**K-najbliži sused algoritam u društvenim mrežama
K-nearest neighbor algorithm in social networks**

Danijela Stepanović

***Elektrotehnički fakultet, Univerzitet u Beogradu***

**Sadržaj –** U ovom radu je opisano kako se k-najbliži sused algoritam može upotrebiti u društvenim mrežama za preporuku tagova.

**Abstract -** This paper describes how k-nearest neighbor algorithm can be used in social networks for tag recommendation**.**

1. uvod

K-najbliži sused algoritam (skraćeno kNN algoritam) spada u jedan od deset najkorišćenijih algoritama za pronalaženje skrivenog znanja [1].

Društvena mreža je društvena struktura sastavljena od pojedinaca koji se nazivaju “čvorovi“. Čvorovi su povezani sa jednim ili više specifičnih tipova međuzavisnosti, kao što su prijateljstvo, srodstvo, vizije, ideje, znanje... Društvene mreže su postale jedan od najčešće korišćenih oblika komunikacije u 21. veku.

Folksonomija ili kolaborativni sistem tagovanja je sistem koji omogućava internet korisniku da deli, obeležava i pretražuje online resurse kroz personalizovane labele, tj. tagove [2]. Nekoliko kolaborativnih sistema tagovanja je steklo ogromnu popularnost i privuklo milione korisnika. Flickr dozvoljava korisnicima da postavljaju i dele slike. YouTube omogućava korisnicima da pretražuju, postavljaju i dele video sadržaje. Postoje i kolaborativni sistemi tagovanja koji su se specijalizovali u muzici, biznis dokumentima, studentskim radovima...

U osnovi kolaborativnih sistema tagovanja je: korisnik tagovima opisuje resurs. Tag može biti deskriptivan, subjektivan, organizacioni, ali može i da nema značenje. Korisnik upotrebom tagova omogućava sistemu da zaključi interese korisnika. Tag omogućava korisniku da organizuje web resurse za kasniju upotrebu. Resursi lako mogu biti sortirani, skupljeni ili preuzeti. Resursi mogu biti označeni sa više tagova. Na taj način olakšavamo korisniku pronalaženje odgovarajućeg resursa. Jedna od najčešće korišćenih vrsta tagovanja je #hashtag [3].

Tagovi predstavljaju jedan od najvećih izazova za algoritme preporuke. Većina kolaborativnih sistema tagovanja dozvoljava tagovanje bez nadzora. Korisnici su slobodni da koriste bilo koji tag da bi opisali resurs. Tagovanje bez nadzora može da dovede do suvišnosti kada više tagova ima isto značenje ili dvosmislenosti kada jedan tag ima više značenja.

Preporukom tagova smanjujemo posao korisnika, jer više nije potrebno da korisnik koristi tastaturu. Dovoljan je samo jedan klik miša. Ovako ohrabrujemo korisnika da češće koristi tagove, stavlja više tagova, ponovo koristi tagove i koristi tagove kojih se sam ne bi setio. Na ovaj način sprečavamo pravopisne greške i samim tim smanjujemo šum podataka. Sprečavamo i korišćenje tagova sa istim značenjem i tagova koji imaju više značenja.

Uvođenjem tagova dovodimo do toga da Folksonomija koristi trodimenzionalnu strukturu podataka. U radu je prikazano kako je moguće modifikovati k-najbliži sused algoritam tako da se može koristi prilikom preporuke tagova. Opisali smo i kako se informacije o korisniku i o resursu mogu primeniti u algoritmu i tako poboljšati preciznost, a smanjiti računarski troškovi.

Ostatak rada je organizovan na sledeći način. Prvo su prikazana postojeća rešenja problema preporuke tagova u Folksonomiji. Istražena je struktura podataka u Folksonomiji i prikazan jedan način redukcije podataka korišćenjem p-core processing metode. Potom je opisan osnovni pristup preporuke tagova u Folksonomiji primenom modifikovanog kNN algoritma.

1. Postojeća rešenja

Kako je popularnost kolaborativnog sistema tagovanja rasla, istraživači su sve više bili zaiteresovani za istraživanje fenomena tagovanja. Za razliku od standardnih algoritama preporuke, koji imaju dvodimenzionalne veze između korisnika i resursa, algoritmi preporuke tagova imaju trodimenzionalne veze između korisnika, resursa i tagova. Istraživanja su pokazala da trenutno većina postojećih sistema koristi samo jednu stavku za preporuku tagova, ili informaciju o korisniku ili informaciju o resursu.

