je obravnavalo ocenjevanje priporočilnih sistemov. Predstavljena je bila metoda izloči enega in meri za računanje napake pri priporočanju MAE in RMSE.
Podatki so bili pridobljeni iz prosto dostopne storitve "Freebase". Spremenjene so bile vrednosti atributov, izločeni so bili nekateri atributi
Skripta colrec je bila prilagojena in uporabljena za analizo metode izbiranja s sodelovanjem na osnovi uporabnikov. Priporočanje metode se je optimiziralo z določanjem praga podobnosti.
Vsebinsko osnovana metoda je bila optimizirana z uporabo različnih funkcij za izračun podobnosti med produkti. Vsak atribut je bil utežen z določeno utežjo. Predstavljeni sta bila algoritma za računanje podobnosti med produkti in napovedovanje ocen.
Vsebinsko osnovana metoda in osnovna metoda sta bili implementirani v programskem jeziku Python. Za shranjevanje ocen je bila uporabljena dvodimenzionalna tabela, ki je del knjižnice numpy.

Ključne besede: priporočilni sistemi, osnovna metoda, metoda izbiranja s sodelovanjem, vsebinsko osnovana metoda, kNN, Jaccard, Ochiai, podobnost, napovedovanje ocene, optimizacija, ocenjevanje priporočilnih sistemov, MAE, RMSE, Freebase

Abstract

In this thesis three different types of reccommender systems were compared: baseline predictor, collaborative filtering, content-based recommeder.
We looked at what recommender systems are and what they are good for. All three methods were addressed and also some others. Pros anc cons of collaborative filtering and content-based recommenders were addressed. The evaluation of recommender systems was addressed.
We took a closer look at collaborative filtering. User-based kNN recommendation was introduced. The term neigbourhood was introduced. The alghoritms for similarity calculation and predicted rating calculation were introduced.
We took a closer look at content based recommender. A high level architecture of content-based systems was introduced. Two functions for product similarity calculation were introduced.
We took a closer look at evaluation of the recommender systems. The leave-one-out method was introduced and also two measures for error calculation MAE and RMSE.
The data was acquired from "Freebase" service. The values of attributes were changed and some of the attributes were eliminated.
Colrec script was adapted and used for analyzing the user-based collaborative filtering. The optimization of recommendation technique was performed by using the threshold.
Content-based technique was optimized by using different finctions for calculating the similarity between products. Every attribute was weighted with a different weight. The alghorithms for calculating the similarity between products and predicting ratings were introduced.
Content-based technique and baseline predictor were implemented by using Python programming language. For ratings storage the two-dimensional array was used, which is included in numpy library.

Keywords: recommender systems, baseline predictor, collaborative filtering, content-based technique, kNN, Jaccard, Ochiai, similarity, rating prediction, optimization, evaluating recommender systems, MAE, RMSE, Freebase

1. Priporočilni sistemi

1.1 Uvod

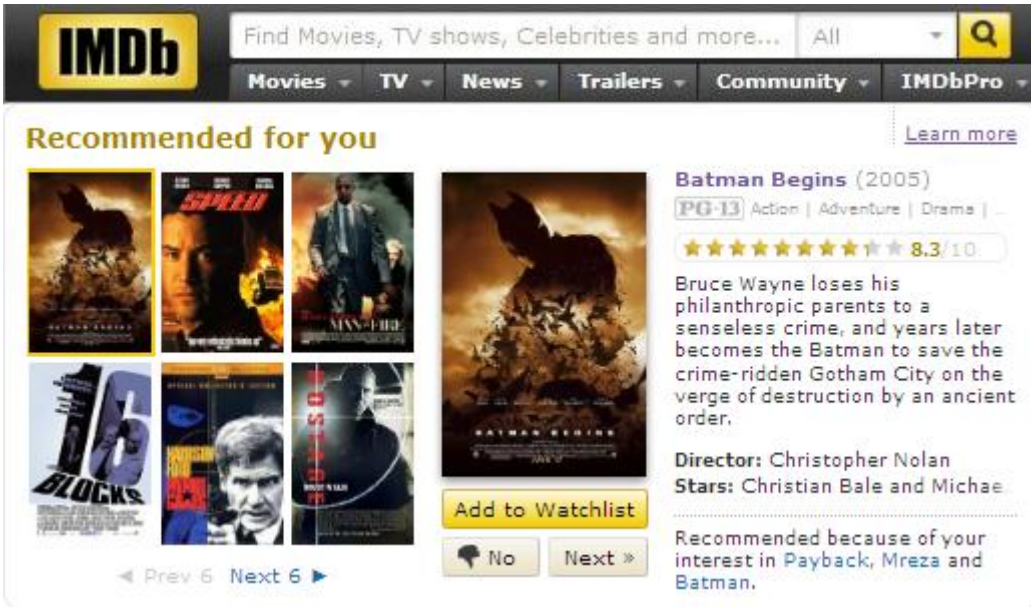
V današnjem času smo ljudje obremenjeni z (pre)veliko količino informacij. S tem problemom se spopadamo na različne načine: se posvetujemo s sorodniki, prijatelji, strokovnjaki ali pa si pomagamo s pomočjo iskanja po spletu. Ker so lahko nasveti drugih oseb neuporabni oz. ker lahko porabimo preveč časa za iskanje primerne informacije na spletu, ali pa nas različne informacije na spletu prej zmedejo kot pa (hitro) privedejo do koristnih informacij , nam pride prav sistem, ki nam pomaga med množico informacij poiskati tiste, ki so za nas najbolj zanimive/koristne.

Vsi smo si že kdaj postavili naslednje vprašanje: Kateri film naj grem gledati v kino ? V kinu je na sporedu več filmov in želimo izbrati tistega, ki nam bo najbolj všeč.

Priporočilni sistem [2, 5] se s pomočjo uporabnikove zgodovine akcij (ogledov, podajanja ocen) priuči profila uporabnika – kaj je uporabniku všeč oz. kaj mu ni všeč. Na osnovi tega profila sistem uporabniku podaja priporočila. Rezultati priporočilnega sistema so lahko učinkoviteje prilagojeni uporabnikovim preferencam, če uporabnik dlje časa uporablja isti priporočilni sistem, saj se z vsako interakcijo izboljšuje model uporabnika.

Oglejmo si primer preprostega priporočilnega sistema.
Spletna stran http://www.imdb.com/ vsebuje podatke o filmih. Uporabnik v polje "Iskanje" vnese naslov filma, ki ga zanima. Prikažejo se osnovni podatki o filmu. Poleg tega se pojavi tudi polje "Uporabniki, ki jim je bil ta film všeč, so si ogledali tudi …", ki vsebuje naslove filmov, ki bi lahko bili za uporabnika zanimivi.
Ko si dalj časa ogledujemo filme na tej spletni strani, se sistem priuči našega okusa, in nam že ob vstopu na spletno stran priporoča filme, ki bi nas utegnili zanimati (polje "Priporočamo Vam").
Opisani elementi so prikazani na sliki 1 in na sliki 2.

Slika 1: Rezultat iskanja filma in področje "Uporabniki, ki jim je bil ta film všeč, so si ogledali tudi …". [8]



Slika 2: Primer priporočilnega sistema (polje "Priporočamo Vam"). V spodnjem desnem kotu je obrazložitev priporočil (na podlagi preteklih uporabnikovih ogledov). [8]

Po uporabnikovi poizvedbi, ki je lahko razčlenjena glede na metodo priporočanja, uporabnikove lastnosti in potrebe, priporočilni sistemi ustvarijo priporočila z uporabo različnih znanj in podatkov o uporabniku, produktih in preteklih transakcijah (zabeleženih interakcijah med uporabnikom in priporočilnim sistemom), shranjenih v prilagojenih podatkovnih bazah. Uporabnik lahko potem brska po priporočilih in zanje podaja svojo oceno (odgovor). Uporabnikova dejanja in odgovore se shrani v podatkovne baze za priporočanje in se jih uporabi za ustvarjanje novih priporočil v naslednjih interakcijah med uporabnikom in sistemom.

Temeljna naloga priporočilnih sistemov je zagotoviti lahko dosegljiva, cenovno ugodna, osebna in visoko kakovostna priporočila velikemu številu uporabnikov.

Priporočila so lahko prilagojena ali neprilagojena. Prilagojena priporočila so tista, pri katerih sistem upošteva uporabnikove preference ali lastnosti. Tako različni uporabniki prejemajo različne predloge.
Neprilagojena priporočila je lažje realizirati, ponavadi jih najdemo v revijah (npr. 10 najbolj gledanih filmov).
Raziskovanje priporočilnih sistemov se osredotoča na prilagojena priporočila, saj obstajajo uporabniki, ki jim neprilagojena priporočila ne koristijo (npr. uporabniku nobeden od 10 najbolj gledanih filmov ni všeč).

Raziskovanje priporočilnih sistemov je sorazmerno novo področje v primerjavi z raziskovanjem ostalih klasičnih informacijskih sistemskih orodij in tehnologij (npr. podatkovne baze, brskalniki). Preprosta in privlačna ideja priporočilnih sistemov se je pojavila v začetku 90. let – zbrati mnenja milijonov uporabnikov na spletu z namenom nudenja pomoči pri iskanju uporabnih in zanimivih informacij.
V zadnjih letih se je zanimanje za priporočilne sisteme močno povečalo, kar kažejo naslednji primeri:

1. Priporočilni sistemi imajo pomembno vlogo na splošno znanih internetnih straneh, kot so Amazon.com, YouTube, Netflix, Yahoo, Tripadvisor, Last.fm in IMDb.
2. Organizirane so konference in delavnice, ki obravnavajo priporočilne sisteme (npr. ACM Recommender Systems – Recsys).
3. Na visokošolskih institucijah po svetu se na dodiplomskem in podiplomskem študijskem programu obravnava priporočilne sisteme.

V tabeli 1 so našteta najbolj pogosta področja, kjer se uporablja priporočilne sisteme. [5]

Področje uporabe priporočilnih sistemov	**Primeri uporabe priporočilnih sistemov**
Zabava	Priporočanje filmov, glasbe
Vsebina	Priporočanje dokumentov, spletnih strani, aplikacij za spletno učenje, filtrov za elektronsko pošto
Spletno oglaševanje	Priporočanje nakupa produkta, npr. knjige, fotoaparata, računalnika
Storitve	Priporočanje potovalnih storitev, strokovnjakov za pogovor, najema stanovanja

Tabela 1: Najbolj pogosta področja, kjer se uporablja priporočilne sisteme.

1.2 Metoda izbiranja s sodelovanjem in vsebinsko osnovana metoda

Poznamo priporočilne metode, ki ne potrebujejo veliko znanja oz. uporabljajo zelo preproste in osnovne podatke, kot so npr. uporabnikove ocene produktov. Ostale metode so veliko bolj odvisne od znanja, npr. uporabljajo podatke o uporabnikih ali produktih, omejitvah, družbenih povezavah ali aktivnostih uporabnikov.

