SBNZ Predlog projekta
Book recommender system

Aleksandar Nikolić SW27/2017, Milan Jokanović SW28/2017

Motivacija

Ideja sistema je da olakša rad bibliotekama i poboljša user experience. Odlučiti se za narednu knjigu za čitanje može biti frustrirajuće, pogotovo kad je izbor praktično neograničen. Ideja ovog sistema je da reši taj problem i preporuči čitaocu narednu knjigu na osnovu njegovih preferencija. Pored preporuke čitaocima, sistem omogućava bibliotekama preporuku koji tip knjige je popularan i šta bi imalo smisla nabaviti u budućnosti.

Pregled problema

Ovaj sistem za cilj ima da pruži korisnicima biblioteka najbolje moguće iskustvo u potrazi za novom knjigom za čitanje. Prednost u odnosu na dosadašnje radove jeste to što ćemo osim preporuka za knjige koje sistem daje na osnovu istorije čitanja korisnika omogućiti i opciju za custom pretragu knjige po parametrima. Ovo će omogućiti korisnicima da dođu do naslova koji nisu u skladu sa onim što su do sad čitali, a postalo im je od interesa. Takođe cilj sistema je da pomogne u administraciji biblioteke, tako što će davati izveštaj o najčitanijim knjigama, piscima i žanrovima preko kojih će admin sistema moći da zaključi koje nove knjige bi mogao da nabavi.

Metodologija rada

- Ulaz
    - U slučaju preporuke preko forme ulaz su podaci sa same forme, sa tim što nisu sva polja u formi obavezna (tj. Korisnik može da ignoriše neki parametar tako što ga neće uneti). Forma će nuditi unos željenih žanrova, pisaca, opseg godina izdavanja, dužinu, uzrast kojem je namenjena, baziranost na pravom događaju, minimalnu prosečnu ocenu, nagrade, broj recenzija, pripadnost serijalu i književnu epohu.
    - U slučaju preporuke za korisnika, ulaz je istorija tog korisnika, recenzije, knjige koje su drugi ljudi dobro ocenili a čitali su knjige kao i nadgledani korisnik itd.
    - Za izveštaj je ulaz generalna popularnost knjiga, pisaca, žanrova.
    - Na osnovu svakog poklapanja atributa knjige sa parametrom sa ulaza, povećava se ocena toj knjizi.
- Izlaz
    - Izlaz iz sistema nakon traženja preporuke putem forme je 5 najbolje rangiranih knjiga.

    - Izlaz iz sistema na osnovu istorije čitanja korisnika je takođe 5 najbolje rangiranih knjiga.
    - Izlaz iz sistema za admina (biblioteku) su izveštaji najpopularnijih knjiga, pisaca i žanrova čiji bi cilj bio da pomognu pri odabiru knjiga za proširenje ponude.
- Baza znanja
    - Sistem će sadržati bazu dostupnih knjiga, koja će biti proširiva.
    - Svaka knjiga će sadržati atribute (žanrovi, autor(i), godina izdavanja, dobijene nagrade, dužinu (broj stranica), period pripadanja, da li je bazirana na pravom događaju, uzrast kojem je namenjena, prosečnu ocenu, broj ocena, pripadnost serijalu)
    - U sistemu će postojati korisnici i za svakog korisnika će se pamtiti koje je knjige pročitao, ocene dodeljene knjigama, najčitaniji pisci i žanrovi, da li voli ili izbegava serijale)
    - Takođe će sistem u globalu pratiti najpopularnije knjige, pisce i žanrove i to znanje će se iskoristiti da ustanovi asistira u daljoj nabavci knjiga.

Pravila

Primer pravila 1 (automatska preporuka za ulogovanog korisnika):

Kada korisnik uđe na stranicu za automatske preporuke, okida se lanac pravila koji će vratiti 5 najbolje rangiranih knjiga koje korisnik nije pročitao na osnovu njegovih karakteristika. Na prvom nivou rezonovanja sistem filtrira knjige na osnovu starosne grupe čitaoca. Na drugom nivou rezonovanja sistem ocenjuje knjige na osnovu žanrova koje je korisnik čitao. Potom ocenjuje pisce koje je čitao. Na četvrtom nivou gleda čitaoce (sličnog uzrasta) koji imaju zajedničke knjige sa subjektom (bonus ako su i subjekat i drugi čitalac dali sličnu ocenu knjizi) i ocenjuje njihove ostale knjige (koje subjekat nije pročitao). Na petom nivou rezonovanja, traži trend u istoriji čitaoca za podatke o dužini knjige, nagradama koje je knjiga osvojila, da li je bazirana na istinitom događaju, pripadnost određenom književnom stilu, postojanje ekranizacije. Parametri na petom nivou su manje bitni i služe za finu korekciju preporuka. Na poslednjem nivou rezonovanja sistem prepoznaje da li je neka knjiga, u 5 najbolje rangiranih, članica nekog serijala, i ako jeste, automatski se zamenjuje sa prvom knjigom tog serijala koju korisnik nije čitao (kako bi korisnik uvek čitao serijal u odgovarajućem redosledu). Čitalac ima mogućnost da kaže sistemu da ignoriše neke od ovih parametara.

