SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 1543

AUTOMATSKO INDEKSIRANJE DOKUMENATA U MODELU VEKTORSKOG PROSTORA

Ana Cvitaš

Zagreb, srpanj 2005.

Sadržaj

Kazalo slika

Kazalo tablica

1. Uvod

U današnje se vrijeme poslovne i upravne organizacije, kao i individualni korisnici susreću sa sve većim količinama neobrađenog teksta. Do sada se većina informacija dobivala na temelju numeričkih i drugih jasnih podataka, dok je značenje samog teksta bilo potpuno zanemareno. Osim toga, velike skupine različitih dokumenata obrađivane su uglavnom ručno, što je imalo velike vremenske zahtjeve, a često uopće nije bilo provedivo u realnom vremenu [1], [2].

Iz ovih se razloga u području dubinske analize[1] podataka (eng. *data mining*) razvila grana za obradu tekstualnih podataka nazvana dubinska analiza teksta (eng. *text mining*). U ovo se područje mogu svrstati različiti postupci, kao što su grupiranje dokumenata, kategorizacija, automatsko stvaranje sažetka, kao i automatsko indeksiranje koje se na neki način može smatrati podvrstom kategorizacije [1].

Velika većina metoda za obradu tekstualnih podataka osmišljena je i optimirana ne temelju svojstava engleskog, a tek ponekad i ostalih svjetskih jezika. Hrvatski je u tom pogledu znatno zapostavljen i nije poznato koliko su metode pogodne za engleski, stvarno primjenjive i na hrvatski jezik.

[1] Rudarenje, otkrivanje znanja

2. Automatsko i ručno indeksiranje

U ovom se poglavlju objašnjava pojam indeksiranja i mogućnosti provedbe tog postupka. Opisane su osnovne razlike, prednosti i mane ručnog i automatskog indeksiranja, kao i mogućnost kombiniranja oba postupka.

2.1. Indeksiranje

Indeksiranje se definira kao proces stvaranja reprezentacije dokumenta. Pojam indeksiranja ne označava samo pronalaženje riječi u tekstu i njihovo slaganje prema abecednom redoslijedu, nego uključuje i analizu sadržaja. Osnovna je razlika u tome da se, čitajući tekst, ne koncentriramo samo na to što piše, nego i o čemu piše u tekstu [3].

2.2. Ručno indeksiranje

Još uvijek se većina dokumenata klasificira i indeksira ručno. Grupe dokumentalista čitaju novopristigle dokumente i na temelju njihovog sadržaja određuju prikladne ključne riječi ili ih svrstavaju u ranije definirane klase. Ovaj se pristup može pokazati kao vremenski vrlo zahtjevan, kompliciran i skup, ali ima i određene prednosti. Kod ručnog indeksiranja nikakvu ulogu ne igra jezik dokumenata, specifičnost terminologije ili rječnika. Indeksiranje nije ovisno ni o kakvoj tehnologiji, a može se i lakše prilagoditi potrebama određenih korisnika [4], [5].

2.3. Poluautomatsko indeksiranje

Ručno indeksiranje se može olakšati primjenom različitih automatskih metoda predlaganja ključnih riječi ili pojmova iz tezaurusa. Na taj način indeksatori dobivaju znatno smanjen popis pojmova među kojima odabiru one bitne za dokument. Konačnu odluku donosi ponovno čovjek, ali se vrijeme i cijena indeksiranja znatno smanjuju. Jedan primjer poluautomatskog indeksiranja je CADIS sustav [6].

2.4. Automatsko indeksiranje

Automatsko indeksiranje obavlja se bez ili uz vrlo neznačajnu intervenciju čovjeka. Rezultati koji se dobivaju su podjednako uspješni kao i kod ručnog indeksiranja. Pogreške su neizbježne, jer čak i rezultati ručnog indeksiranja istog skupa dokumenata kad indeksiraju različiti timovi ljudi nisu jednoznačni.

Automatsko indeksiranje znatno štedi vrijeme, a mnogi skupovi dokumenata bez ovakvih metoda ne bi uopće nikada niti bili obrađeni. Nakon što je sustav za automatsko indeksiranje uspostavljen moguće ga je prilagoditi za rad sa različitim zbirkama dokumenata. Mnogi problemi kao što su specifičnost terminologije, višejezičnost ili višeznačnost, koji kod ručnog indeksiranja nisu prisutni, također se mogu riješiti primjenom odgovarajućih metoda [5].

