SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI SEMINAR

USLUGE OBAVJEŠTAVANJA I PREPORUČIVANJA ZASNOVANO NA DRUŠTVENOJ UMREŽENOSTI KORISNIKA

Mia Augustinović

Mentor: Doc. dr. sc. Vedran Podobnik

Zagreb, svibanj 2012.

Sadržaj

Uvod

Sličnost korisnika se može interpretirati na različite načine raznim metodama. Zadatak seminarskog rada jest da se na temelju grupa koje korisnici preferiraju na društvenoj mreži Facebook možemo utvrditi kolika sličnost postoji između korisnika koje uspoređujemo. Na temelju tih sličnosti može se dobiti pouzdana lista korisnika kojima bi se mogle slati preporuke ili obavijesti, a da ih on/ona uzme u obzir.

Glavna ideja je da se dohvatom svih naziva grupa koje korisnici preferiraju stvara rječnik koji se sastoji od skupa riječi koje se pojavljuju u nazivu tih grupa. Potom se za primjer uzmu dva korisnička profila te grupe koje im se sviđaju. Na temelju rječnika mogu se uvidjeti sličnosti u nazivu grupa; ako postoje dvije grupe s istom tematikom (npr. službena stranica nekog filma i grupa koju su napravili *fanovi*), ali nazivi im se malo razlikuju. Za takve grupe se može reći da su iste te se na temelju takvih usporedbi može zaključiti kako ta dva korisnika imaju slične interese.

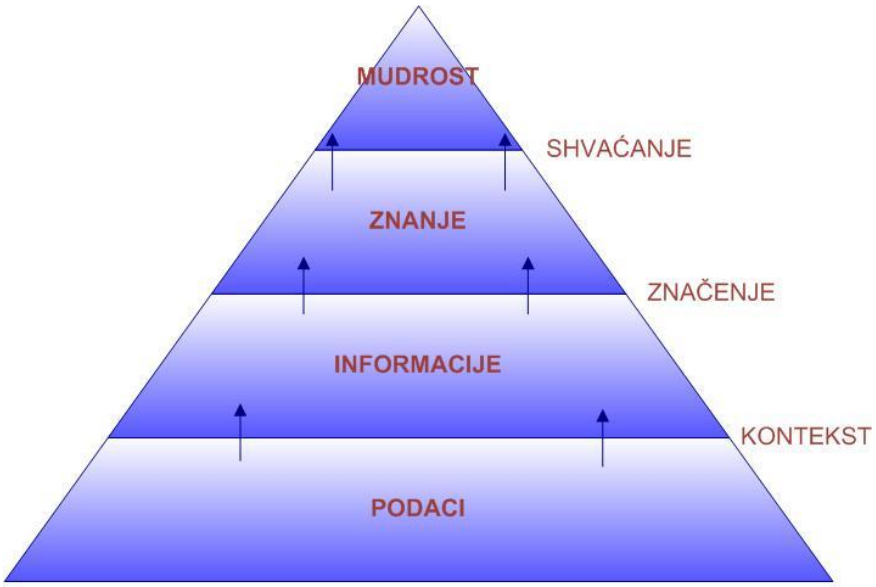
Seminarski rad u prvom poglavlju daje uvod u semantičko pretraživanje podataka te što za korisnika znače pojmovi podaci, informacije, znanje i mudrost. Nadalje, slijedi opis dubinske analize podataka te koji alati za analizu postoje i na koji način se koriste. Drugo poglavlje opisuje metodu preporučivanja zasnovanu na sadržaju; njezinu formalnu definiciju te primjer korištenja i ograničenja sustava. Zadnje poglavlje donosi opis studijskog slučaja koji je napravljen u sklopu rada.

1. Semantičko pretraživanje podataka

Semantičko pretraživanje informacija u tekstualnim dokumentima je zadatak discipline dubinske analize teksta ili rudarenja tekstualnih podataka [1] .