Sistem koji automatski predlaže tagove za Youtube video resurse je baziran jedino na audiovizuelnom sadržaju resursa [4]. Svi tagovi naučeni u toku treninga se čuvaju u memoriji. Prilikom postavljanja novog videa, resurs se mapira u labele tako što se upoređuje sa svim video sadržajima koji postoje na sajtu. Na osnovu otkrivenih labela resursu se dodeljuju tagovi iz memorije.

Sistem koji automatski predlaže tagove za Instagram slike koristi jedino tagove prethodno upotrebljene od strane korisnika. Na ovaj način sistem nema mogućnost učenja novih reči.

Twitter sistem za automatsko predlaganje tagova je baziran na najčešće korišćenim tagovima drugih korisnika [5]. Tagovi zajedno sa ocenom popularnosti se čuvaju u memoriji.

Flickr sistem za automatsko predlaganje tagova je baziran na kolektivnom znanju [6]. Korisnici tagovima opisuju resurs i obezbedjuju kontekstualne i semantičke informacije. Sistem upoređuje opisane resurse i na osnovu toga predlaže tagove.

Predloženo rešenje nadmašuje postojeća rešenja, jer tagovi zavise i od resursa i od korisničkog profila.

1. Model POdataka u Folksonomiji

U ovoj sekciji definisaćemo trodimenzionalnu strukturu podataka koja se koristi u Folksonomiji i objasniti kako je moguće zameniti je sa dvodimenzionalnim projekcijama. Potom ćemo istražiti tehniku redukovanja skupa podataka pomoću p-core processing metode koja smanjuje količinu podataka uklanjajući korisnike, resurse i tagove koji se pojavljuju manje od predefinisanog broja puta.

3.1 Model podataka

Struktura podataka u Folksonomiji se razlikuje od standardne strukture podataka koja se koristi u algoritmima preporuke. Folksonomija se može opisati preko:

 𝐷=⟨𝑈,𝑅,𝑇,𝐴⟩ (1)

pri čemu, U predstavlja skup korisnika, R skup resursa, T skup tagova, a A skup anotacija koje su predstavljene kao korisnik-resurs-tag tripleti.

 A⊆{⟨u,r,t⟩ : u∈U, r∈R, t∈T} (2)

Folksonomija se može posmatrati kao trojni hibridni graf kod koga su korisnici, tagovi i resursi čvorovi i temena trougla, dok anotacije predstavljaju veze između korisnika, tagova i resursa [7].

Slika 1. Grafički prikaz Folksonomije

Dimenzionalnost možemo smanjiti tako što ćemo umesto resursa i taga koristiti njihovu vezu. Težinu relacije između resursa i taga računamo kao broj korisnika koji su postavili određeni tag na resurs. Na ovaj način smo izgubili informacije, ali smo dobili dvodimenzionalnu matricu koja se lako primenjuje na k-najbliži sused algoritam.

3.2 Uklanjanje irelevantnih podataka

Koristeći p-core processing metodu korisnici, resursi i tagovi se uklanjaju iz skupa podataka tako da ostatak skupa podataka garantuje da svaki korisnik, resurs i tag se pojavljuje u najmanje p postova. Svaki post sadrži informacije o korisniku, resursu i tagovima koje je korisnik postavio na resurs.

Algoritam iterira kroz postove i broji pojavljivanja svakog korisnika, resursa i taga. Ako broj pojavljivanja ovih stavki ne zadovoljava neophodnu definisanu vrednost p, svi postovi koji sadrže stavku se uklanjaju iz skupa podataka. Kako uklanjanje posta na osnovu pojavljivanja neke stavke utiče na broj pojavljivanja ostalih stavki tog posta potrebno je više puta iterirati kroz skup podataka. Rezultat je manji skup podataka.

Postoji nekoliko razloga za korišćenje p-core processing metode. Uklanjanjem neproduktivnih korisnika i nepopularnih tagova i resursa smanjujemo šum podataka. Uklanjanjem retkih stavki, veličina podataka može drastično da se smanji. Ovako omogućavamo primenu data mining tehnika koje bi inače bile računarski nemoguće.