Metoda izbiranja s sodelovanjem (collaborative recommendation technique)

Osnova ideja priporočilnih sistemov, ki uporablja metodo izbiranja s sodelovanjem [2], je naslednja: če so imeli uporabniki v preteklosti podoben okus (npr. ogledali so si enake filme), bodo imeli podoben okus tudi v prihodnosti. Ker izbira produktov, za katere upamo, da bodo uporabnika zanimale, vključuje izbiranje najbolj verjetnih produktov iz velike množice produktov, in ker uporabniki med seboj na nek način "sodelujejo", to metodo imenujemo izbiranje s sodelovanjem (collaborative filtering).

Vsebinsko osnovana metoda (content-based recommendation technique)

Vsebinsko osnovani priporočilni sistemi [2, 5] so osnovani na dostopnosti opisov produktov in profila uporabnika, ki določi pomembnost lastnostim produkta. Tovrstni sistem se nauči priporočati produkte, ki so bili uporabniku všeč v preteklosti. Podobnost med produkti se izračuna na osnovi lastnosti primerjanih produktov.

Poleg teh dveh metod poznamo še naslednje štiri metode:

- Metoda, osnovana na znanju (knowledge-based recommendation technique) [2] – Na področju, kot je npr. zabavna elektronika, je prisotno veliko število kupcev, ki opravijo le enkraten nakup. Iz tega sledi, da ne moremo uporabiti zgodovine uporabnika (ker ni na voljo), ki pa je predpogoj za pristope izbiranja s sodelovanjem in vsebinsko osnovanih sistemov. Lahko pa je dostopna bolj podrobna in bolje organizirana vsebina, vključno s tehničnimi lastnostmi, podatki o kakovosti, znanjem o uporabnikih in produktih za priporočanje, znanjem o področju.
- Demografska metoda (Demographic recommendation technique) [5] - Priporočilni sistem, osnovan na demografski metodi, priporoča produkte na osnovi demografskega profila uporabnika. Glavna predpostavka je, da morajo biti za različne demografske niše podana različna priporočila.
- Družbeno osnovana metoda (Community-based recommendation technique) [5] - Priporočilni sistem, ki uporablja družbeno osnovano metodo, priporoča produkte z

uporabo preferenc/ocen uporabnikovih prijateljev. Zbira in ureja podatke o povezavah med uporabniki in o preferencah uporabnikovih prijateljev. Dokazano je, da ljudje bolj zaupajo priporočilom svojih prijateljev, kot pa priporočilom sebi podobnih, a neznanih posameznikov. Metoda upošteva staro reklo: "Povej mi, kdo so tvoji prijatelji, in povedal ti bom, kdo si".

- Hibridne metode (Hybrid recommender) [2, 5] - Ti priporočilni sistemi so osnovani na združitvi različnih metod. Hibridni sistem z združitvijo metod A in B poizkuša izkoristiti prednosti metode A, da odpravi šibkosti metode B. S tem lahko dosežemo boljša/bolj točna priporočila.

1.2.1 Primerjava metode izbiranja s sodelovanjem in vsebinsko osnovane metode

Obe metodi imata svoje prednosti in slabosti. Povzemamo jih v tabeli 2 (povzeto po [1, 2, 3, 5]).

Metoda	**Prednosti**	**Slabosti**
Metoda izbiranja s sodelovanjem	Odkrivanje niš. Ni potrebno znanje o področju. Prilagodljiva: kakovost priporočil se s časom izboljšuje. Zadošča implicitni odgovor. Presenečenja v priporočilih. Nauči se marketinškega segmenta. Lahko razumljiva. Lahko priporoča produkte z različno vsebino.	Pomanjkanje informacij (hladni zagon) za nove uporabnike. Pomanjkanje informacij (hladni zagon) za nove produkte. Problem "črne ovce" (uporabniki, ki niso podobni nobeni skupini uporabnikov). Kakovost odvisna od velike zbirke zgodovinskih podatkov. Potrebuje skupino uporabnikov. Potrebuje neko obliko odgovora z oceno. Problem stabilnosti proti nestabilnosti. Problemi majhnega števila ocen.

		Ne uporablja drugih virov znanja. Ni razlag rezultatov. Priporočanje preveč podobnih produktov.
Vsebinsko osnovana metoda	Ni potrebno znanje o področju. Neodvisna od (ostalih) uporabnikov. Prilagodljiva: kakovost priporočil se s časom izboljšuje. Zadošča implicitni odgovor. Mogoče primerjanje produktov. Razlaga priporočil. Priporočanje novih produktov. Ne potrebuje veliko uporabnikov za dosego solidne točnosti priporočanja.	Pomanjkanje informacij (hladni zagon) za nove uporabnike. Kakovost odvisna od velike zbirke zgodovinskih podatkov. Problem stabilnosti proti nestabilnosti. Potreben opis vsebine/produkta. Ni presenečenj. Priporočanje preveč podobnih produktov. Površinska analiza vsebine.

Tabela 2: Primerjava metod za priporočanje.

Pristopi, ki uporabljajo metodo izbiranja s sodelovanjem, ne izkoriščajo/potrebujejo znanja o produktih samih. Očitna prednost te strategije je, da teh podatkov ni potrebno vnašati v sistem ali vzdrževati. Po drugi strani je uporaba teh lastnosti za priporočanje produktov, ki so bili uporabniku v preteklosti všeč, lahko bolj učinkovito.

Metoda izbiranja s sodelovanjem je odvisna od tega, da čimveč uporabnikov oceni čimveč enakih produktov in ima težave, kadar malo uporabnikov oceni enake produkte.

Metoda izbiranja s sodelovanjem je najbolj primerna v primeru, ko veliko število uporabnikov oceni veliko produktov iz majhne in nespremenljive množice.

Metoda izbiranja s sodelovanjem najbolje deluje pri uporabniku, ki je podoben velikemu številu ostalih uporabnikov.

Vsebinsko osnovane metode imajo prav tako težave pri začetnem priporočanju, saj morajo najprej zbrati dovolj ocen za zanesljivo priporočanje.

1.3 Ocenjevanje priporočilnih sistemov

Pogosta vprašanja:

- Katere vrste raziskovanja so primerne za ocenjevanje priporočilnih sistemov ?
- Kako so lahko priporočilni sistemi ocenjeni z preizkušanjem sistema na v preteklosti zbranih podatkih ?
- Katera mera je primerna za ocenjevanje priporočilnih sistemov ?
- Kakšne so omejitve obstoječih metod za ocenjevanje, kadar je potrebno upoštevati pogovorno ali poslovno vrednost priporočilnih sistemov ?
- Kako lahko dejansko izmerimo kakovost priporočil priporočilnih sistemov ?

Ocenjevanje [2, 5] je potrebno za preverjanje primernosti izbire metode za priporočanje. Zanima nas, kakšno napako naredi metoda pri napovedovanju ocene. To preverimo tako, da primerjamo dejansko oceno in napovedano oceno.

Ocenjevanje ne obsega samo računanja napake pri priporočanju. Oceniti je potrebno tudi izvedbo priporočilnih sistemov. Preveriti je potrebno parametre, kot so uteženi pragovi, število sosedov, ki zahtevajo stalno prilagajanje in preračunavanje.

Nazadnje se oceni priporočilni sistem še iz vidika uporabnika. Izvedemo nadzorovan poskus, pri katerem uporabniki rešujejo različne naloge z različnimi različicami sistema. Nato je možno analizirati uporabnikove izpolnjene naloge in s pomočjo razdeljenih vprašalnikov ustvariti poročilo o uporabnikovi izkušnji. Pri tovrstnih poskusih se lahko zbira kvalitativne in kvantitativne informacije o sistemu.

2. Priporočilni sistem na osnovi metode izbiranja s sodelovanjem

2.1 Priporočanje na osnovi uporabnikov z metodo najbližjih sosedov

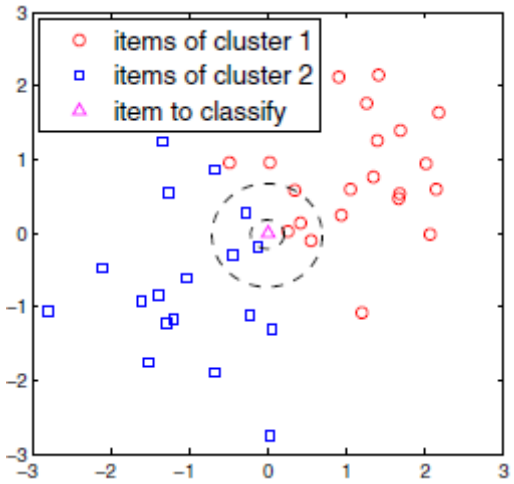
Med pristopi priporočanja metod izbiranja s sodelovanjem, so najbolj priljubljeni tisti, ki so osnovani na metodi izbiranja najbližjih sosedov [5], zaradi njihove preprostosti, učinkovitosti in sposobnosti podajanja točnih in osebnih priporočil. Pri tej metodi so uporabnikove ocene produktov shranjene v sistemu in neposredno uporabljene za napovedovanje ocen novih produktov.

Metoda poišče podobne uporabnike uporabniku, za katerega podajamo priporočilo (aktivni uporabnik), glede na zgodovino podanih ocen. Napovedano oceno za še nepreizkušeni produkt se izračuna z uporabo ocen, ki so jih za ta produkt podali aktivnemu uporabniku podobni uporabniki.

Poleg sistemov, osnovanih na uporabnikih, poznamo tudi sisteme, osnovane na produktih, ki prav tako uporabljajo metodo najbližjih sosedov. Ti sistemi aktivnemu uporabniku napovedujejo oceno za produkt na osnovi ocen, ki jih je ta uporabnik podal temu produktu podobnim produktom. Produkta sta si pri tovrstnih sistemih podobna, če je več uporabnikov sistema podobno ocenilo ta dva produkta.

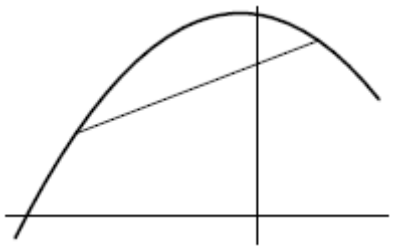
2.1.1 Izbira velikosti soseščine

Točnost napovedovanja je odvisna od velikosti soseščine [2,5]. Z izbiro premajhne velikosti soseščine se lahko zgodi, da ni mogoče podati napovedi, ker v podatkih nimamo med seboj zelo podobnih uporabnikov. Z izbiro prevelike soseščine pa upoštevamo uporabnike, ki se med seboj zelo razlikujejo, kar vodi k večji napaki pri priporočanju in k negativnemu vplivu na čas izvajanja. Iz teh dveh razlogov ne izberemo takšne soseščine, ki bi zajemala vse uporabnike. Primer izbire soseščine prikazuje slika 3.



Slika 3: Izbor velikosti soseščine.

Z večanjem velikosti soseščine se točnost napovedovanja obnaša kot konkavna funkcija [6] (Slika 4).