1.1. Podaci, informacije i znanje o korisniku

Ljudsko shvaćanje znanja može se opisati hijerarhijskim ustrojem koje kreće od najsiromašnijeg kontekstnog opisa – *podataka*, prema sve bogatijem – *informacijama, znanju i mudrosti* (engl. *Data, Information, Knowledge, Wisdom*, DIKW hijerarhija). DIKW hijerarhija (Slika 1) kao temelj uzima podatke koji predstavljaju neki jednostavan zapis, npr. brojku dobivenu mjerenjem ili opažanjem. Ova hijerarhijska razina ne sadrži mjernu jedinicu ili bilo kakav drugi kontekstni atribut koji bi dodao šire značenje zapisu. Sljedeća razina hijerarhije su informacije koje podacima dodaju kontekstni atribut ekvivalentan mjernoj jedinici. Sljedeća razina, znanje, donosi odgovor na pitanje kako koristiti informaciju. Najviša razina razumijevanja podataka, mudrost, zasad je svojstvena samo ljudima i pruža odgovor na pitanje kada koristiti informaciju [2] [3] .



Slika 1. DIKW hijerarhija

1.2. Dubinska analiza podataka

Disciplina dubinske analize teksta (engl. *text mining*) sastavni je dio discipline koja se naziva dubinska analiza podataka (engl. *data mining*), a bavi se sadržajno utemeljenom (engl. *content based*) obradom nestrukturiranih tekstualnih dokumenata i izdvajanjem korisne informacije iz njih [4] . Dubinska analiza podataka je računalni način obrade podataka koji podrazumijeva razne postupke koji imaju za cilj dobivanje što korisnijih informacija iz podataka. Sam termin rudarenja podataka može se objasniti kao proces pronalaženja korisnog znanja ili informacija, tj. otkrivanja znanja iz velike količine podataka.

Cilj pretraživanja informacija jest vratiti kao rezultat pretraživanja na postavljen korisnički upit što više dokumenata relevantnih za korisnički upit i pri tome vratiti što manje dokumenata koji nisu relevantni.

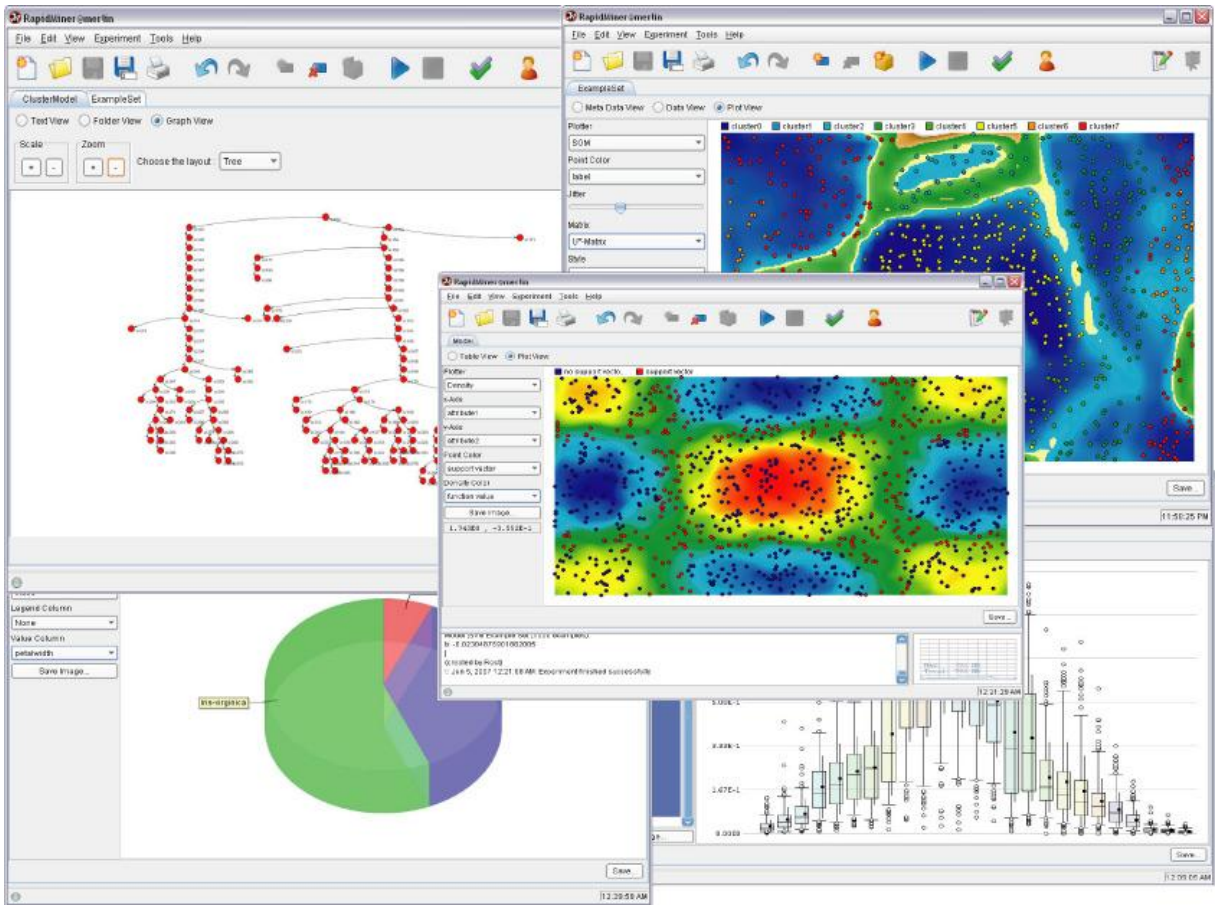
Zahvaljujući najnovijim tehničkim postignućima u procesiranju podataka, povećanom kapacitetu memorija (više spremljenih podataka) i boljoj povezivosti računala, pretraživanje podataka postalo je vrlo važno. Upravo je rudarenje podataka to koje pomaže otkriti važne informacije i znanje utkano u podatke, uvelike pridonoseći donošenju odluka, poslovanju i znanosti.