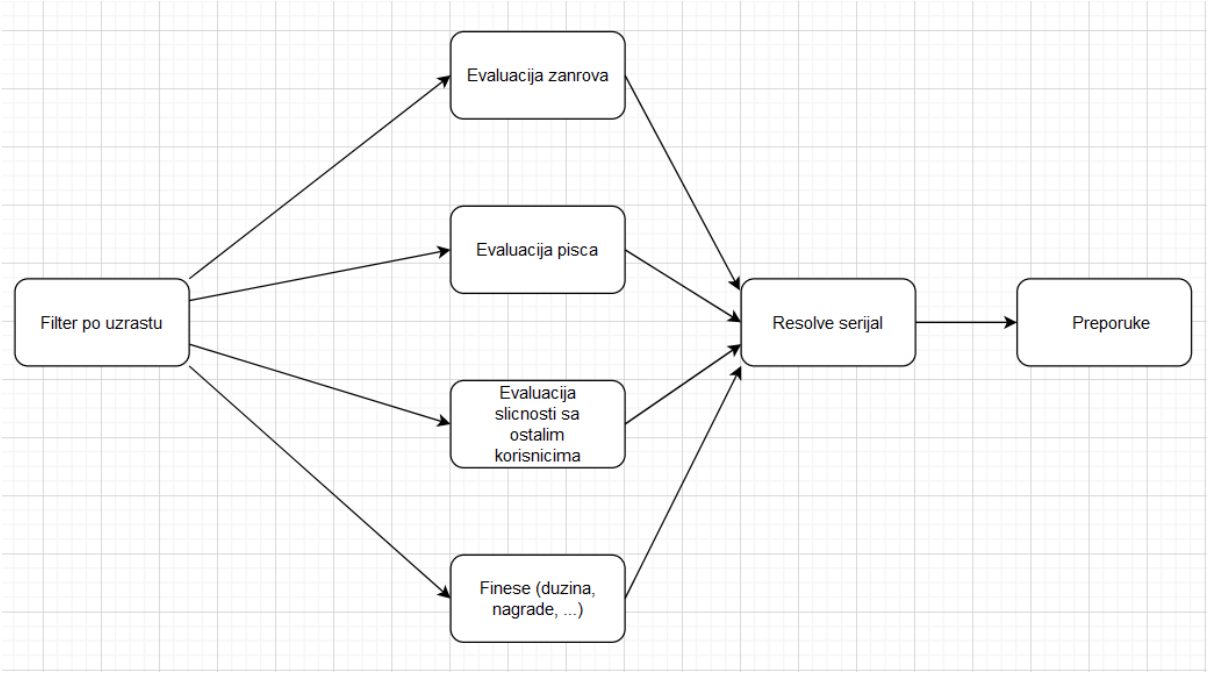
Primer: Recimo da je čitalac star 13 godina. U prvom koraku se filtriraju knjige za ovu starosnu grupu. Recimo da je korisnik čitao i visoko ocenio serijale “Percy Jackson” (fantastika, grčka mitologija, akcija, avantura), “Narnia” (fantastika, avantura) i “The tapestry series” (fantastika, avantura, school, akcija). Uočavamo da svi ovi serijali imaju žanr fantastike i avanture, te ostalim knjigama sa tim žanrovima povećavamo ocenu. Potom gledamo pisce i povećavamo ocenu njihovim ostalim knjigama, težina ovog koraka je srazmerna broju pročitanih knjiga određenog pisca. Potom gledamo druge korisnike i recimo da imamo korisnike koji su pročitali i visoko

ocenili sve ovo plus “Harry Potter” serijal i “Alex rider” serijal. Ovde vidimo da serijal “Harry Potter” ima jako poklapanje po žanrovima, i vidimo poklapanje ocena sa serijalima koje je i subjekat pročitao. Nakon ovog nivoa gledaju se finesse do tad rangiranih knjiga (dužina, nagrade, …). Finalno se vrši filter za serijal i korisnik dobija listu 5 najbolje ocenjenih knjiga za njega u tom trenutku. To bi u ovom slučaju bio recimo “Harry Potter” serijal, gde bi sistem preporučio prvu knjigu serijala koju subjekat nije pročitao, iako je neka kasnija potencijalno imala bolju ocenu.