Slika 4: Konkavna funkcija.

Velikost soseščine se lahko omeji na dva načina: z izbiro praga podobnosti ali pa z upoštevanjem k najbližjih sosedov.

Najbolj primeren prag ali vrednost za k se določi s prečnim preverjanjem.

2.1.2 Računanje podobnosti med uporabniki

Računanje podobnosti ima dvojno vlogo v priporočilnih metodah, osnovanih na soseščini:

1. Dopušča izbiro zaupanja vrednih sosedov, katerih ocene so uporabljene pri napovedovanju.
2. Zagotavlja različne načine določanja pomembnosti sosedov v napovedi, t.j. služi kot utež za podobnost.

Računanje podobnosti je pomemben korak pri razvoju priporočilnega sistema na osnovi soseščine, saj lahko to precej vpliva na njegovo točnost in zmogljivost.

Za izračun podobnosti med dvema uporabnikoma uporabimo Pearsonov količnik (enačba 1) [2].

$$sim(a,b) = \frac{\sum_{p \in P}(r_{a,p} - \overline{r_a})(r_{b,p} - \overline{r_b})}{\sqrt{\sum_{p \in P}(r_{a,p} - \overline{r_a})^2}\sqrt{\sum_{p \in P}(r_{b,p} - \overline{r_b})^2}}$$

Enačba 1: Pearsonov količnik.

Razlaga elementov v enačbi 1

- sim(a,b) … podobnost med uporabnikoma a in b
- $P = \{p_1, \ldots, p_m\}$ … množica produktov
- $r_{i,j}$ … ocena, ki jo je podal uporabnik i produktu j
- $\overline{r_i}$ … povprečna ocena uporabnika i

Pearsonov količnik lahko zavzame vrednosti od vključno +1 (močna pozitivna soodvisnost) do vključno -1 (močna negativna soodvisnost).

2.1.3 Napovedovanje ocen

Za napovedovanje [2], v kolikšni meri bo produkt p všeč uporabniku, se moramo odločiti, katere ocene sosedov bomo uporabili (N) in kako močno bomo upoštevali njihova mnenja (uteži).
Za izračun napovedi, ki upošteva tudi razdaljo najbližjih N sosedov (podobnost oz. utež) in uporabnikovo povprečje, se uporabi enačbo 2:

$$pred(a,p) = \overline{r_a} + \frac{\sum_{b \in N} sim(a,b) * (r_{b,p} - \overline{r_b})}{\sum_{b \in N} sim(a,b)}$$

Enačba 2: Uteževanje in normiranje napovedane ocene.

Tako lahko izračunamo napovedi za vse produkte, ki jih uporabnik še ni preizkusil. Na seznam napovedi vključimo tiste z visokimi izračunanimi vrednostmi.

3. Vsebinsko osnovan priporočilni sistem

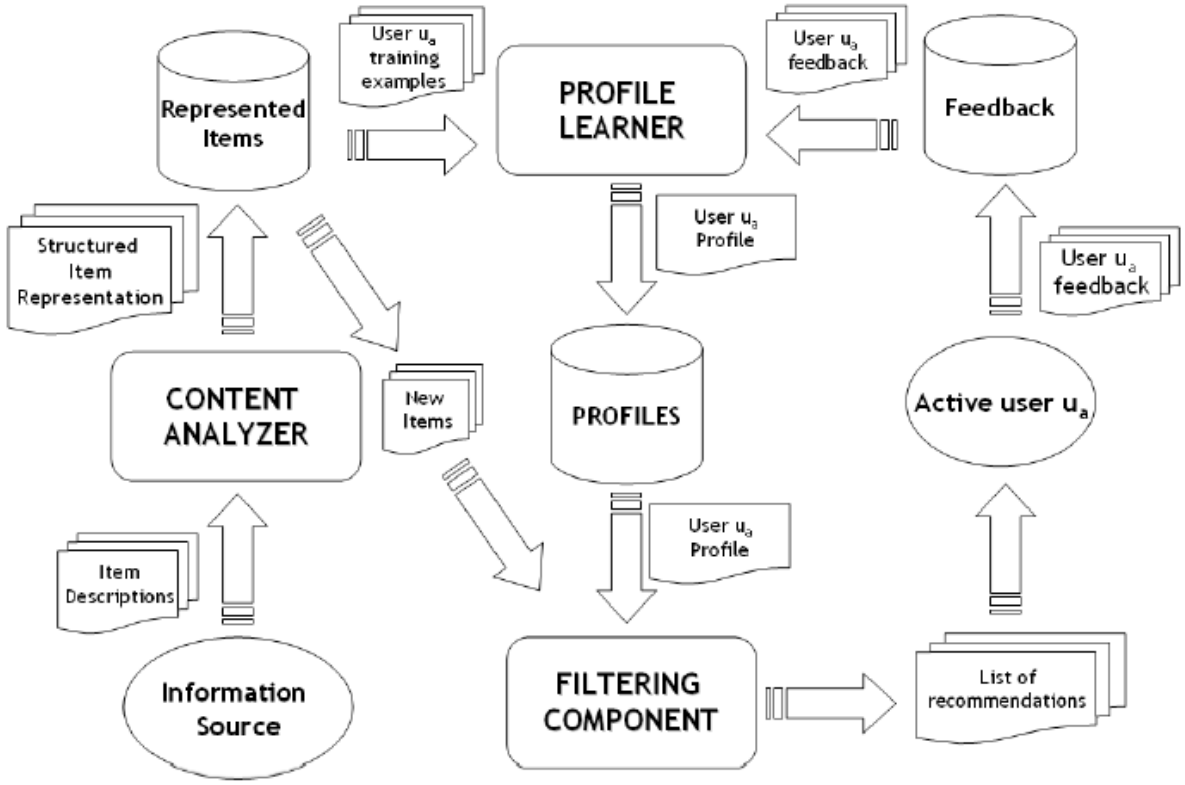
Za uporabo metod izbiranja s sodelovanjem moramo poznati ocene uporabnikov, ni nam pa potrebno nič vedeti o priporočanih produktih. Glavna prednost tega je, da se izognemo dragi nalogi preskrbovanju sistema s podrobnimi in najnovejšimi opisi produktov. Slabost pa je, da samo z uporabo metode izbiranja s sodelovanjem ni mogoče po občutku izbrati kandidatov za priporočanje, na osnovi njihovih lastnosti in točno določenih preferenc uporabnika: v realnem svetu bi enostavno priporočili film F uporabniku U, če vemo da:

a) je film F komedija,
b) uporabnik U rad gleda komedije.

Priporočilni sistem lahko opravi tovrstno nalogo, če pozna dva kosa informacije: opis lastnosti produkta in profil uporabnika, ki nekako opisuje (pretekla) zanimanja uporabnika, morda v povezavi z prednostnimi značilnostmi produkta. Naloga priporočanja je nato sestavljena iz določanja produktov, ki najbolj ustrezajo uporabnikovim preferencam. Ta proces se običajno imenuje vsebinsko osnovano priporočanje [2,. Čeprav je tovrstni pristop odvisen od dodatnih informacij o produktih in preferencah uporabnikov, ne potrebuje veliko uporabnikov ali zgodovine ocenjevanja, t.j. seznam priporočil se lahko ustvari, četudi je v bazi samo en uporabnik.

3.1 Visoko nivojska arhitektura vsebinsko osnovanih priporočilnih sistemov

Visoko nivojska arhitektura vsebinsko osnovanih priporočilnih sistemov je prikazana na sliki 5 (povzeto po [5]).



Slika 5: Visoko nivojska arhitektura vsebinsko osnovanih priporočilnih sistemov.

Proces priporočanja se izvede v treh korakih. S posameznim korakom rokuje ločena komponenta:

- Pregledovalec vsebine (*content analyzer)* – Glavna zadolžitev enote je predstavitev vsebine produktov, ki jo dobi od virov z informacijami (*information source)* v obliki, ki je primerna za naslednji korak obdelave. To predstavitev nato na vhod dobita komponenta za učenje profila in komponenta za izbiranje.
- Komponenta za učenje profila *(profile learner)* – Ta enota zbira podatke o preferencah uporabnika in poskuša posplošiti te podatke, da lahko ustvari profil uporabnika.
- Komponenta za izbiranje (*filtering component)* – Ta enota izkorišča profil uporabnika za predlaganje primernih produktov s primerjanjem predstavitve profila s predstavitvijo priporočanih produktov.

1. Prvi korak priporočilnega procesa izvaja pregledovalec vsebine. Iz vira z informacijami prejme opise produktov (*item descriptions*), iz katerih ustvari oblikovano predstavitev produktov (*structured item description*). To predstavitev se nato shrani v vir za shranjevanje predstavitev produktov (*represented items*).
2. Da se lahko ustvari in posodablja profil aktivnega uporabnika (*active user* u_a), se na nek način zbira njegove odzive in nato zapiše v vir za shranjevanje odgovorov (*feedback*). Ti odzivi, imenovani označbe ali odgovori (*user* u_a *feedback*) so skupaj s sorodnimi opisi

produkta uporabljeni med procesom učenja modela, ki je uporaben za napovedovanje dejanske koristnosti novih predstavljenih produktov.

3. Priporočilni sistem se profila uporabnika priuči tako, da komponenta za učenje profila prejme učno množico za uporabnika u_a (*user u_a training examples*), ki vsebuje ocene, ki jih je uporabnik u_a podal različnim predstavitvam produktov, in nad njimi uporabi nadzorovane učne algoritme za ustvarjanje napovednega modela – profila uporabnika (*user u_a profile*) – ki je običajno shranjen v viru za shranjevanje profilov (*profiles*) za kasnejšo uporabo v komponenti za izbiranje. Komponenta za izbiranje ob novi predstavitvi produkta (*new items*) z uporabo profila uporabnika u_a določi, ali produkt ustreza uporabnikovim preferencam. Če ustreza, uvrsti produkt na seznam priporočil (*list of recommendations*). Uporabnik nato oceni produkte na seznamu priporočil. Drugi in tretji se ponovno izvedeta, tokrat z uporabo novih znanj o uporabniku, pridobljenih z uporabnikovimi odgovori.

3.2 Predstavitev vsebine

Produkti so predstavljeni z množico lastnosti, imenovan atributi, značilke ali profili produkta.[2]

Slika 6 prikazuje primer predstavitve produktov.

product	release-date	directors
Mission To Mars	2000	Brian De Palma
Minority Report	2002	Steven Spielberg
The One	2001	James Wong
I Robot	2004	Alex Proyas
Hulk	2003	Ang Lee

Slika 6: Primer predstavitve podatkov o filmu.

Ko so uporabnikove preference opisane z uporabo lastnosti produktov, je naloga priporočilnega sistema povezati opis lastnosti produkta in opis preferenc uporabnika.
Primer opisa preferenc uporabnika je v tabeli 3.

user-id	**product**	**release-date**	**directors**
63080018	…	1989	Steven Spielberg, James Cameron

Tabela 3: Primer opisa uporabnikovih preferenc.