Najbolji se rezultati ostvaruju kombiniranjem ranije spomenutih metoda. Jedan od načina kombiniranja je korištenje sustava za automatsku kategorizaciju dokumenata, koji uči na temelju ručno indeksiranih primjera i nove tekstove klasificira sam usporedbom s naučenim [5].

Algoritmi za automatsko indeksiranje jednaki su onima za klasifikaciju uz uvjet da isti dokument može biti svrstan u nekoliko različitih klasa [7].

3. Vrste automatskog indeksiranja

Osnovna podjela postupaka za automatsko indeksiranje odnosi se na metode ekstrakcije i dodjeljivanja ključnih riječi. Svaka od tih metoda ima svoje prednosti i nedostatke što će biti detaljnije opisano u ovom poglavlju.

Još jedna podjela metoda za automatsko indeksiranje je na lingvističke i statističke metode. Iako se statističke znatno češće primjenjuju, i lingvističke metode mogu doprinijeti uspješnosti postupka.

3.1. Ekstrakcija ključnih riječi (eng. keyword extraction)

Većina postojećih metoda za automatsko indeksiranje koristi kao indeks pojmove iz prirodnog jezika koji se već nalaze u tekstu dokumenta koji se indeksira. Ti pojmovi se odnose na samostalne riječi, ali i složenije fraze. Osnovni problem je u određivanju pojmova koji su dovoljno bitni za sadržaj dokumenta i definiranje njihove važnosti, odnosno težine. Proces koji pomaže u određivanju tih pojmova sastoji se od sljedećih koraka [8]:

1. Leksička analiza, odnosno identifikacija individualnih riječi u tekstu.
2. Eliminacija čestih riječi nebitnih za sadržaj (stop riječi).
3. Morfološka normalizacija ili svođenje riječi na osnovni oblik
4. Opcionalno formiranje fraza
5. Opcionalna zamjena riječi i fraza pomoću tezaurusa
6. Računanje težine za svaki pojam

Najveći utjecaj za dobivanje uspješnijih rezultata imaju eliminacija stop riječi, morfološka normalizacija i formiranje fraza, koji će biti objašnjeni kasnije. Na taj način se eliminiraju previše općeniti ili specifični pojmovi.

Osnovna prednost ovakvih metoda su brzina i fleksibilnost. Izbor pojmova za stvaranje indeksa je znatno veći jer nije ograničen nekim unaprijed definiranim skupom, što omogućuje veću preciznost i više različitih pogleda na bazu podataka. Zbog nepostojanja tog skupa moguć je prijenos i korištenje istog sustava za indeksiranje na različitim skupovima dokumenata, a nije ograničen samo na jedno područje.

Uz opisane prednosti ekstrakcija ključnih riječi ima i određene nedostatke. Ovakve metode nemaju mogućnost razlikovanja višeznačnosti kao što su homonimi, pa se na određeni upit mogu pojaviti dokumenti sadržaja potpuno nevezanog uz traženu temu. Jedan dio tog problema može se riješiti korištenjem ne samo samostalnih riječi nego i fraza. Na taj način, svaka riječ fraze svojim značenjem utječe na ostale i nastala kombinacija riječi više nije višeznačna. Osim zbog navedenog problema, kvalitetno indeksiranje dodatno je otežano zbog velike specifičnosti pojedinih riječi i fraza za značenje dokumenta [7].

3.2. Dodjeljivanje ključnih riječi (eng. keyword assignement)

Kod indeksiranja dodjeljivanjem ključnih riječi one se preuzimaju iz nekog izvora koji nije sam dokument. Njih može odrediti sam indeksator na temelju vlastitog iskustva ili mogu biti preuzete iz nekog kontroliranog rječnika kao što je tezaurus. Dodjeljivanje ključnih riječi iz kontroliranog rječnika temelji se na znanju o tipičnim uzorcima u dokumentu, odnosno pojavljivanju pojedinih riječi ili njihovih kombinacija i njihovoj povezanosti s dodijeljenim pojmovima koji tvore indeks. Pridijeljene riječi nazivaju se deskriptori.