Ulaz: Korisnik star 13 godina, pročitao serijale “Percy Jackson” (Rick Riordan, ocena 9, {fantastika, grčka mitologija, akcija, avantura}), “The tapestry series” (Henry H. Neff, ocena 8, {fantastika, avantura}) i “Narnia” (C. S. Lewis, ocena 10, {fantastika, avantura, school, akcija}).
Tok odlučivanja:

1. Filtriranje knjiga na osnovu uzrasta, ostaju knjige za tinejdžere i young adults
2. Povećanje skora knjigama sa target žanrovima po sledećoj formuli
   +0.5 svim knjigama koje sadrže žanrove fantastike, avanture, akcije, school, grčka mitologija. A potom povećanje skora svim knjigama koje istovremeno sadrže ove žanrove i to +0.1*n gde je n broj željenih žanrova koje knjiga istovremeno sadrži.
   Primer: ako knjiga ima žanrove akcija i avantura njen skor će biti 0.5+0.5+0.2
3. Povećanje skora knjigama koje je napisao pisac čije je knjige čitalac već pročitao i to po sledećoj formuli
   broj_pročitnih_knjiga_od_pisca*(0.01*prosečna_korisnikova_ocena_pisca)
   Primer: “The tapestry” serijal ima 5 knjiga i ovo su jedine knjige Henry H. Neff-a koje je korisnik pročitao, ostalim knjigama ovog pisca bi se dodelio skor (5*0.01*8 = 0.4).
4. Recimo da imamo još 2 korisnika uzrasta 15 i 17 koji su pročitali iste serijale kao subjekat plus “Harry Potter” serijal. Ova 2 korisnika su ocenili ove serijale na sledeći način PJ (5,8), TTS (9,8), Narnia (7,9). Sistem dodeljuje skor svakoj drugoj knjizi koje su ova 2 korisnika pročitali i to na sledeći način.
   (broj_zajedničkih_knjiga/100)*(1/prosek_razlika_ocena_subjekta_korisnika(1))
   Primer: prosek_razlika za prvog korisnika je 2.66, a za drugog 0.66.
   Za prvog korisnika bi išlo ((18/100)*(1/2.66)=0.067).
   Za drugog korisnika ((18/100)*(1/1) = 0.18).
   Ovi skorovi bi se dodali drugim knjigama koje su ova 2 korisnika pročitali u ovom slučaju knjige serijala “Harry Potter”.
5. Za parametre: dužina knjige, nagrade koje je knjiga osvojila, da li je bazirana na istinitom događaju, pripadnost određenom književnom stilu, postojanje ekranizacije, sistem dodeljuje skor svakoj knjizi po sledećoj formuli
   0.01*broj_pročtanih_knjiga_posmatranog_parametra (while n<10)+
   0.005*broj_pročtanih_knjiga_posmatranog_parametra(while n>10 && n<50) +
   0.001*broj_pročtanih_knjiga_posmatranog_parametra(while n > 50)
   Primer: Korisni je pročitao 18 knjiga srednje dužine (100-500 stranica) te sve knjige ove dužine dobijaju skor 10*0.01 + 8*0.005 = 0.14

6. Recimo da je prethodno ocenjivanje rezultovalo da najbolju ocenu ima knjiga “Harry Potter and the Deathly Hallows”. Sistem čitaocu preporučuje “Harry Potter and the sorcerer's stone” zato što je to prva knjiga serijala koju nije čitao. Takođe ako se još neka “Harry Potter” knjiga nađe u top 5, sistem će je izbaciti i ubaciti prvu narednu po skoru.