3.3 Računanje podobnosti glede na vsebino

Vsebinsko osnovani priporočilni sistemi priporočila podajajo tako, da izračunajo, v kolikšni meri je še nepreizkušeni produkt podoben produktom, ki so bili uporabniku v preteklosti všeč, glede na njihove lastnosti.[2]
Zanker in ostali avtorji so leta 2006 predlagali pristop, pri katerem so uporabljene različne funkcije podobnosti za različne lastnosti produkta. Te lastnosti in njihove uteži se nato uporabi za izračun povprečne podobnosti med produkti.
Tudi pri vsebinsko osnovanem priporočilnem sistemu lahko uporabimo metodo najbližjih sosedov za izbiro (primerne) soseščine podobnih produktov. Napoved za še nepreizkušeni produkt se izračuna tako, da uporabimo ocene najbližjih sosedov. Poleg velikosti soseščine so možne še ostale spremembe, kot npr. binarizacija ocen, uporaba najmanjšega praga podobnosti ali uteževanje ocen na osnovi stopnje podobnosti.

3.3.1. Jaccardov indeks

Jaccardov indeks (znan tudi pod imenom Jaccardov koeficient podobnosti) [9] meri podobnost med dvema množicama podatkov. Izračuna se ga tako, da se deli velikost preseka množic z velikostjo unije množic. Njegova vrednost lahko zavzame vrednosti med 0 in 1.

Enačba 3 prikazuje izračun Jaccardovega koeficienta.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Enačba 3: Jaccardov koeficient

Poleg Jaccardovega koeficienta poznamo tudi Ochiaiev koeficient [10]. Enačba 4 prikazuje izračun Ochiaievega koeficienta.

$$K = \frac{n(A \cap B)}{\sqrt{n(A) \times n(B)}}$$

Enačba 4: Ochiaiev koeficien.

4. Ocenjevanje priporočilnih sistemov

Priporočilni sistemi lahko uporabljajo različne metode za priporočanje, zato moramo znati poiskati najboljšo. Priporočilni sistem ocenimo tako, da izračunamo, kakšno napako naredi pri priporočanju.[2]
Danes je splošno sprejeto dejstvo, da so točne napovedi nujne, a ne zadostne za razvoj dobrega orodja za priporočanje. Priporočilni sistemi so odvisni tudi od interakcije uporabnika z računalniškimi sistemi in ostalimi uporabniki. Zato se pri ocenjevanju priporočilnih sistemov postavlja tudi vprašanja, kot so:

- Ali se uporabnikom zdi interakcija s priporočilnim sistemom uporabna ?
- Ali so zadovoljni s kakovostjo prejetih priporočil ?
- Kaj vodi ljudi k prispevanju znanja, kot so ocene in komentarji, ki povečajo kakovost priporočil sistema ?
- Kaj je uporabnikom všeč pri prejemanju priporočil ?
- Ali jim je všeč stopnja naključnosti in novosti, ali samo dejstvo, da jim je prihranjeno iskanje produkta ?

4.1 Metoda izloči enega in mera MAE za računanje napake pri priporočanju

Napako pri priporočanju se meri z metodo izloči enega (leave-one-out)[5], ki deluje tako, da iz podatkov izloči eno dejansko vrednost. Nato poskuša z uporabo metode, katero ocenjujemo, izračunati napovedano vrednost. Ko je napovedana vrednost izračunana, se lahko izračuna napako pri priporočanju. Napovedano vrednost se nato primerja z dejansko vrednostjo tako, da se izračuna absolutno razliko med njima. Metodo se izvede po vseh ocenah, povprečje napak (razlik) pa se imenuje povprečna absolutna napaka (MAE – Mean Absolute Error) [4]. Enačba 5 prikazuje izračun napake MAE.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|f(i) - \hat{f}(i)|$$

Enačba 5: Mera MAE.

$f(i)$ … dejanska vrednost

$\hat{f}(i)$ … napovedana vrednost

N … Število primerjanih vrednosti

Poleg mere MAE poznamo tudi mero koren srednje kvadratne napake (RMSE – Root Mean Squared Error) [5], ki daje več poudarka oz. kaznuje velike razlike med dejansko in napovedano vrednostjo. Enačba 6 prikazuje izračun napake RMSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|} \sum_{(u,i)\in\mathcal{T}} (\hat{r}_{ui} - r_{ui})^2}$$

Enačba 6: Mera RMSE.

$\mathcal{T}$ … Število primerjav napovedane in dejanske vrednosti

$\hat{r}_{ui}$ … Napovedana vrednost

r_{ui} … Dejanska vrednost

5. Seznanjanje s podatki in njihova obdelava

5.1 Predstavitev podatkov

Podatki vsebujejo:

- 100.000 ocen,
- 6.267 uporabnikov,
- 4.404 produktov.

Tabela 4 prikazuje, s katerimi atributi so opisani produkti, in kakšne so njihove lastnosti.

Atribut	**Ena vrednost/Več vrednosti**	**Manjkajoče vrednosti (N/4404)**
naslov produkta	ena vrednost	0
leto izdaje	ena vrednost	0
režiserji	več vrednosti	8
producenti	več vrednosti	2067
igralci	več vrednosti	58
scenaristi	več vrednosti	2137
snemalci	več vrednosti	2368
uredniki	več vrednosti	2416
skladatelji	več vrednosti	2223
spregovorjeni jeziki	več vrednosti	1729
starostna meja	več vrednosti	2028
predvideni proračun	ena vrednost	2632
države snemanja	več vrednosti	1648
trajanje	ena vrednost	10
mesta snemanja	več vrednosti	4402
filmske institucije	več vrednosti	4372
lokacije snemanja	več vrednosti	3425
žanri	več vrednosti	15
žanri, podrobneje opredeljeni	več vrednosti	1667
filmske serije	več vrednosti	4227
pisci zgodbe	več vrednosti	3388
nadaljevanja	več vrednosti	3965
predhodniki	več vrednosti	4078
tematike	več vrednosti	3584
formati filma	več vrednosti	4004

distributerji	več vrednosti	2084
produkcijske hiše	več vrednosti	2818
slogani filma	več vrednosti	3202
filmski festivali	več vrednosti	3820
izvršni producenti	več vrednosti	3027
izbira igralcev	več vrednosti	3053
oblikovalci produkcije	več vrednosti	2605
oblikovalci	več vrednosti	2717
oblikovalci scen	več vrednosti	2762
ocenjeni zaslužek	ena vrednost	4374

Tabela 4: Atributi in njihove lastnosti.

5.2 Pridobivanje atributov preko prosto dostopne storitve "Freebase"

"Freebase" [7] je prosto dostopna storitev, ki shranjuje podatke v različna skladišča. Vsako skladišče ima shranjene podatke o točno določenem področju (npr. skladišče za film, skladišče za glasbo, …).
Trenutno "Freebase" shranjuje oblikovane podatke 23 milijonov entitet. Entiteta je posamezna oseba, prostor ali stvar. "Freebase" združuje entitete v graf. Vsaka entiteta ima enolično identifikacijo, po kateri entitete ločimo med seboj. Entitete so opisane z različnimi tipi (npr. ljudje, lokacije, knjige, filmi, itd.), ki vsebujejo dodatne lastnosti, kot npr. datum rojstva za ljudi ali žanr za filme. Tipi in lastnosti so združene v shemo (scheme).
V storitev "Freebase" lahko kdorkoli prispeva podatke, lahko pa se tudi ustvari lastno shemo v bazi (base), če "Freebase" še ne vsebuje sheme za tematiko, ki je zanimiva za posameznika.

Atribute in njihove vrednosti smo pridobili z uporabo datotek, ki jih imenujemo "Freebase data dumps" in vsebujejo podatke o določenem področju. Datoteke se nahajajo znotraj skladišč storitve "Freebase".

Naša datoteka s podatki vsebuje podatke o filmih, zato je bilo potrebno iz storitve "Freebase" najprej prenesti arhiv datotek s podatki (film.tar.bz2).

Večino podatkov smo pridobili iz datoteke film.tsv. Vzorec te datoteke prikazuje slika 7.

name	id	initial_release_date	directed_by	produced_by	written_by	cinematography	edited_by	music	language	rating	estimated_	country	starring	runtime
Dating: Do	/m/01xv8f	1949	Ted Peshak	David A. Smart	Dick Creyke				English Language			United Sta	/m/0g9ftm(	/m/04q64bt
While You	/m/01gzmk	21.4.1995	Jon Turteltaub	Joe Roth,Roge	Daniel G. Sull	Phedon Papamic	Bruce Gre	Randy Ede	English Language	PG (USA)	/m/0k80rd	United Sta	/m/0bfm92	/m/0k80r2
Bride of th	/m/01l40	11.5.1956	Ed Wood, Jr.	Donald E. McC	Ed Wood, Jr.,	William C. Thom	Warren Ad	Frank Wor	English Language			United Sta	/m/0jvszp,	/m/0jvs_3,/n
For Love o	/m/0191l8	18.10.2000	Joseph Sargent						Spanish Language,English L	PG-13 (USA)		United Sta	/m/0cg6s9	/m/0bvqwy2
Chyornaya	/m/021yc04	1985	Sergei Tarasov						Russian Language			Russia		

Slika 7: Primer podatkov iz datoteke film.tsv.

Na sliki 7 lahko vidimo, da imajo nekateri atributi le eno vrednost, drugi imajo več vrednosti (npr. atribut language). Vidimo, da atributi vsebujejo tudi neznane vrednosti.
Nekateri atributi pa niso predstavljeni z dejanskimi vrednostmi, ampak z njihovimi identifikacijami (npr. atribut runtime). Pri teh je potrebno za dostop do dejanskih vrednosti pregledati še drugo datoteko tipa .tsv. Tako lahko npr. v datoteki film_cut.tsv preberemo, da identifikacija /m/04q64bt predstavlja 13 minutni film. Tabela 5 prikazuje uporabljene atribute z identifikacijskimi vrednostmi in datoteke, v katerih se lahko prebere dejansko vrednost identifikacije.

Atribut	**Uporabljene datoteke .tsv**
predvideni proračun	currency.tsv, dated_money_value.tsv
Trajanje	film_cut.tsv
Distributerji	film_film_distributor_relationship.tsv

Tabela 5: Uporabljeni atributi z identifikacijskimi vrednostmi in uporabljene datoteke .tsv.

Poleg podatkov o filmih smo potrebovali tudi podatke o denarnih vrednostih in valutah. Ti podatki se nahajajo v datotekah currency.tsv, in dated_money_value.tsv.

5.3 Spreminjanje vrednosti atributov

Vrednosti smo spremenili pri naslednjih atributih:

- spregovorjeni jeziki,
- države snemanja,
- nadaljevanja,
- predhodniki.