Ovom metodom rješava se problem višeznačnosti, što je posebno značajno prilikom obrade višejezičnih tekstova. Osim toga omogućava se povezivanje i filtriranje dokumenata na temelju pridijeljenih klasa. Važno svojstvo ovakvog indeksiranja je znatno veća općenitost jer se slične značajke pojedinih dokumenata preslikavaju u zajedničke, odnosno opće karakteristike. Kontrolirani rječnik također rješava problem previše općenitih ili specifičnih pojmova koji se koriste u indeksu.

U odnosu na ekstrakciju ključnih riječi znatno je smanjeno svojstvo fleksibilnosti s obzirom na potrebe korisnika. S promjenom kolekcije dokumenata ili potreba pretraživanja potrebno je također obnoviti kontrolirani riječnik [7].

Dodjeljivanje ključnih riječi na temelju kontroliranog rječnika usko je povezano sa pojmom kategorizacije [2] . Kategorizacija se odnosi na svrstavanje dokumenata u dvije ili više klasa ili kategorija. To je nadzirani proces koji zahtijeva učenje na skupu primjera za svaku kategoriju [1]. U slučaju indeksiranja svakom tekstu može biti pridijeljen jedan ili više

[2] klasifikacija

deskriptora. Oni se ponašaju na isti način kao kategorije kod procesa kategorizacije.

3.2.1. Tezaurus

Pojam tezaurusa pojavljuje se u različitim kontekstima. Postoje također različite klasifikacije i vrste tezaurusa, a onaj koji je bitan za svrhu indeksiranja naziva se konceptualni ili tematski tezaurus.

Svojstvo koje konceptualni tezaurus izdvaja kao posebnu vrstu leksikografskih priručnika je njegovo konceptualno, odnosno tematsko ustrojstvo, za razliku od abecednog koje se inače koristi. Osim toga, tezaurus karakterizira i semantička organizacija, koja se obično prikazuje kroz hijerarhijsku strukturu. Pojedini pojmovi sudjeluju u izgradnji nekog šireg i općenitijeg koncepta. Tezaurus se često pogrešno poistovjećuje s rječnikom sinonima. Razlika je u tome što on osim sinonima opisuje i mnoge druge značajno složenije leksičke relacije.

Osim opisanih svojstava, tezaurus sadrži abecedni indeks i mnoge druge informacije na nižoj razini [9].

3.2.2. Indeksiranje na temelju pravila

U sustavima temeljenim na pravilima automatski se stvaraju jednostavna pravila za kategorizaciju ili indeksiranje. Svako pravilo se stvara na temelju različitih kombinacija ili relacija pojmova u tekstu. Algoritam pronalazi takvo pravilo koje zadovoljava što više pozitivnih i što manje negativnih primjera. Nakon stvorenog pravila svi objekti koji ga zadovoljavaju se više ne uzimaju u obzir. Dalje se traži pravilo za ostale dokumente, sve dok taj skup ne postane potpuno prazan. Dodatna pravila mogu nastati postavljanjem novih uvjeta kategorizacije. Ti se uvjeti mogu odnositi na određene riječi koje moraju biti prisutne ili odsutne u tekstu, međusobnu udaljenost riječi, veličinu početnog slova i mnoga druga ograničenja.

Sva pravila su vidljiva i moguće ih je mijenjati radi reguliranja uspješnosti rada sustava. Svaka promjena deskriptora ili kategorije stvara novo ili modificira neko od starih pravila.

Stvaranje skupa pravila može se započeti s najopćenitijim ili najspecifičnijim pravilom, kao i kombinacijom ovih metoda. Cilj postupka je dobiti skup pravila koja zadovoljavaju sve pozitivne i ne zadovoljavaju niti jedan negativan primjer. Kod podataka s prisutnošću šuma to često nije moguće i klase se ne mogu jasno odvojiti.

Naučeni primjeri mogu poprimiti oblik pravila ili stabla. Pravila su izrazi u propozicijskoj logici oblika *ako-onda*, a rezultat može biti istina ili laž. Stabla odluke sastoje se od čvorova i grana. Svaki čvor predstavlja neku odluku koja se dalje grana na sve moguće ishode.