Slika 1. Tok izvršavanja pravila

Primer pravila 2 (izveštaj najpopularnijih knjiga u biblioteci):

Kada admin zatraži izveštaj za najpopularnije knjige okida se lanac pravila koji vraća 5 najpopularnije rangiranih knjiga na nivou cele biblioteke. Na prvom nivou rezonovanja određujemo pripadnost knjige po dužini na osnovu broja stranica i generišemo ocenu svakog od tipova. Na ovom nivou takođe računamo prosečnu ocenu za serijalizovane knjige i računamo generalnu prosečnu ocenu knjiga kao i prosečan broj pregleda. Na drugom nivou rezonovanja sistem pronalazimo koeficijent za svaki tip dužine knjige, za serijalizacije knjige i za favorit knjige. Na ovom nivou rezonovanje takođe dajemo koeficijent za broj pregleda na osnovu odnosa broja pregleda i prosečnog broja pregleda. Na trećem nivou rezonovanja generišemo osnovu ocenu knjige kao sumu prethodno određenih koeficijenata i proizvoda prosečne ocene knjige sa količnikom broja recenzija i broja pregleda. Na četvrtom nivou recenzije određujemo ocenu žanra na osnovu ocena knjiga sa trećeg nivoa rezonovanja, pri tom veću će ocenu dobijati od knjiga koje imaju manje žanrova, na ovom nivou ćemo takođe izdvojiti za koliko čitaoca je pisac omiljen kao i ocenu knjiga koje pisac napisao. Na petom nivou rezonovanja dajemo ocenu piscu na osnovu ocena knjiga koje je napisao i broja čitalaca za koje je on omiljen. Na šestom nivou rezonovanja dobijamo konačnu ocenu za knjigu spojem izvedene

ocene knjige sa ocenom žanrova i ocenom za pisca. Na sedmom nivou rezonovanja ćemo na osnovu dobijen konačne ocene proslediti adminu top 5 knjiga.
Tipovi knjiga po dužini:

1. Kratke do 100 stranica
2. Srednje od 100 do 500 stranica
3. Dugačke od 500 stranica na više

Primer: Prilikom traženja najboljih 5 knjiga za izveštaj svaka knjiga će dobijati pojedinačnu ocenu i recimo da posmatramo “Lord of the Rings The Return of the King” koji ima 347 strana on bi spadao u knjige srednje dužine. U ovom slučaju nas interesuje kakve je zainteresovanost za čitanje knjiga srednje dužine, ovo dobijamo posmatranjem karekteristika svih knjiga u biblioteci.

Ulaz: Sve knjige iz sistema. Uzimamo redom knjige kako bi davali ocene.

1. Određujemo tipove knjiga na osnovu broja stranica, zatim kreiramo ocene za sva 3 tipa knjige kao sumu proizvoda prosecna ocene knjige i broja pregleda svih knjiga koje odgovaraju datom tipu. Istu proceduru radimo kako bi smo generisali ocenu za knjige koje pripadaju serijalima. Za serijale neće biti potreban korak za određivanje kategorija pošto se to nalaziti u samom objektu knjige, računica ide po istoj formuli kao i za ocenu tipova dužine, ali ovde imamo samo 1 ocenu za to da li je knjiga deo serijala. Takođe na ovom nivou računamo generalnu prosečnu ocenu i prosečan broj pregleda svih knjiga.
2.
    a. Sad kad svaki od tipova dužine ima ocenu dajemo im posebni koeficijent gde će dužina sa najboljom ocenom dobiti 0.5, sa srednjom 0 i ona sa najgorom ocenom -0.5. Slično kako imamo ocenu i za to da li je knjiga deo serijala ili ne i za to ćemo dati koeficijent gde ako knjige nije deo serijala on je 0, a ako jeste proveravamo da li je ocena za knjigu koje su deo serijala veći od proizvoda generalne prosečne ocene i prosečnog broja pregleda svih knjiga i ako jeste dajemo koeficijent 0.5, a ako nisu onda je -0.5.
    b. Kako bi smo izračunali ocenu za knjigu takođe će nam biti potrebno da odredimo koeficijent za broj pregleda knjige. Ovo postižemo tako što ćemo posmatrati da li je broj pregleda veći od prosećnog broja pregleda koji smo prethodno izračunali. Iz ovoga dobijamo 3 pravila koji se dešavaju u sledećem redosledu. Ako je broj pregleda 2 puta veći od prosečnog broja pregleda onda je koeficijent 2, ako je samo veći, ali ne i 2 puta veći onda je 1 i ako je broj pregleda manji od prosečnog broja onda je 0.
    c. Zelimo da dajemo prednost knjigama koje su favorit velikom broju čitalaca. Ovo postižemo ako knjiga ima broj favorit koji iznosi bar ⅓ broja pregleda i ako je broj pregleda veći od prosečnog broja pregleda ili ako je broj favorita iznosi bar ⅔ broja pregleda kad je broj pregleda manji od prosečnog, u ovim slučajevima koeficijent za favorite dobija

vrednost 1, dok ako nijedan od ova dva slučajeva nije prošao onda je koeficijent 0.