5.3.1 Spregovorjeni jeziki ter države snemanja

Iz vrednosti atributa smo ustvarili tri skupine: English Language (v filmu se govori le angleško), English Language + Other (v filmu se poleg angleščine pojavi še vsaj en drugi jezik) in Other (v filmu se angleško ne spregovori), saj angleški jezik v podatkih prevladuje (angleški jezik se pojavi v 93 %, francoski jezik v 0,6 %, nemški jezik v 0,3 %), in nas zanima le, ali je angleški jezik vsebovan v vrednosti tega atributa.

Pri atributu države snemanja prevladuje država Amerika (Amerika se pojavi v 85 %, Velika Britanija v 4,5 %, Nemčija v 1,9 %), zato (enako kot pri atributu spregovorjeni jeziki) iz vrednosti atributa ustvarimo tri skupine : USA, USA + Other ter Other.

5.3.2 Nadaljevanja in predhodniki

Atributa nadaljevanja in predhodniki smo združili v atribut nadaljevanja in predhodniki. Z njim naštejemo filme, ki so del neke filmske serije. Zanima nas, ali je primerjani film del neke filmske serije.
Primer: Filma Johnny English in Johnny English Reborn sta del filmske serije o Johnnyu Englishu.

5.4 Izločanje atributov

Izločanje atributov ima dve glavni prednosti: izboljšamo lahko točnost napovedovanja in olajšamo si nalogo nastavljanja uteži posameznemu atributu.

Atribut lokacije snemanja smo izločili, saj se na istih lokacijah snemajo zelo različni filmi. Posledično bi za dva produkta, ki sta si dejansko zelo različna, lahko izračunali, da sta si podobna. Iz enakega razloga smo izločili še atributa formati filma ter filmski festivali.

Atributa mesta snemanja ter filmske institucije smo izločili, saj vsebujeta preveč neznanih vrednosti.

Atribut snemalci smo izločili tako, da smo naključno izbrali 25.000 ocen izmed 100.000 in pogledali, kako dobro napoveduje osnovni priporočilni sistem (Tabela 6). Nato smo preizkusili še delovanje vsebinsko osnovanega priporočilnega sistema, če uporabimo le ta atribut (Tabela 7). Atribut smo izločili, če je napaka (MAE) pri priporočanju večja ali enaka 0,78.
Mejo 0,78 smo izbrali zaradi velike razlike med napako MAE, ki jo naredi osnovni priporočilni sistem in napako MAE, ki jo naredi vsebinsko osnovani priporočilni sistem.
Z enakim postopkom smo izločili še atribute izbira igralcev, oblikovalci produkcije, oblikovalci ter oblikovalci scen.

Priporočilni sistem	MAE	RMSE	Število primerjav
osnovni	0,699	0,916	25000

Tabela 6: Napaka pri priporočanju osnovnega priporočilnega sistema.

Atribut	MAE	RMSE	Število primerjav
leto izdaje	0,770	1,026	24488
režiserji	0,697	1,051	5339
producenti	0,757	1,076	6139
igralci	0,758	1,064	15615
scenaristi	0,715	1,066	3188
snemalci	0,790	1,125	4237
uredniki	0,770	1,114	4463
skladatelji	0,767	1,086	6196
spregovorjeni jeziki	0,758	1,015	16906
starostna meja	0,745	1,006	15014
države snemanja	0,759	1,018	17015
trajanje	0,769	1,024	24448
žanri	0,751	1,011	24088
žanri, podrobneje opredeljeni	0,741	1,003	16725
filmske serije	0,440	0,822	610
pisci zgodbe	0,577	0,928	956
nadaljevanja in predhodniki	0,461	0,853	980
tematike	0,760	1,074	2856
distributerji	0,754	1,036	13614
produkcijske hiše	0,757	1,046	9809
izvršni producenti	0,745	1,056	4268
izbira igralcev	0,798	1,121	5626
oblikovalci produkcije	0,789	1,127	3559
oblikovalci	0,797	1,126	3760
oblikovalci scen	0,799	1,121	3595

Tabela 7: Napaka vsebinsko osnovanega priporočilnega sistema pri posameznem atributu.

Atributa predvideni proračun in ocenjeni zaslužek filma vsebujeta (denarne) vrednosti, ki so odvisne od časa (npr. vrednost dolarja leta 1914 se razlikuje od vrednosti dolarja leta 2010). Ker se v diplomski nalogi ne ukvarjamo s preračunavanjem pretekle vrednosti valute v današnjo, ta dva atributa izpustimo.

V diplomski nalogi se prav tako ne ukvarjamo z obdelavo naravnega jezika, zato izpustimo atribut slogani filma.

6. Izvedba metod in testiranje priporočilnih sistemov

6.1 Osnovna metoda

Osnovna metoda [Ricci] za napovedovanje uporablja povprečja. Napovedano oceno izračuna z uporabo enačbe 1.
Z μ označimo povprečno oceno vseh ocen v podatkih. Osnovno napoved za oceno, ki jo bo uporabnik u podal produktu i, označimo z b_{ui}. Parametra b_u in b_i označujeta odstopanje uporabnika in produkta od povprečja.

$$b_{ui} = \mu + b_u + b_i$$

Enačba 7: Osnovna napoved za oceno, ki jo bo uporabnik u podal produktu i.

Pri računanju b_i ne upoštevamo ocene uporabnika u, če je gledal produkt i, ker opazujemo uporabnikovo odstopanje od povprečja (b_u) in odstopanje družbe (ostali uporabniki, ki so gledali produkt i) od povprečja (b_i).

6.2 Metoda izbiranja s sodelovanjem na osnovi uporabnikov

Računanje podobnosti uporabnikov se izračuna z uporabo Pearsonovega količnika.

Metoda za priporočanje produkta uporablja metodo najbližjih sosedov (izvedena z določanjem praga), osnovano na uporabnikih, in vrne 0, če ne najde med seboj podobnih uporabnikov, drugače vrne napovedano oceno.

Testiranje priporočilnega sistema se izvede z metodo izloči enega.

6.2.1 Izvirna skripta colrec in njena prilagoditev

Skripta colrec vsebuje različne metode izbiranja s sodelovanjem. Deluje na podatkih, ki so zapisani v obliki, kot jo prikazuje slika 8.

```
% simple ratings database file
% items section lists all available items
% users section gives user id at the beginning of the line (first line)
% and then follows a tab-delimited list of given user's ratings
% of all available items in the order they are listed in the items section

% ratings are given as numeric values
% ? is a special mark denoting user has not rated this item

% lines starting with % and empty lines are ignored


[items]

The Da Vinci Code
Ratatouille
Sister Act
Arthur et les Minimoys (Arthur and the Invisibles)
The School of Rock
Men in Black - Special Edition
Two Ninas
Miss Cast Away
Rock My World
Men In Black II


[users]

6615,           4.0,5.0,5.0,5.0,5.0,?,?,?,?,?

4564,           ?,?,?,?,?,3.0,4.0,1.0,3.0,3.0
```

Slika 8: Oblika podatkov za izvirno skripto.

Izvirno skripto smo najprej prilagodili, da deluje na zapisu trojk (identifikacijska številka uporabnika, produkt, ocena), ki je pogosteje uporabljen za zapis ocen. Tovrstni zapis prikazuje slika 9.

```
6615    The Da Vinci Code       4.0
6615    Ratatouille     5.0
6615    Sister Act      5.0
6615    Arthur et les Minimoys (Arthur and the Invisibles)      5.0
6615    The School of Rock      5.0
4564    Men in Black - Special Edition  3.0
4564    Two Ninas       4.0
4564    Miss Cast Away  1.0
4564    Rock My World   3.0
4564    Men In Black II 3.0
```

Slika 9: Oblika podatkov za prilagojeno skripto.

Nadalje smo spremenili podatkovno strukturo za hranjenje ocen. Izvirna skripta je v ta namen uporabljala seznam, mi pa uporabljamo slovar, ker je iskanje po slovarju oz. dostop do podatkov v slovarju hitrejši. Ključ slovarja je terka (identifikacijska številka uporabnika, produkt), vrednost slovarja pa ocena.

6.3 Vsebinsko osnovana metoda

Atributa filmske serije ter nadaljevanja in predhodniki sta se izkazala kot zelo dobra (Tabela 7). Atribut filmske serije je boljši od atributa nadaljevanja in predhodniki, zato za računanje podobnosti in napovedovanje ocene uporabimo samo ta atribut. Če nam ne uspe napovedati ocene, poskusimo še z uporabo atributa nadaljevanja in predhodniki. Če še vedno ni možno napovedati ocene, izračunamo podobnost in napovedano oceno po vseh preostalih atributih.
Podobnost produktov se izračuna s pomočjo funkcij za izračun podobnosti, vsako podobnost po določenem atributu pa se uteži glede na to, v kolikšni meri je produkt koristen za napovedovanje. Z enačbo 8 se izračuna podobnosti med dvema produktoma.
Podobnosti ne moremo izračunati, če produkt pri atributu, po katerem računamo podobnost, vsebuje neznano vrednost. V tem primeru za podobnost vrnemo vrednost None.

$$similarity = \frac{\sum_{a \in A} weight(a) \times similarity(a)}{\sum_{a \in A} similarity(a)}$$

Enačba 8: Izračun podobnosti.

Praga podobnosti pri tem priporočilnem sistemu nismo nastavili, pri računanju upoštevamo vse produkte.
Oceno se napove z uporabo enačbe 9. Napovedano oceno se izračuna tako, da se za posamezen produkt poišče njemu podobne produkte. Ko najdemo podoben produkt, pogledamo, kakšna je njegova ocena, in jo utežimo s podobnostjo. Na koncu napovedano oceno normiramo (delimo s podobnostjo). Še vedno se lahko zgodi, da napovedi ne moremo podati (ni podobnih produktov oz. podobnosti ni mogoče izračunati), v tem primeru sporočimo, da ne moremo podati ocene.

$$predictedRating = \frac{\sum similarity \times rating}{\sum similarity}$$

Enačba 9: Napoved ocene.

Testiranje priporočilnega sistema se izvede z metodo izloči enega.

7. Optimizacija delovanja priporočilnih sistemov

7.1 Priporočilni sistem, osnovan na metodi izbiranja s sodelovanjem

7.1.1 Izbiranje praga podobnosti

Za izboljšanje napovedovanja priporočilnega sistema je potrebno poiskati prag podobnosti, pri katerem priporočilni sistem naredi najmanjšo napako pri napovedovanju.

Prag smo določili tako, da smo priporočilni sistem testirali na vseh 100.000 ocenah in merili njegovo napako (MAE) pri različnih pragovih. Rezultati testiranja so prikazani v tabeli X.

Prag podobnosti	**MAE**	**RMSE**	**Število primerjav**
1,0	0,878	1,142	66827
0,9	0,861	1,119	71638
0,8	0,846	1,098	75433
0,7	0,830	1,078	78682
0,6	0,813	1,056	81874
0,5	0,793	1,031	84949
0,4	0,774	1,006	87511
0,3	0,756	0,982	89252
0,2	0,745	0,968	90275
0,1	0,741	0,963	90695
0,05	0,740	0,962	90827
…	0,740	0,962	…
0,004	0,740	0,962	90914
0,003	0,741	0,962	90916

Tabela 8: Izbiranje praga podobnosti.