Osnovna prednost ovakve metode su ukupna cijena i vrijeme implementacije, no u nekim slučajevima to ipak nije isplativo, što će biti pokazano kasnije [10], [7].

3.2.3. Statističke metode

U sustavima temeljenim na statistici, različiti algoritmi određuju stupanj sličnosti među dokumentima. Oni dokumenti koji imaju mnogo zajedničkih pojmova, mogu se smatrati sličnima. Pretpostavlja se da oni predstavljaju isti koncept i zato su kategorizirani zajedno. Novi dokumenti klasificiraju se ili indeksiraju na temelju sličnosti sa skupom za učenje [10]. Skup za učenje predstavlja skup dokumenata koji su klasificirani ili indeksirani ručno i služi za učenje sustava.

Tehnike diskriminacije

Diskriminacijske analize pokušavaju naći funkciju koja najbolje odvaja dvije klase, dok linearne diskriminacijske analize traže linearnu kombinaciju obilježja koja ih najbolje dijeli. U kategorizaciji teksta se za svaku klasu traži funkcija koja najbolje odvaja objekte pridružene nekoj klasi od onih ostalih.

Jedna od uobičajenih tehnika je linearni diskriminator po najmanjim kvadratima. Ovdje se stvara hiperravnina za koju je suma kvadrata udaljenosti od vrijednosti obilježja najmanja. Svaki novi dokument je predstavljen točkom u n-dimenzionalnom prostoru. Odgovarajuća klasa se dodjeljuje ovisno o tome na kojoj se strani hiperravnine nalazi.

Druga tehnika je logistička diskriminacijska analiza ili logistička regresija. Obično se započinje sa linearnom diskriminacijskom funkcijom, koja se dalje iterativno mijenja kako bi najbolje odvojila dvije klase.

Alternativne diskriminacijske tehnike imaju pravilo za svaku klasu. Pravilo se bazira na zajedničkoj distribuciji pozitivnih primjera i zajedničkoj distribuciji negativnih primjera klase. Rezultat je težinski vektor gdje svaka komponenta predstavlja neko obilježje i pripadnu težinu za određenu klasu. Novi dokument se također predstavlja kao vektor, koji se uspoređuje sa svim

težinskim vektorima. Kada rezultat (sličnost ili udaljenost vektora) prijeđe neki prag, novi dokument se može svrstati u tu klasu.

Za određivanje težinskih vektora često se koriste Rocchio algoritam ili Widrow-Hoff algoritam (LMS algoritam) [7].

Klasifikator k najbližih susjeda (eng. k nearest neighbour - kNN)

Klasifikator k najbližih susjeda (kNN) radi samo na temelju pozitivnih primjera. Kada se pojavi novi dokument, njegov vektor se uspoređuje s vektorima ostalih već klasificiranih primjera. Pretpostavlja se da slični dokumenti pripadaju istoj klasi. Klasifikator pronalazi najbliži primjer i ako sličnost prelazi određeni prag, novom dokumentu će biti pridijeljena klasa tog pronađenog primjera. Alternativno, može se tražiti i *k* najbližih primjera.

Klasifikator kNN ima mnoge prednosti. Jedna od njih je da može učiti na temelju klasa koje se međusobno preklapaju. Osim toga, ako je primjer dobro odabran, klasifikator može donijeti odluku o novom dokumentu na temelju samo tog jednog primjera. Pošto se ne stvara poseban opis klase, nego se novi dokumenti uspoređuju sa svakim primjerom pojedinačno, moguće je mnogo fleksibilnije spajanje sličnih vektora.

Osim prednosti prisutni su i mnogi nedostaci. Na primjer, vrijeme učenja je zanemarivo, ali klasifikacija ili indeksiranje novog dokumenta je izrazito vremenski zahtjevno. Također postoje problemi oko pohrane velike količine primjera. Dodatan problem predstavlja činjenica da je kNN klasifikator vrlo osjetljiv na šum u podacima [7].