1. Preporuka tagova u folksonomiji

Preporuka tagova u Folksonomiji smanjuje napore korisnika jer više nije potrebno da korisnik koristi tastaturu. Dovoljan je i samo jedan klik miša. Na ovaj način ohrabrujemo korisnika da taguje više resursa, stavlja više tagova, ponovo koristi stare tagove i možda koristi tagove kojih se sam ne bi setio. Eliminisanjem unosa tagova putem tastature smanjujemo pojavljivanje pravopisnih i interpunkcijskih grešaka, kao i ponavljanje tagova koji dovode do šuma u podacima.

Resursi su najčešće tagovani zbog kasnije upotrebe. Tagovanje resursa na pametan način olakšava proces njihovog kasnijeg pronalaženja. Kada su resursi tagovani tako da tag predstavlja opis resursa, preporuka taga može dovesti do većeg zadovoljstva korisnika sajtom. Ovakav pristup je od izuzetnog značaja u slučaju kada resursi ne mogu biti lako procenjeni od stane računara.

**4.1 Osnovni algoritmi preporuke**

Tradicionalni algoritmi preporuke kao ulaz dobijaju korisnika *u*, dok izlaz predstavlja skup stavki *I.* Algoritam preporuke tagova u Folksonomiji kao ulaz dobija i informacije o korisniku i informacije o resursu. Izlaz i dalje predstavlja skup stavki, pri čemu u ovom slučaju stavke su tagovi. Jedna od poteškoća sa kojom se susreće algoritam preporuke tagova je kako iskoristiti i informacije o korisniku i o resursu. Najjednostavnija strategija preporuke tagova jeste predložiti najčešće korišćene tagove od stane korisnika. Međutim ovakva strategija ignoriše informacije i o korisniku i o resursu.

Algoritam za preporuku tagova u Folksonomiji zahteva korišćenje informacija o korisniku i resursu, tako da preporučeni skup sadrži tagove koji su relevantni za resurs i ujedno predstavljaju korisnikove navike prilikom tagovanja.

**4.2 K-najbliži sused algoritam**

K-najbliži sused algoritam predstavlja instance-based learning algoritam tj. sve aproksimacije funkcija se vrše lokalno i sva računanja se odlažu do klasifikacije [8]. K-najbilži sused algoritam se koristi kada imamo veliku količinu redudantnih podataka i spada u jedan od najprostijih algoritama.

Koristi se za:

1. klasifikaciju
2. regresiju

Prilikom klasifikacije ulaz algoritma se sastoji od k najbližih primera naučenih tokom treninga algoritma, dok izlaz predstavlja kojoj klasi pripada objekat. Objekat se klasifikuje po tome koja klasa je najzastupljenija u skupu od k najbližih suseda. Prilikom klasifikacije koristi se supervised learning metoda tj. klasama se dodeljuju metrike ili labele radi lakše implementacije algoritma.

Prilikom regresije ulaz algoritma se sastoji od k najbližih primera naučenih tokom treninga algoritma, dok izlaz predstavlja vrednost objekta. Vrednost se dobija kao srednja vrednost najbližih suseda.

K-najbliži sused algoritam se može modifikovati tako da koristi informacije i o korisniku i o resursu. Tradicionalno algoritam nalazi skup korisnika sličnih pretraživanom korisniku.

Možemo modifikovati algoritam tako da ignoriše korisnike koji nisu tagovali pretraživani resurs. Kada utvrdimo koji su slični korisnici, algoritam razmatra samo tagove koji su postavljeni na pretraživani resurs od strane sličnih korisnika i računa težine tih tagova.

**Input:** , query korisnik; *,* query resurs;

 ***k***, broj suseda; ***n***, broj predloženih tagova

**Outut:** , skup predloženih tagova

**foreach**  *koji je pribeležio* **do**

= similarity*(u,)*

**end**

*N* je skup k najbližih suseda od

**foreach**  **do**

 **foreach** *t koje je u postavio na* **do**

 **end**

**end**

Sortiraj tagove po vrednostima

 je skup gornjih n tagova

**return**

**4.3 User model**

Model korisnika omogućava izračunavanje sličnosti između korisnika koja je potrebna za k-najbliži sused algoritam. Svaki korisnik se može modelirati kao vektor koji zavisi od težina tagova. Težina taga predstavlja važnost nekog taga.