Pri priporočilnem sistemu na osnovi metode izbiranja s sodelovanjem smo izbrali prag 0,004, saj je to zadnji (testirani) prag, pri katerem je bila napaka MAE najmanjša. S pragom, manjšim od 0,004, se napaka prične povečevati.

7.2 Vsebinsko osnovan priporočilni sistem

7.2.1 Funkcije za izračun podobnosti

Za izračun podobnosti produktov smo uporabili več funkcij.

Med običajnimi funkcijami smo izbrali Jaccardov koeficient in Ochiaiev koeficient.

Določili smo tudi posebno funkcijo za izračun podobnosti, ki deluje na naslednji način:

- če ima katerikoli izmed primerjanih produktov pri atributu, po katerem računamo podobnost, neznano vrednost, podobnosti ne moremo izračunati, zato funkcija vrne vrednost None.
- če imata oba produkta pri primerjanem atributu eno vrednost in sta ti dve vrednosti enaki, funkcija vrne 1, drugače vrne vrednost 0.
- če ima prvi produkt pri primerjanem atributu eno vrednost in drugi produkt dve vrednosti, izračunamo presek vrednosti. Če je velikost preseka enaka 0, vrne funkcija vrednost 0, drugače vrne vrednost 0,9.
- če ima prvi produkt pri primerjanem atributu eno vrednost in drugi produkt več kot dve vrednosti, izračunamo presek vrednosti. Če je velikost preseka enaka 0, vrne funkcija vrednost 0, drugače vrne vrednost 0,6.
- če ima prvi produkt pri primerjanem atributu dve vrednosti in drugi produkt eno vrednost, izračunamo presek vrednosti. Če je velikost preseka enaka 0, vrne funkcija vrednost 0, drugače vrne vrednost 0,9.
- če ima prvi produkt pri primerjanem atributu dve vrednosti in drugi produkt dve vrednosti, izračunamo presek vrednosti. Če je velikost preseka enaka 0, vrne funkcija vrednost 0, če je velikost preseka enaka 1, funkcija vrne vrednost 0,8,drugače vrne vrednost 1.
- če ima prvi produkt pri primerjanem atributu dve vrednosti in drugi produkt več kot dve vrednosti, izračunamo presek vrednosti. Če je velikost preseka enaka 0, vrne funkcija vrednost 0, če je velikost preseka enaka 1, funkcija vrne vrednost 0,8, drugače vrne vrednost 1.
- če ima prvi produkt pri primerjanem atributu več kot dve vrednosti in drugi produkt eno vrednost, izračunamo presek vrednosti. Če je velikost preseka enaka 0, vrne funkcija vrednost 0, drugače vrne vrednost 0,6.
- če ima prvi produkt pri primerjanem atributu več kot dve vrednosti in drugi produkt dve vrednosti, izračunamo presek vrednosti. Če je velikost preseka

enaka 0, vrne funkcija vrednost 0, če je velikost preseka enaka 1, funkcija vrne vrednost 0,8, drugače vrne vrednost 1.

- če ima prvi produkt pri primerjanem atributu več kot dve vrednosti in drugi produkt več kot dve vrednosti, izračunamo presek vrednosti. Če je velikost preseka enaka 0, vrne funkcija vrednost 0, če je velikost preseka enaka 1, vrne funkcija vrednost 0,8, če je velikost preseka enaka 2, vrne funkcija vrednost 0,9, drugače vrne vrednost 1.

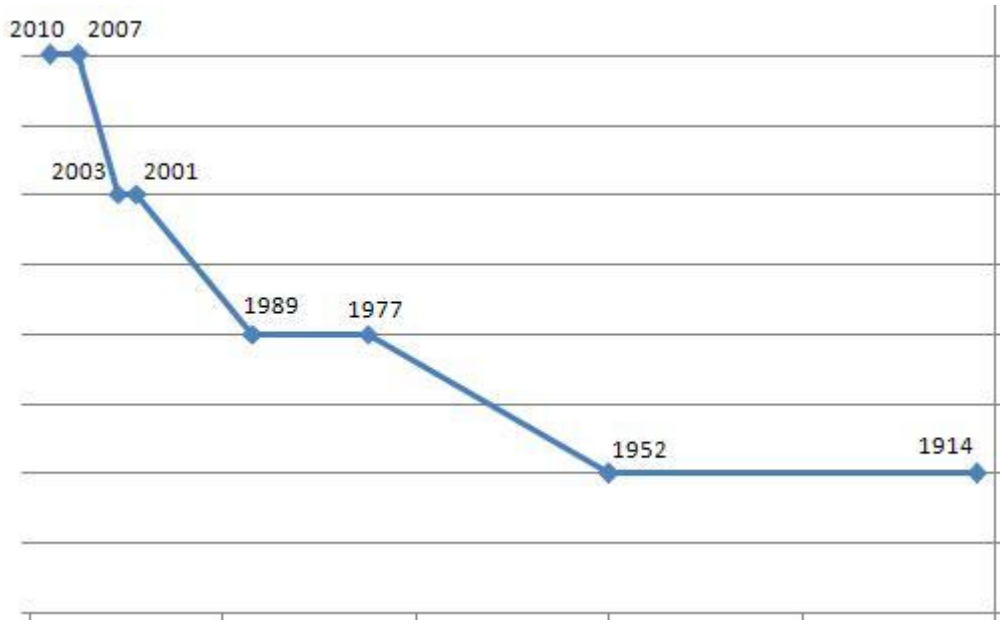
Pri tej posebni funkciji podobnosti se lahko nastavlja vrednosti, ki jih funkcija vrača.

Posebne funkcije podobnosti smo določili tudi za naslednje atribute::

- leto izdaje,
- spregovorjeni jeziki,
- starostna meja,
- države snemanja,
- trajanje,
- filmske serije,
- nadaljevanja in predhodniki.

7.2.1.1 Leto izdaje

Funkcija za izračun podobnosti po atributu leto izdaje je prikazana na sliki 10.



Slika 10: Funkcija za izračun podobnosti po atributu leto izdaje.

V ravnini 2010 do 2007 se zajame novejše filme. V ravnini 2003 do 2001 se zajame npr. trilogijo Gospodar prstanov. V ravnini 1989 do 1977 se npr. zajame prve tri dele filmske serije Indiana Jones in prve tri dele filmske serije Star Wars (pri obeh filmskih serijah sta sodelovala igralec Harrison Ford in režiser/scenarist George Lucas), v ravnini 1952 in 1914 se zajame črno-bele filme.

Za produkta, ki sta v isti ravnini, funkcija vrne vrednost 1, drugače izračuna razliko med letoma, razliko deli s 100, odšteje izračunano vrednost od 1 in vrne podobnost med produktoma. Če za enega izmed produktov ne poznamo vrednosti atributa, funkcija vrne vrednost None.

7.2.1.2 Spregovorjeni jeziki in države snemanja

Funkcija za izračun podobnosti po atributih spregovorjeni jeziki/države snemanja deluje na naslednji način:

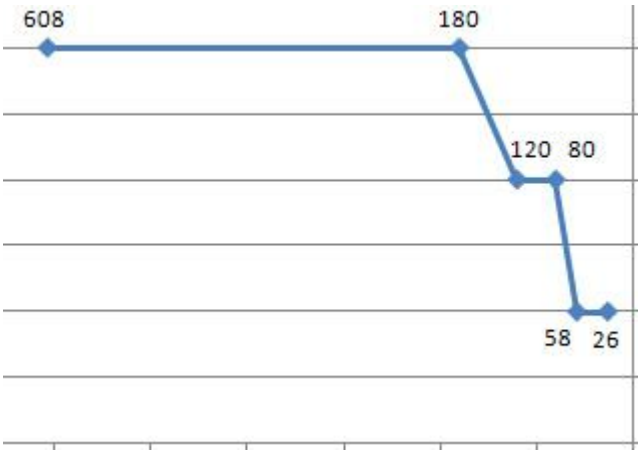
- če ima eden od obeh produktov pri tem atributu neznano vrednost, funkcija vrne vrednost None.
- če je vrednost atributa pri obeh produktih English Language, funkcija vrne vrednost 1.
- če je vrednost atributa pri obeh produktih English Language + Other, funkcija vrne vrednost 0.8.
- če je vrednost atributa pri obeh produktih Other, funkcija vrne vrednost 0.9.
- če je vrednost atributa pri enem produktu English Language in pri drugem produktu English Language + Other, funkcija vrne vrednost 0.7.
- če je vrednost atributa pri enem produktu English Language in pri drugem produktu Other, funkcija vrne vrednost 0.
- če je vrednost atributa pri enem produktu English Language + Other in pri drugem produktu Other, funkcija vrne vrednost 0.6.

7.2.1.3 Starostna meja

Funkcija za izračun podobnosti preveri, ali imata produkta vsaj eno starostno mejo enako. Če jo imata, vrne vrednost 1, sicer vrne vrednost 0. V primeru neznanih vrednosti, funkcija vrne vrednost None.

7.2.1.4 Trajanje

Funkcija za izračun podobnosti po atributu trajanje je prikazana na sliki 11.



Slika 11: Funkcija za izračun podobnosti po atributu trajanje.

V ravnini 608 do 180 se zajame zelo dolgo trajajoče filme. V ravnini 120 do 80 se zajame povprečno trajajoče filme. V ravnini 58 do 25 se zajame kratke filme.

Za produkta, ki sta v isti ravnini, funkcija vrne vrednost 1, drugače izračuna razliko med letoma, razliko deli s 100, odšteje izračunano vrednost od 1 in vrne podobnost med produktoma. Če je podobnost manjša od 0, funkcija vrne vrednost 0. Če za enega izmed produktov ne poznamo vrednosti atributa, funkcija vrne vrednost None.

7.2.1.5 Filmske serije

Funkcija za izračun podobnosti po atributu filmske serije najprej preveri, ali imata oba produkta pri tem atributu neznano vrednost. V tem primeru vrne vrednost None. Drugače izračuna presek obeh vrednosti. Če presek obstaja, vrne vrednost 1, drugače vrne vrednost 0.
Atribut lahko vsebuje neznano vrednost, eno vrednost ali več vrednosti.
Primer: Film American Pie Presents: Beta House vsebuje pri atributu filmske serije vrednosti American Pie Presents film series,American Pie film series, film American Pie 2 pa vrednosti American Pie film series. Če želimo ta atribut pravilno izkoristiti (filma sta si očitno podobna), moramo preveriti, ali obstaja med obema vrednostma presek. Če obstaja, rečemo, da sta si filma podobna. V nasprotnem primeru si filma nista podobna.

7.2.1.6 Nadaljevanja in predhodniki

Če je vrednost tega atributa pri obeh produktih neznana, funkcija vrne vrednost None. Če je pri enem od obeh produktov vrednost neznana, funkcija vrne vrednost 0. Če nobeden od teh dveh pogojev ni izpolnjen, funkcija preveri, ali obstaja presek med obema vrednostma. Če obstaja, funkcija vrne vrednost 1. Če tudi ta zadnji pogoj ni izpolnjen pa funkcija preveri še, ali se naslov drugega produkta pojavi v vrednosti atributa nadaljevanja in predhodniki prvega produkta. Če se pojavi, funkcija vrne vrednost 1, drugače vrne vrednost 0.