3. Sad generišemo osnovnu ocenu knjige po formuli:
   ocena_knjige * (broj_recenzije / broj_pregleda) + koeficijent_favorita + koeficijent_broja_pregleda+ koeficijent_tipa_dužine + koeficijent_serijal + 1
   Dodajemo +1 na kraju formule kako bi opseg ocene ostao 0-10.
4.
   a. Sad kad imamo osnovnu ocenu knjige računamo ocenu žanrova na osnovu ocene knjiga koje imaju taj žanr po formuli:
   sum(osnovna_ocena_knjige / broj_žanrova_knjige)
   Sortiranjem žanrova po ovoj oceni dobijamo kakva je popularnost žanra.
   b. Za konačnu ocenu knjige želeli bismo da uključimo i ocenu za pisca, ali pre nego što ocenimo pisca prvo želimo da damo koeficijent za pisce koji su favoriti za bar 10% čitalaca u biblioteci. Koeficijent će imati vrednost 0 ili 1 u zavisnosti od toga da li je broj favorita veći od 10% ili ne.
   c. Takođe da bi smo ocenili pisca moramo da nadjemo ukupnu ocenu za knjige koje su napisane od strane tog pisca ovo dobijamo po formuli suma(osnovna_ocena_knjige) / broj_napisanih_knjiga
5.
   a. Sad kad smo pripremili sve podatke dajemo ocenu i za pisce po formuli:
   0.9 * ocena_napisanih_knjiga + koeficijent_favorit_pisac
   b. Kako smo prethodno našli ocenu zanrova i iz toga mogli da pronađemo koja im je prosećna ocena, sad ćemo uz pomoć te 2 vrednosti da genrišemo i ocenu žanr za konkretne knjige. Pravilo glasi da ako je ocena žanra veća od prosećne onda se dodaje +1 u suprotnom ništa, ovo se ponavlja za sve žanrove koji se nalaze u knjizi.
6. Konačnu ocenu knjige se dobija po formuli:
   0.8 * osnovna_ocena_knjige + 0.1 * ocena_pisca + (ocena_žanrova_knjige) / broj_žanrova_knjige

Konačna ocena knjige će se čuvati u okviru modela knjige kao sistemska ocena. Sortiranje svih knjiga po sistemskoj oceni će nam dati 5 najboljih.

Poređenje sa github-om

Za ocenu 5: Vecina pravila će imati neki uslov za njeno izvršavanje te će samim tim biti sačinjeno od when i then dela.
Za ocenu 6: U primeru pravila 2 stavka 2.b imamo pravila za davanje koeficijenta za broj pregleda knjige ideja je da se samo 1 prvailo za zadavanje koeficijenata okida na osnovu broja pregleda i prosečnog broja pregleda. Na primer ako je broj pregleda

manji od prosečnog broja pregleda okidamo pravilo koje postavlja koeficijent na 0 i ostala pravila se neće okinuti.
Za ocenu 7 i 8: Pravilo 2 stavka 5.b - Prilikom računanja ocene žanra za knjigu postavićemo redosled okidanja pravila na osnovu ocene konkretnog žanra, a accumulate imamo tako što ćemo proveravati da li se žanr nalazi među žanrovima knjige, logika će biti kompleksna u smislu da nećemo ručno ništa navoditi nego ćemo samo proveravati postojanje žanra u listi žanrova knjige.
Za ocenu 9:
1) U primeru pravila 1 pod stavkom 2, ocene se dodeljuju na osnovu žanrova koje knjiga poseduje, a potom se dodeljuje dodatni bonus ako knjiga sadrži više žanrova istovremeno.
2) U primeru pravila 1 pod stavkom 6, ako se zaključi da je knjiga deo serijala gleda se da li je prva u serijalu. Ako nije, query-jem se dobavlja poslednja knjiga tog serijala koju je korisnik pročitao i potom se postavlja naredna.
Za ocenu 10: U primeru pravila 1 pod stavkom 4, query-ima će biti dobavaljane knjige koje su i subjekat i drugi korisnik pročitali kao i knjige koje je korisnik pročitao, a subjekat nije.

Literatura

Online Book Recommendation System using Collaborative Filtering (With Jaccard Similarity)- https://iopscience.iop.org/article/10.1088/1742-6596/1362/1/012130/meta