1.3. Alati za dubinsku analizu podataka

Danas postoje razni alati za dubinsku analizu podataka. Ovdje su obuhvaćena dva alata; RapidMiner [5] i Lucene [7] . U nastavku je dan opis pojedinog alata.

1.3.1. Alat RapidMiner

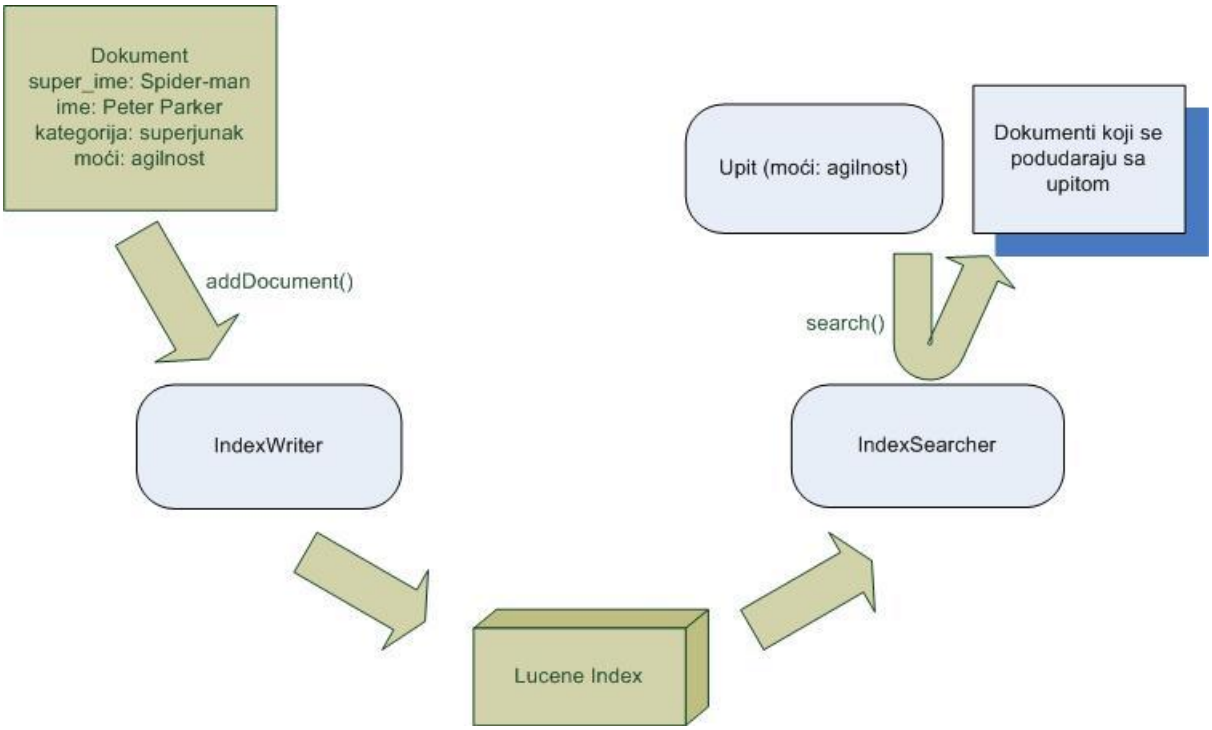
Glavni produkt Rapid-I, RapidMiner je svjetski vodeći sustav otvorenog koda za otkrivanje znanja i dubinsku analizu podataka. RapidMiner predstavlja fleksibilno Java okružje za otkrivanje znanja u bazama podataka, strojnom učenju i rudarenju tekstualnih podataka. To je suvremeni sustav za dubinsku analizu podataka koje se odlikuje kvalitetnim korisničkim sučeljem. Može se koristiti kao zasebna aplikacija ili se može integrirati u već postojeći produkt [6] .