Bayesov naivni klasifikator

Za svaku klasu odabire se mali skup obilježja. Vjerojatnost da neki novi dokument pripada nekoj od klasa određuje se na temelju vjerojatnosti da su njegove karakteristike vezane uz tu klasu. Računanje vjerojatnosti pripadnosti novog dokumenta može se pojednostavniti korištenjem Bayesovog teorema, koji pretpostavlja da su obilježja međusobno nezavisna. Pripadnost klasi dodjeljuje se ako je vjerojatnost pripadnosti iznad određenog praga ili ako se radi o prvih *k* predloženih klasa.

Bayesov teorem glasi:

$$P\left(C_k = 1 \mid w_1 = x_1,..., w_p = x_p\right) = \frac{P\left(w_1 = 1,..., w_p = x_p \mid C_k = 1\right) \cdot P\left(C_k = 1\right)}{P\left(w_1 = x_1,.., w_p = x_p\right)},$$

gdje je C_k klasa, $P(C_k=1)$ a priori vjerojatnost da će C_k klasa biti pridijeljena, x događaj, a $w_1,\ldots,w_p$ skup od p obilježja.

Maronov model (1961.) uzima u obzir samo prisutnost pojma. Ako se pretpostavi nezavisnost obilježja onda formula glasi:

$$P\left(C_k = 1 \mid D_m\right) = P\left(C_k = 1\right) \cdot \prod_j \frac{P\left(w_j = 1 \mid C_k = 1\right)}{P\left(w_j = 1\right)},$$

gdje je D_m dokument, $P(w_j=1|C_k=1)$ vjerojatnost da se svojstvo w_j pojavljuje u tekstu iz skupa za učenje bitnom za klasu C_k, a $P(w_j=1)$ vjerojatnost da se svojstvo w_j pojavljuje u cijelom skupu dokumenata.

Fuhrov (1989) i Lewisov (1992) model uzimaju u obzir prisutnost i odsutnost pojma. Uz pretpostavku nezavisnosti obilježja ova formula glasi:

$$P(C_k = 1 \mid D_m) =$$
$$= P(C_k = 1) \cdot \prod_j \left(\frac{P(w_j = 1 \mid C_k = 1) \cdot P(w_j = 1 \mid D_m)}{P(w_j = 1)} + \frac{P(w_j = 0 \mid C_k = 1) \cdot P(w_j = 0 \mid D_m)}{P(w_j = 0)} \right),$$

gdje je $P(w_j=0|C_k=1)$ vjerojatnost da se svojstvo w_j pojavljuje u tekstu iz skupa za učenje bitnom za klasu C_k, $P(w_j=0)$ vjerojatnost da se obilježje w_j ne pojavljuje u ukupnom skupu dokumenata, $P(w_j=1|D_m)$ vjerojatnost da je obilježje w_j prisutno u dokumentu D_m, a $P(W_j=0|D_m)$ vjerojatnost da nije prisutno [7].

3.2.4. Neuronske mreže

Neuronska mreža se sastoji od međusobno povezanih ulaznih i izlaznih čvorova. Ulaznim čvorovima se pridjeljuje vrijednost obilježja na temelju kojih se aktiviraju daljnji čvorovi. Svaka jedinica izračunava određenu vrijednost koju, ako zadovoljava određene kriterije, dalje prenosi na svoje izlaze. Vrijednosti izlaznih čvorova određuju klasu kojoj određeni slučaj pripada. Zbog velike povezanosti mreže, pogreška u nekoliko pojmova je zanemariva, pa se neuronske mreže pokazuju kao jako dobre u slučajevima šuma ili pogrešaka.

Mreže uče na skupu primjera u više uzastopnih iteracija. Radi podešavanja parametara i smanjenja mogućnosti pogreške koristi se algoritam napredovanja unatrag (eng. *Backpropagation algorithm*) [11].

Veliki problem kod neuronskih mreža je njihova kompleksnost, pa je potrebno napraviti jako dobru selekciju značajki. To je jedan od razloga zašto se neuronske mreže, unatoč dobrim rezultatima, rijetko koriste kao klasifikatori tekstova [7].

3.3. Lingvističke i statističke metode

Lingvističke metode indeksiranja se mogu upotrijebiti za poboljšanje performansi statističkih metoda. One funkcioniraju na temelju lingvističke obrade tekstova i određuju pojmove koji su bitniji za formiranje indeksa pridjeljujući im veće težine. Takvi pojmovi najčešće su imenice ili grupe imenica, a cilj lingvističkih metoda je da ih prepoznaju u tekstu. Dodatna prednost je prepoznavanje fraza, što znatno poboljšava kvalitetu indeksiranja.