7.2.2 Izbiranje najbolj primerne funkcije za izračun podobnosti po posameznem atributu

Pri atributih, kjer lahko uporabimo več funkcij podobnosti, je potrebno za natančnejše napovedovanje izbrati najbolj primerno za računanje podobnosti.
Naloge smo se lotili tako, da smo iz naših podatkov s 100.000 ocenami naključno izbrali 25.000 ocen. Nato smo za posamezen atribut preizkusili tri različne funkcije (Jaccard, Ochiai in posebna funkcija) in izmerili njihovo napako pri priporočanju (MAE). Rezultati so zbrani v tabeli 9. Odebeljene vrednosti MAE označujejo najmanjšo storjeno napako pri priporočanju.

Atribut	Jaccard	Ochiai	Posebna funkcija
režiserji	MAE: 0.6976	MAE: 0.6971	**MAE: 0.6966**
	RMSE: 1.0517	RMSE: 1.0510	RMSE: 1.0506
producenti	**MAE: 0.7571**	MAE: 0.7576	MAE: 0.7573
	RMSE: 1.0756	RMSE: 1.0750	RMSE: 1.0744
igralci	**MAE: 0.7584**	MAE: 0.7594	MAE: 0.7639
	RMSE:1.0645	RMSE: 1.0649	RMSE: 1.0684
scenaristi	**MAE: 0.7155**	MAE: 0.7164	MAE: 0.7157
	RMSE: 1.0662	RMSE: 1.0665	RMSE: 1.0664
snemalci	MAE: 0.7908	MAE: 0.7905	**MAE: 0.7903**
	RMSE: 1.1252	RMSE: 1.1250	RMSE: 1.1249
uredniki	MAE: 0.7722	MAE: 0.7709	**MAE: 0.7699**
	RMSE: 1.1180	RMSE: 1.1153	RMSE: 1.1142
skladatelji	MAE: 0.7679	MAE: 0.7668	**MAE: 0.7666**
	RMSE: 1.0868	RMSE: 1.0856	RMSE: 1.0855
žanri	**MAE: 0.7513**	MAE: 0.7531	MAE: 0.7558
	RMSE: 1.0111	RMSE: 1.0123	RMSE: 1.0151
žanri, podrobneje opredeljeni	**MAE: 0.7412**	MAE: 0.7425	MAE: 0.7475
	RMSE: 1.0029	RMSE: 1.0033	RMSE: 1.0080
pisci zgodbe	MAE: 0.5802	MAE: 0.5786	**MAE: 0.5774**

	RMSE: 0.9318	RMSE: 0.9302	RMSE: 0.9281
tematike	MAE: 0.7629	MAE: 0.7610	**MAE: 0.7606**
	RMSE: 1.0766	RMSE: 1.0735	RMSE: 1.0736
distributerji	MAE: 0.7583	MAE: 0.7558	**MAE: 0.7539**
	RMSE: 1.0428	RMSE: 1.0384	RMSE: 1.0358
produkcijske hiše	MAE: 0.7599	MAE: 0.7588	**MAE: 0.7570**
	RMSE: 1.0499	RMSE: 1.0476	RMSE: 1.0457
izvršni producenti	**MAE: 0.7446**	MAE: 0.7453	MAE: 0.7454
	RMSE: 1.0561	RMSE: 1.0554	RMSE: 1.0547
izbira igralcev	MAE: 0.8012	MAE: 0.8005	**MAE: 0.7986**
	RMSE: 1.1247	RMSE: 1.1224	RMSE: 1.1207
oblikovalci produkcije	**MAE: 0.7894**	MAE: 0.78960	MAE: 0.7897
	RMSE: 1.1266	RMSE: 1.1264	RMSE: 1.1263
oblikovalci	MAE: 0.8000	MAE: 0.79908	**MAE: 0.7974**
	RMSE: 1.1288	RMSE: 1.1269	RMSE: 1.1262
oblikovalci scen	MAE: 0.8008	MAE: 0.7999	**MAE: 0.7993**
	RMSE: 1.1216	RMSE: 1.1206	RMSE: 1.1208

Tabela 9: Izbiranje najbolj primerne funkcije za izračun podobnosti po posameznem atributu.

7.2.3 Določanje uteži za posamezen atribut

Vsak atribut prispeva različen delež pri podajanju priporočila. Za izboljšanje točnosti napovedovanja je potrebno atribute, ki pozitivno prispevajo k točnosti napovedovanja, bolj upoštevati (dobijo večjo utež) kot tisti, ki nič ne prispevajo oz. negativno prispevajo k točnosti napovedovanja.

Zopet smo se naloge lotili tako, da smo uporabili 25.000 ocen izmed 100.000. Vsem atributom, razen atributoma filmske serije ter nadaljevanja in predhodniki, smo nastavili utež 50 in izračunali napako pri priporočanju. Rezultat je prikazan v tabeli 10.

Atribut	**Utež**	**MAE**	**RMSE**	**Število primerjav**
vsi (razen filmske serije ter nadaljevanja in predhodniki)	50	0.7615	1.0221	24488

Tabela 10: Vsi atributi (razen atributov filmske serije ter nadaljevanja in predhodniki) z utežjo 50.

Nato smo posameznemu atributu znižali utež (na 1 in 25) in zvišali utež (75 in 100) in opazovali, kaj se dogaja z napako. Uporabljene uteži imajo vrednosti med vključno 0 in 100 in so med seboj različne.

V tabeli 11 so prikazani rezultati, atributi so razporejeni po velikosti uporabljene uteži. Odebeljene vrednosti MAE označujejo najmanjšo storjeno napako pri priporočanju.

Atribut	**Testirana utež = 1**	**Testirana utež = 25**	**Testirana utež = 75**	**Testirana utež = 100**	**Uporabljena utež**
žanri	MAE: 0.7636	MAE: 0.7625	MAE: 0.7606	**MAE: 0.7598**	100
	RMSE: 1.0247	RMSE: 1.0233	RMSE: 1.0211	RMSE: 1.0204	
režiserji	MAE: 0.7622	MAE: 0.7618	MAE: 0.7611	**MAE: 0.7608**	96
	RMSE: 1.0230	RMSE: 1.02250	RMSE: 1.0218	RMSE: 1.0215	
igralci	MAE: 0.7621	MAE: 0.7618	MAE: 0.7611	**MAE: 0.7609**	92
	RMSE: 1.0229	RMSE: 1.0224	RMSE: 1.0218	RMSE: 1.0216	
starostna meja	MAE: 0.7620	MAE: 0.7617	MAE: 0.7612	**MAE: 0.7611**	88
	RMSE: 1.0227	RMSE: 1.0223	RMSE: 1.02188	RMSE: 1.0217	
žanri, podrobneje opredeljeni	MAE: 0.7618	MAE: 0.7616	MAE: 0.7613	**MAE: 0.7612**	84
	RMSE: 1.0223	RMSE: 1.0222	RMSE: 1.0212	RMSE: 1.0219	
spregovorjeni jeziki	MAE: 0.7615	MAE: 0.7615	**MAE: 0.7614**	MAE: 0.7615	75
	RMSE: 1.0221	RMSE: 1.0221	RMSE: 1.0221	RMSE: 1.0221	
distributerji	MAE: 0.76149	MAE: 0.76147	**MAE: 0.76145**	MAE: 0.76146	70
	RMSE: 1.02204	RMSE: 1.02206	RMSE: 1.02214	RMSE: 1.02220	
države snemanja	**MAE: 0.7613**	MAE: 0.7614	MAE: 0.7615	MAE: 0.7615	23
	RMSE: 1.0220	RMSE: 1.0221	RMSE: 1.0221	RMSE: 1.02222	
tematike	**MAE: 0.7614**	MAE: 0.7614	MAE: 0.7615	MAE: 0.7615	20
	RMSE: 1.0220	RMSE: 1.0220	RMSE: 1.0221	RMSE: 1.0221	
produkcijske hiše	**MAE: 0.7613**	MAE: 0.7614	MAE: 0.7615	MAE: 0.7616	17
	RMSE: 1.0220	RMSE: 1.0220	RMSE: 1.0221	RMSE: 1.0222	

pisci zgodbe	**MAE: 0.7613**	MAE: 0.7614	MAE: 0.7615	MAE: 0.7616	14
	RMSE: 1.0220	RMSE: 1.0220	RMSE: 1.0221	RMSE: 1.0222	
producenti	**MAE: 0.76126**	MAE: 0.76135	MAE: 0.76156	MAE: 0.76167	11
	RMSE: 1.0219	RMSE: 1.0220	RMSE: 1.0222	RMSE: 1.0223	
scenaristi	**MAE: 0.7612**	MAE: 0.7613	MAE: 0.7616	MAE: 0.7617	9
	RMSE: 1.0219	RMSE: 1.0220	RMSE: 1.0222	RMSE: 1.0223	
skladatelji	**MAE: 0.7612**	MAE: 0.7613	MAE: 0.7616	MAE: 0.7617	7
	RMSE: 1.0219	RMSE: 1.0220	RMSE: 1.0222	RMSE: 1.0223	
uredniki	**MAE: 0.7611**	MAE: 0.7613	MAE: 0.7616	MAE: 0.7617	4
	RMSE: 1.0219	RMSE: 1.0220	RMSE: 1.0222	RMSE: 1.0224	
izvršni producenti	**MAE: 0.7612**	MAE: 0.7613	MAE: 0.7616	MAE: 0.7617	3
	RMSE: 1.0218	RMSE: 1.0220	RMSE: 1.0222	RMSE: 1.0223	
trajanje	**MAE: 0.7601**	MAE: 0.7609	MAE: 0.7618	MAE: 0.7620	2
	RMSE: 1.0209	RMSE: 1.0216	RMSE: 1.0225	RMSE: 1.0227	
leto izdaje	**MAE: 0.7591**	MAE: 0.7605	MAE: 0.7620	MAE: 0.7624	1
	RMSE : 1.0196	RMSE: 1.0210	RMSE: 1.0228	RMSE: 1.0234	

Tabela 11 : Določanje uteži za posamezen atribut.

7.2.3.1 Razlaga uporabljenih uteži

Atributi žanri, režiserji, igralci, starostna meja ter žanri, podrobneje opredeljeni naredijo najmanjšo napako pri testirani uteži 100. Nato primerjamo med seboj vrednosti napak teh atributov.

Primer: Pri atributu žanri je manjša napaka kot pri atributu režiserji. Zato dobi atribut žanri večjo utež.

Enako velja za atributa spregovorjeni jeziki ter distributerji, pri katerih se najmanjšo napako naredi z uporabo uteži 75.