Slika 2. Prikaz sučelja RapidMiner alata [8]

1.3.2. Alat Lucene

Lucene predstavlja Javinu biblioteku za pretraživanje koja omogućuje jednostavno dodavanje tražilice u bilo koju aplikaciju. Posljednjih godina Lucene je postao vrlo popularna i najkorištenija knjižnica za pretraživanje informacija. Jedan od razloga zbog čega je Lucene tako popularan je njegova jednostavnost. Nisu potrebna velika znanja o tome kako Luceneovo indeksiranje informacija i njihovo pronalaženje radi kako bi se započelo s radom s njim. Lucene se iznenađujuće koristi na mnogo različitih mjesta kao što su NetFlix, MySpace, LinkedIn, Fedex, Apple [7] [9] . Slika (Slika 3) prikazuje način korištenja alata Lucene. Kao ulaz koristi se dokument s raznim podacima, koji se dodaju u indeks riječi. Nakon što se indeks popunio s podacima, nad njim se izvršava određeni upit koji daje kao rezultat dokumente koji se podudaraju s upitom.



Slika 3. Primjer uporabe alata Lucene

2. Metoda preporučivanja zasnovana na sadržaju

Sadržajno utemeljena obrada podrazumijeva obradu tekstualnih dokumenata isključivo na osnovi sadržaja dokumenata, a ne metapodataka.

U sustavima za preporučivanje koji se temelje na sadržajnom preporučivanju (engl. *content-based recommendation*), korisnost $u(c, s)$ od stavke s za korisnika c je procijenjena na temelju korisnosti $u(c, s_i)$ koje je korisnik c pridijelio stavkama $s_i \in S$ koje su slične stavci s.

Sadržajno preporučivanje ima korijene u pretraživanje i filtriranju informacija. Upravo zbog razvijenosti navedenih sustava, mnogi današnji sustavi koji se temelje na sadržajnom preporučivanju navedenih sustava fokusiraju se na preporučivanje stavaka koje sadrže tekstualnu informaciju, poput dokumenata ili web-stranica.

Poboljšanje u odnosu na sustave za pretraživanje informacija očituje se u korištenju osobnih profila korisnika koji sadrže strukturirane informacije o korisnikovu ukusu, potrebama i preferencijama. Informacije koje služe za profiliranje korisnika mogu se sakupljati eksplicitno, tj. putem upitnika, ili implicitno – učenjem kroz korisničke transakcije u nekom vremenskom periodu.

2.1. Formalna definicija

Neka je $Content(s)$ profil stavke, tj. skup atributa koji karakteriziraju stavku s. Profil je uobičajeno izračunat izdvajanjem skupa značajki i koristi se za računanje prikladnosti stavke za svrhe preporučivanja. Kako su sustavi sa sadržajnim preporučivanjem najčešće dizajnirani za preporučivanje tekstualnih stavki, sadržaj u ovim sustavima najčešće je opisan ključnim riječima. Važnost riječi k_j u dokumentu d_j određena je težinskom mjerom w_{ij}, koja se može definirati na nekoliko načina. Jedna od najpoznatijih mjera za specificiranje težine ključnih riječi jest TF-IDF (*term frequency/inverse document frequency*) mjera koja je definirana na sljedeći način: pretpostavimo da je N ukupan broj dokumenata koji se mogu preporučiti korisniku i da se ključna riječ k_j pojavljuje u n_i od njih. Nadalje, pretpostavimo da je $f_{i,j}$ broj pojavljivanja ključne riječi k_i u dokumentu d_j. Tada je $TF_{i,j}$, učestalost ključne riječi k_i u dokumentu d_j, definirana kao:

$$TF_{i,j} = \frac{f_{i,j}}{max_z f_{z,j}} \tag{1}$$

gdje je maksimum izračunat nad učestalostima $f_{z,j}$ svih ključnih riječi k_z koje se pojavljuju u dokumentu d_j. Međutim, ključne riječi koje se pojavljuju u mnogo dokumenata nisu korisne u razlučivanju između relevatnog i nerelevantnog dokumenta. Stoga, često se koristi mjera IDF_i (*inverse document frequency*) u kombinaciji sa $TF_{i,j}$ (*term frequency*). Mjera IDF_i definirana je kao:

$$IDF_i = \log \frac{N}{n_i}. \tag{2}$$

Tada, TF-IDF težina ključne riječi k_i u dokumentu d_j je definirana kao:

$$w_{i,j} = TF_{i,j} \times IDF_i \tag{3}$$

i sadržaj dokumenta d_j je definiran kao:

$$Content(d_j) = \left(w_{i,j}, \dots, w_{k,j}\right). \tag{4}$$

Neka je *ContentBasedProfile(c)* profil korisnika c koji sadrži ukuse i preferencije tog korisnika. Ti profili su dobiveni analiziranjem sadržaja stavki koje su prethodno viđene i ocijenjene. *ContentBasedProfile(c)* može biti definiran kao težinski vektor $(w_{c1}, \dots, w_{ck})$, u kojem svaka težina w_{ci} označava važnost ključne riječi k_i za korisnika *c*.

U sustavima za sadržajno preporučivanje, funkcija korisnosti $u(c, s)$ definirana je kao:

$$u(c, s) = score\left(ContentBased\ Profile(c), Content(s)\right). \tag{5}$$

U drugim literaturama, funkcija korisnosti $u(c, s)$ može biti definirana pomoću kosinusne funkcije (vektorski prostor modela):

$$u(c, s) = \cos(\vec{w}_c, \vec{w}_s) = \frac{\vec{w}_c \cdot \vec{w}_s}{\| \vec{w}_c \|_2 \times \| \vec{w}_s \|_2} = \frac{\sum_{i=1}^{K} w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^{K} w_{i,c}^2} \sqrt{\sum_{i=1}^{K} w_{i,s}^2}}, \tag{6}$$

gdje je *K* ukupan broj ključnih riječi u sustavu [10] .

2.2. Primjer

Kao primjer može se uzeti sljedeće: ako korisnik c čita mnogo *online* članaka na temu bioinformatike, sustav za sadržajno preporučivanje će uspješno korisniku preporučiti sve ostale članke vezane uz tu temu. Razlog toga je podudarnost sadržaja članka, koji će u tom slučaju sadržavati mnogo pojmova veznih uz bioinformatiku (npr. genom, sekvence). Prema tome, preporučiteljski sustav će koristeći kosinusnu funkciju ili povezane slične mjere dodijeliti veću korisnost $u(c, s)$ onim člancima s koji imaju visokoponderirane bioinformatičke pojmove u $\vec{w}_s$ i nisku korisnost tamo gdje bioinformatički pojmovi imaju manju težinu (značenje).

2.3. Ograničenja

Sadržajno preporučivanje je metoda koju je vrlo jednostavno implementirati, ali posjeduje određena ograničenja:

- Ograničena analiza sadržaja (engl. *limited content analysis*)

  Tehnike za sadržajno preporučivanje su ograničene značajkama objekata koji se preporučuju. Kako bi se raspolagalo dovoljnom količinom značajki, sadržaj mora biti u formi koja se može automatski parsirati na računalu ili sve značajke moraju biti u tekstualnom obliku. Dok opisana tehnika dobro radi u otkrivanju značajki iz tekstualnih dokumenata, neke druge domene predstavljaju svojstven problem. Na primjer, metodu automatskog izvlačenja značajki je teže primjeniti na multimedijske podatke, poput slika, video i audio zapisa.

  Dodatni problem vezan uz ograničenu analizu sadržaja jest taj da ukoliko su dvije različite stavke prestavljene istim skupom značajki, njih nije moguće razlučiti kao zasebne. Stoga sustavi temeljeni na sadržajnom preporučivanju ne mogu napraviti razliku između dobro i loše napisanog tekstualnog dokumenta ako oni koriste iste izraze.