 (3)

Na ovaj način je prikazana povezanost korisnika samo sa tagom. Na sličan način može se modelirati korisnik i preko zavisnosti od resursa.

Najbolje utvrđivanje sličnosti između korisnika se dobija ako modeliramo korisnika preko parova tag-resurs, pri čemu par specificira koji tag je postavljen na resurs.

 (4)

 težina je jednaka 1 ako je korisnik postavio tag na resurs , inače je jednaka 0.

Kada smo korisnike prikazali pomoću vektora, sličnost možemo dobiti kao kosinus ugla koji zaklapaju vektori korisnika.

1. Moguće modifikacije i unapređenja

U standardnom pristupu predlaganja tagova skup predloženih tagova ne bi sadržao tagove koje je korisnik već koristio. Međutim u kolaborativnim sistemima za tagovanje korisnici često ponovo koriste stare tagove. Prethodno korišćeni tag je važan trag za algoritam preporuke tagova. Boosting faktor *b* može da se koristi za promovisanje tagova na korisničkom profilu. Kao dodatni korak modifikovanom k-najbliži sused algoritmu, b se dodaje težini taga ako je korisnik postavio taj tag na neki drugi resurs.

Pseudokod:

**foreach do**

 **if**  **then**

 e**nd**

**end**

1. Zaklučak

U radu je predloženo korišćenje kNN algoritma za preporuku taga u Folksonomiji. Zbog jedinstvene strukture Folksonomije, modifikacije su potrebne da bi se sistem prilagodio algoritmu. Rešenje problema preporuke tagova obezbeđuje da tag zavisi i od resursa i od korisničkog profila. Na ovaj način sugestije tagova su preciznije i samim tim se povećava zadovoljstvo korisnika. Modifikacije korisničkog modela Folksonomije su dovele do smanjenja potrebnih računanja i memorije, jer smo uklonili redudantne podatke. U radu je data i jedna od mogućih modifikacija predloženog rešenja koja uzima u obzir i tagove koje je korisnik već koristio.

Predloženo rešenje prevazilazi trenutna rešenja ovog problema, jer uzima u obzir i korisnikove informacije i resurs. Za razliku od postojećih rešenja predloženo rešenje ne koristi memoriju za čuvanje već korišćenih tagova.

Predloženo rešenje pokazuje izvestan upotrebni potencijal. Naročito u slučaju kada se obrađuje velika količina redudantnih podataka. Takođe, rešenje je primenljivo u slučaju kada su resursi slike, video sadržaji ili muzika tj. resursi su takvi da se sporo obrađuju od strane kompjutera. U tom slučaju tagovi su takvi da predstavljaju sadržaj resursa.

literatura

1. Xindong W. „Top 10 Algorithms in Data Mining“, University of Vermont, USA, 17.11.2014.
2. Wikipedia, Folksonomy, dostupno na:

 <http://en.wikipedia.org/wiki/Folksonomy>.

1. Wikipedia, Hashtag, dostupno na:

 <http://en.wikipedia.org/wiki/Hashtag>.

1. G. Toderici, H.Aradhye, M.Pasca, L.Sbaiz, and J.Yagnik, „Finding meaning on YouTube: Tag recommendation and category discovery“, CVPR, pp 3447-3454, 2010.
2. E.Zangerle, W.Gassler, and G.Specht, „Recommending #-tags in Twitter“, UMAP 2011, pp 67-78, Gerona, Spain, 2011.
3. B. Sigurbjörnsson, and Roelof van Zwol, „Flickr tag recommendation based on collective knowledge“, WWW '08, pp 327-336, New York, USA, 2008.
4. Z. Zhang, and C. Liu, „A hypergraph model of social tagging networks“, Journal of Statistical Mechanics: Theory and Experiment, Vol. 2010, No.10, 2010.
5. Wikipedia, K-nearest neighbors algorithm, dostupno na: <http://en.wikipedia.org/wiki/KNearestNeighborsAlgorithm>.