Ravno obratno pa velja za preostale atribute, pri katerih se najmanjšo napako naredi z uporabo uteži 1. Atribut, pri katerem se naredi manjšo napako, dobi manjšo utež.

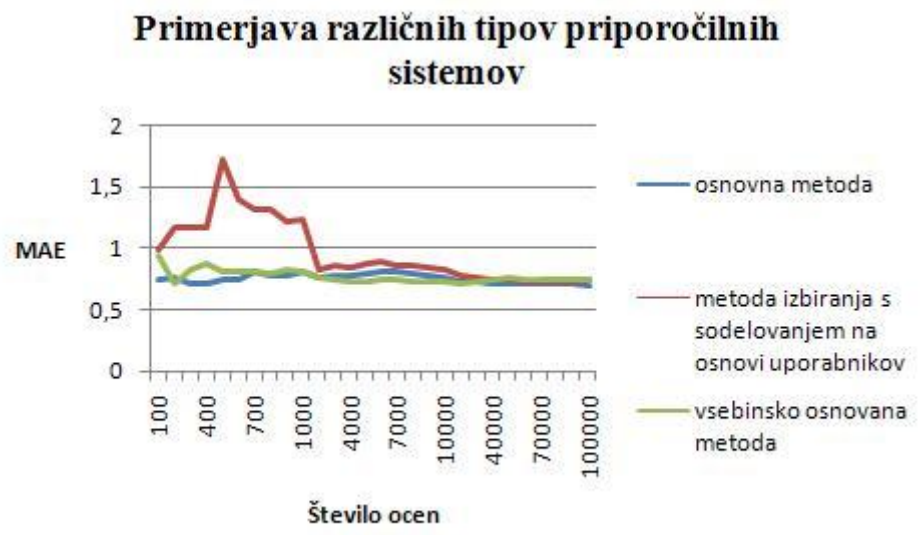
Primer: Pri atributu leto izdaje je manjša napaka kot pri atributu trajanje. Zato dobi atribut leto izdaje manjšo utež.

8. Primerjava različnih tipov priporočilnih sistemov

Primerjamo tri metode za priporočanje: osnovno metodo, metodo izbiranja s sodelovanjem na osnovi uporabnikov in vsebinsko osnovano metodo. Zanima nas, kako učinkovite so te tri metode, ko se povečuje število ocen.

Ocene smo naključno izbirali tako, da ima vsak uporabnik vsaj 5 ocen. S tem omilimo problem hladnega zagona za nove uporabnike.

Primerjava različnih tipov priporočilnih sistemov je prikazana na sliki 12.

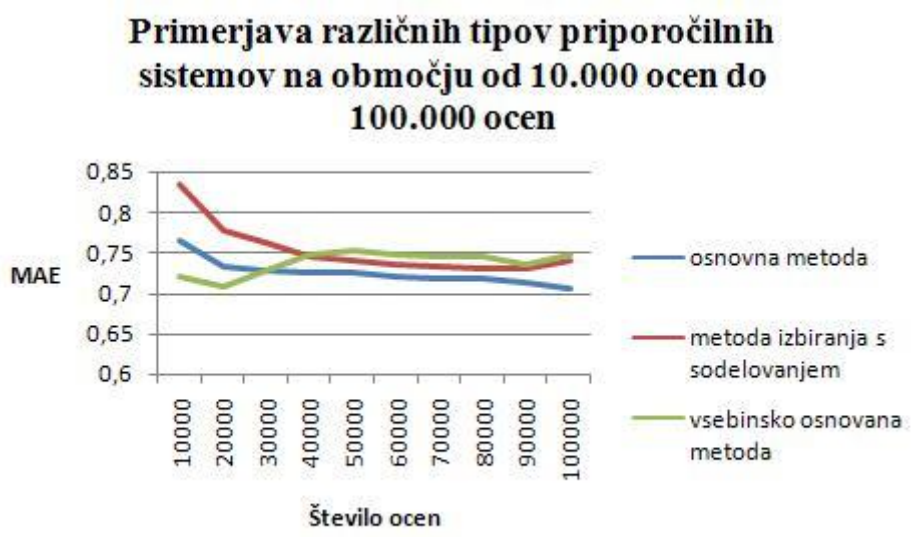


Slika 12: Primerjava različnih tipov priporočilnih sistemov.

Metoda izbiranja s sodelovanjem na osnovi uporabnikov pri majhnem številu ocen ne deluje (opazno naraščanje/padanje napake MAE). Delovati začne šele pri 2000 ocenah. Ta ugotovitev potrjuje trditev, da je slabost metode izbiranja s sodelovanjem problem majhnega števila ocen.

Delno smo pokazali prednost metode izbiranja s sodelovanjem, da se kakovost priporočil z večanjem števila ocen izboljšuje (od 6.000 ocen do 90.000 ocen napaka pada).

Pobližje si poglejmo, kaj se dogaja z metodami med 10.000 ocenami in 100.000 ocenami (Slika 13).



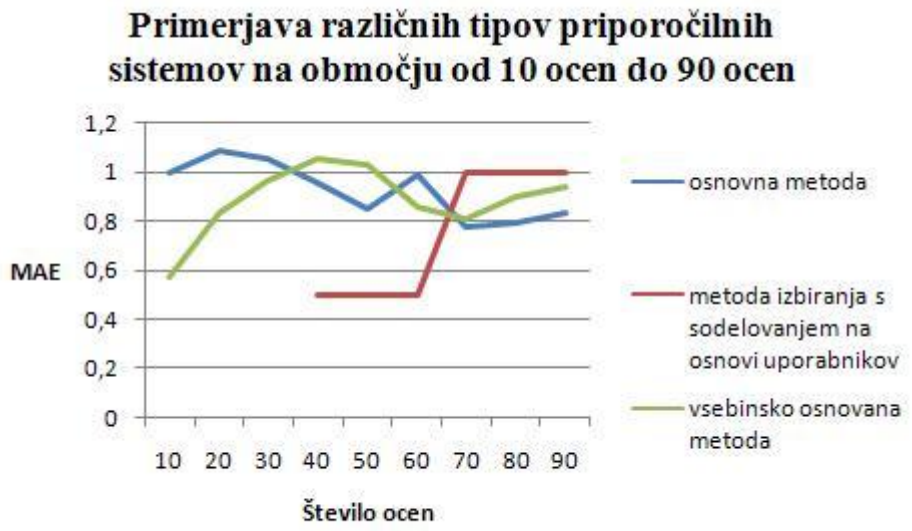
Slika 13: Primerjava različnih tipov priporočilnih sistemov na območju od 10.000 ocen do 100.000 ocen.

Presenetljiv rezultat je, da z uporabo osnovne metode naredimo najmanjšo napako. Pričakovali bi, da naprednejše metode (osnovane na znanju) napovedujejo bolje od osnovnih.

Pri majhnem številu ocen (Slika 14) opazimo, da se z uporabo metode izbiranja s sodelovanjem na osnovi uporabnikov ne da napovedati nobene ocene, ker v podatkih nimamo podobnih uporabnikov, ko je v podatkih 10, 20 in 30 ocen. S tem smo uspeli pokazati, da sta slabosti metode izbiranja s sodelovanjem na osnovi uporabnikov pomanjkanje informacij (hladni zagon) za nove uporabnike in problem "črne ovce".

Pokazali smo, da vsebinsko osnovana metoda ni odvisna od preostalih uporabnikov (v primerjavi z metodo izbiranja s sodelovanjem na osnovi uporabnikov nima težav z priporočanjem pri 10, 20 in 30 ocenah). Enako velja za osnovno metodo.

Nismo uspeli pokazati, da kakovost priporočil vsebinsko osnovanih priporočilnih sistemov s časom narašča.



Slika 14: Primerjava različnih tipov priporočilnih sistemov na območju od 10 ocen do 90 ocen.

9. Zaključek

V diplomski nalogi smo si pogledali primerjavo treh različnih vrst priporočilnih sistemov: osnovno metodo, metodo izbiranja s sodelovanjem na osnovi uporabnikov in vsebinsko osnovane metode.

Ugotovili smo, da metoda izbiranja s sodelovanjem na osnovi uporabnikov ne deluje dobro, kadar je na razpolago malo ocen. Delno smo uspeli dokazati, da se kakovost priporočil z večanjem števila ocen pri metodi izbiranja s sodelovanjem izboljšuje. Pokazali smo, da sta slabosti metode izbiranja s sodelovanjem na osnovi uporabnikov pomanjkanje informacij (hladni zagon) pri novih uporabnikih in problem "črne ovce". Pokazali smo, da vsebinsko osnovana metoda ni odvisna od preostalih uporabnikov (v primerjavi z metodo izbiranja s sodelovanjem na osnovi uporabnikov nima težav z priporočanjem pri 10, 20 in 30 ocenah). Enako smo dokazali tudi za osnovno metodo.

Pri razvoju priporočilnih sistemov je potrebno ogromno časa nameniti nastavljanju parametrov in testiranju, zato je ostalo še mnogo stvari nepreizkušenih, s katerimi bi lahko izboljšali točnost napovedovanja vsebinsko osnovanih metod:

- nastavljanje parametrov pri posebnih funkcijah podobnosti,
- zmanjšanje primerjane množice pri tistih atributih, kjer je najboljša funkcija za izračun podobnosti, funkcija Jaccard. Primerjani množici bi zmanjšali tako, da bi upoštevali le tiste vrednosti, ki so kontroverzne (prejemajo zelo različne ocene) in tiste vrednosti, ki se v podatkih največkrat pojavijo,
- uporaba večje uteži pri posameznem produktu.

Dostop do podatkov bi lahko pohitrili z uporabo podatkovne baze.

Primerjava priporočilnih sistemov bi bila lahko izvedena še glede na:

- število uporabnikov,
- hladni zagon.

10. Literatura in viri

[1] Burke, R.: Hybrid Recommender Systems: Survey and Experiments. UMUAI 12 (4), 331-370 (2002)

[2] D. Jannach et al., Recommender Systems: An Introduction, Cambridge University Press, 2011

[3] D. Jannach & G. Friedrich: Recommender Systems Tutorial (prosojnice), 2011

[4] I.Kononenko in M. Robnik Šikonja: Inteligentni sistemi. Založba FE in FRI, Ljubljana, 2010

[5] Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor, Recommender Systems Handbook, Springer 2010

[6] Gabrijel Tomšič, Neža Mramor-Kosta, Bojan Orel: Matematika I; Ljubljana, Fakulteta za elektrotehniko in računalništvo, 2004

[7] (2012) Podatki o "Freebase".
Dostopno na: http://wiki.freebase.com/wiki/What_is_Freebase%3F

[8] (2012) Spletna stran z podatki o filmih IMDb.
Dostopno na: http://www.imdb.com/

[9] (2012) Wikipedia – Jaccardov koeficient.
Dostopno na: http://en.wikipedia.org/wiki/Jaccard_index

[10] (2012) Wikipedia – Ochiaiev koeficient.
Dostopno na: http://en.wikipedia.org/wiki/Cosine_similarity#Ochiai_coefficient