- Prevelika specijalizacija (engl. *overspecialization*)

  Ako sustav može preporučivati samo stavke koje su veoma slične značajkama korisnikovog profila, korisniku će se uvijek preporučivati samo stavke koje su međusobno veoma slične. Na taj način korisnik neće nikad moći izaći iz kruga poznatih stavki i nikada mu se neće preporučiti nešto nasumično. Raznolikost

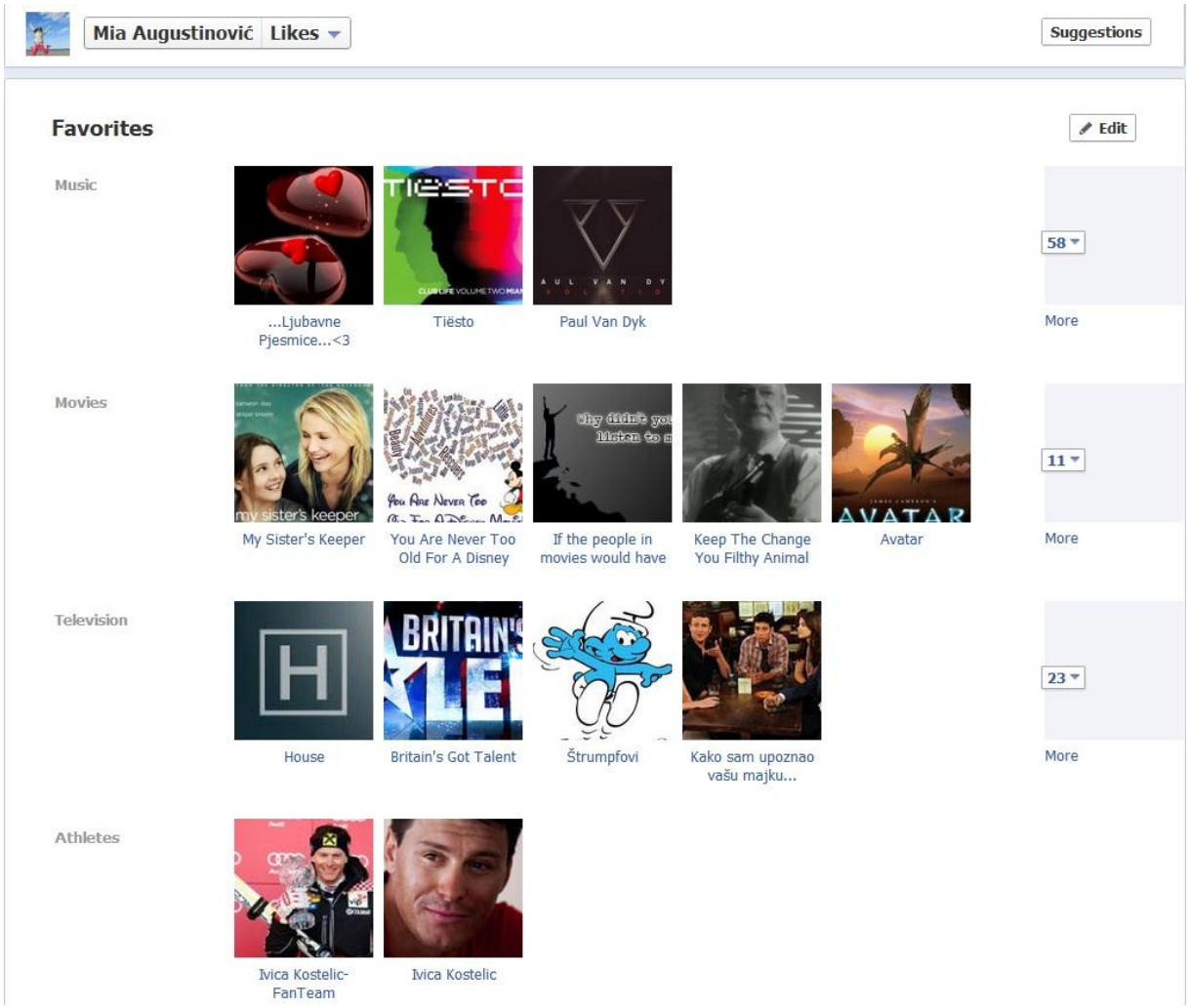
preporuka je vrlo poželjna karakteristika sustava za preporučivanje. U idealnom slučaju, korisniku bi se trebao priakzati spektar opcija, a ne homogena skupina stavki.

- Problem novog korisnika (engl. *new user problem*)

  Korisnik mora ocijeniti dovoljan broj stavki kako bi sustav za preporučivanje temeljen na sadržaju bio u mogućnosti razumijeti korisnikove preferencije i prezentirati korisniku vjerodostojne preporuke. Stoga, novi korisnik koji je podijelio mali broj ocjena neće moći dobiti relevantne preporuke [10] .

3. Studijski slučaj: analiza sadržaja Facebook profila alatom Lucene

Temeljem aktivnosti korisnika moguće je provesti analizu korisničkih profila na društvenoj mreži Facebook, tj. preferiranjem (engl. *Like*) određenih grupa (Slika 4). Na temelju tih grupe može se napraviti usporedba dva korisnika; može se uvidjeti kolika je sličnost između njih.

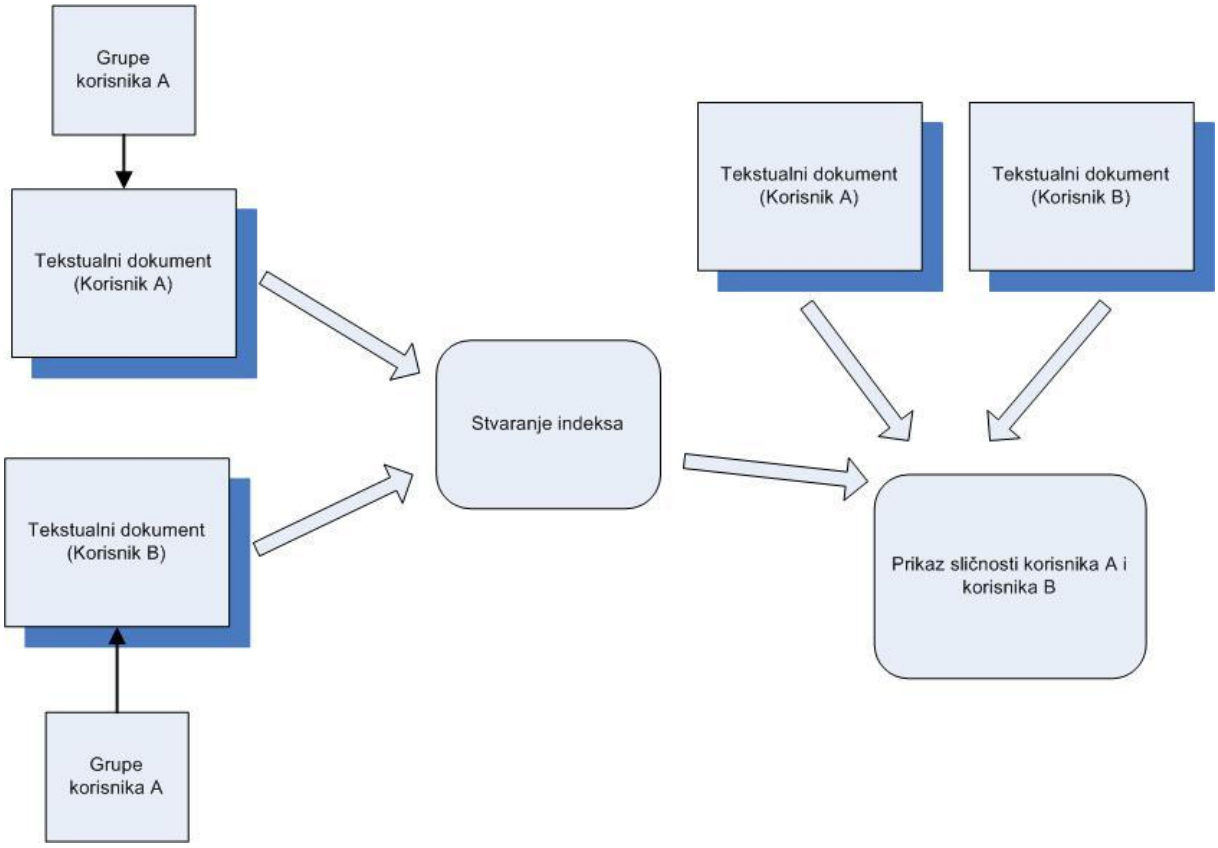


Slika 4. Prikaz dijela profila korisnika sa grupama koje mu se sviđaju

Na *Facebook Developers* [11] stranicama nalaze se upute kako napraviti aplikaciju na društvenoj mreži Facebook [12] [13] . Za izgradnju sustava za usporedbu korisnika te uviđanja njihove sličnosti potrebno je izraditi aplikaciju koja analizira te preuzima podatke s korisničkog profila. Podaci koji se mogu dobiti s nekog korisničkog profila su razni, no u

ovom slučaju, za usporedbu dvaju korisnika, koriste se tzv. *likes*, odnosno grupe koje ti korisnici preferiraju.

Na temelju izvučenih podataka, odnosno imena grupa koje se spremaju u tekstualne datoteke, stvara se rječnik pojmova s riječima iz naziva grupa. Nakon što se indeks popunio riječima, slijedi upit na temelju kojeg se usporede tekstualne datoteke dvaju korisnika koji se promatraju. Na temelju usporedbe tekstualnih datoteka s indeksom dobiva se broj pogodaka, odnosno broj riječi koje se podudaraju, na temelju kojih se računa sličnost ta dva korisnika (Slika 5).



Slika 5. Prikaz načina na koji je implementiran zadatak

4. Zaključak

Seminarski rad daje pregled dubinske analize podataka, objašnjava ljudsko shvaćanje znanja koje je prikazano hijerarhijom od najsiromašnijeg kontekstnog opisa prema sve bogatijem te opisuje alate za dubinsku analizu podataka, RapidMiner i Lucene. Također, opisana metoda preporučivanja zasnovana je na sadržaju; njezina formalna definicija, ograničenja sustava te primjer korištenja.

Uz današnju tehnologiju i raznolike alate moguće je analizirati i uspoređivati korisnike nekog sustava na razne načine. U seminarskom radu prikazan je jedan način usporedbe korisnika društvene mreže Facebook na temelju osobnih preferencija pojedinog korisnika. S obzirom da se društvene stranice zasnivaju na načelu eksplicitne društvene mreže (što znači da su korisnici društvenih stranica definirani svojim (polu)javnim profilima koji sadrže informacije o pripadnosti grupama te poveznicama s drugim korisnicima društvene mreže s kojima je promatrani korisnik povezan), bilo je moguće usporediti dva korisnika na temelju pripadnosti pojedinim grupama, te na kraju dobiti sličnost između njih.

Nastavak istraživanja će biti usmjeren na razradu modela za usporedbu korisnika koja će se temeljiti na vezama s većim značenjem.

Literatura

[1] Jasminka Dobša, predavanje „Semantičko pretraživanje informacija u tekstualnim dokumentima“, http://www.matematika.hr/_download/repository/Dobsa_HMD.pdf

[2] Ackoff, R.L. From Data to Wisdom, Journal of Applied System Analysis, Vol. 16, 1989, 3-9.

[3] Vedran Podobnik, „Višeagentski sustav za pružanje telekomunikacijskih usluga zasnovan na profilima korisnika“, Doktorska disertacija, Sveučilište u Zagreb Fakultet elektrotehnike i računarstva, Zagreb, Hrvatska, 2010.

[4] Jasminka Dobša, „Text mining using concept indexing“, Doktorat, Sveučilište u Zagrebu Fakultet organizacije i informatike, Varaždin, 2006.

[5] RapidMiner, http://rapid-i.com/content/view/181/190/

[6] DataminingTools Inc, „RapidMiner: Introduction to datamining“, http://www.slideshare.net/dataminingtools/rapidminer-introduction-to-datamining

[7] Lucene, http://lucene.apache.org/core/

[8] Machine learning open source software, Project details for RapidMiner, http://mloss.org/software/view/27/

[9] M. McCandless, E. Hatcher, O. Gospodnetic, Lucene in Action. Greenwich: Manning, 2010.

[10] Adomavicius G., Tuzhilin A., Toward the next generation of recommender systems: A survey of the State-of-the-art and possible extensions, IEEE Transactions on knowledge and data engineering, Vol. 17, No. 6, 2005, 734-749.

[11] Facebook Developers, http://developers.facebook.com/docs/guides/web/

[12] Apps for Facebook, https://developers.facebook.com/docs/guides/canvas/

[13] Facebook Graph API, http://developers.facebook.com/docs/reference/api/