Lingvističke metode se kombiniraju sa statističkima prilikom određivanja težine, odnosno važnosti pojedinih pojmova. Doprinos lingvističkih metoda je u pronalaženju vrsta i grupa riječi, ali i mjesta pojavljivanja pojedinih pojmova u tekstu [12].

4. Prikaz teksta

Za uspješniju obradu teksta potrebno je obaviti predprocesiranje koje se najčešće sastoji od morfološke normalizacije i eliminacije stop riječi. Čak i nakon ovih postupaka tekst obično sadrži previše pojmova da bi svi ušli u proces učenja, pa je potrebno odrediti one pojmove koji su karakteristični za dotični tekst.

Skup tekstova se prikazuje u obliku matrice pojmova i dokumenata, koja se određuje uz pomoć težina riječi. Težina predstavlja važnost određene riječi za tekst i može se odrediti pomoću više različitih metoda.

Svaki je pojedini dokument u matrici pojmova i dokumenata predstavljen pomoću vektora. Na temelju sličnosti dva vektora određuje se i sličnost pripadnih dokumenata.

4.1. Morfološka normalizacija

U svim jezicima se javlja problem pojavljivanja istih pojmova u različitim oblicima, kao što su na primjer jednina ili množina, razna vremena i slično. Na taj će se način isti pojam prepoznati kao nekoliko različitih riječi i tako tretirati u cijelom procesu. Zbog toga je jedan važan korak predprocesiranja morfološka normalizacija, odnosno svođenje riječi na osnovni oblik. Još jedna prednost ovog procesa je smanjenje broja različitih značajki koje sudjeluju u daljnjoj obradi, odnosno smanjenje dimenzionalnosti vektorskog prostora, opisanog kasnije. Jedan primjer morfološke normalizacije je sljedeći:

zakon, zakonom, zakoni, zakona → zakon.

Jedan od osnovnih problema koji otežava rad s hrvatskim jezikom je činjenica što se radi o izrazito morfološki bogatom jeziku. Zbog toga se može pretpostaviti da bi morfološka normalizacija u slučaju obrade tekstova na hrvatskom jeziku imala još značajniji utjecaj nego pri radu s mnogim drugim jezicima.

4.2. Stop riječi

Za potrebe indeksiranja i ostalih oblika obrade teksta, sve riječi u tekstu nisu jednako značajne. Takve neznačajne riječi nazivaju se stop riječi i tvore stop listu. Stop lista služi za eliminiranje riječi koje ne nose nikakvo značenje za sadržaj teksta. Dodatna prednost korištenja ove metode je automatsko smanjenje dimenzionalnosti na vrlo jednostavan način.

Postoji više metoda za stvaranje stop liste, a najčešća se odnosi na pronalaženje najfrekventnijih riječi. Jedan način za konstrukciju ove liste je korištenje nekog općeg korpusa, koji sadrži mnoga područja. Stop lista se može kreirati i korištenjem korpusa koji se zapravo obrađuje. Na taj se način mogu dobiti u listi i riječi koje možda nisu toliko česte u cjelokupnom jeziku, ali se učestalo pojavljuju u dokumentima koji se trenutno obrađuju i za njih nemaju nikakvo značenje. Kriterij za određivanje riječi za stop listu su ili zadani minimalni prag frekvencije ili broj riječi.

Još jedna metoda za stvaranje stop liste je odabir kratkih riječi. Kako bi se spriječilo stavljanje bitnih, ali kratkih riječi u tu listu, formira se anti-stop lista. Osim navedenih kriterija za stvaranje stop liste, može se koristiti i informacija o vrsti riječi [7].

Slijedi primjer nekoliko stop riječi za hrvatski jezik:

a, ali, dakle, danas, ipak, ispod, moj, niti, onako, prije, takav, uzalud...

4.3. Odabir značajki

4.3.1. Frekvencija pojmova i dokumenata

Najjednostavniji način za odabir adekvatnih pojmova je računanje frekvencija pojavljivanja tog pojma u pojedinom tekstu i u preostalim dokumentima. Pojam se smatra važnim ako se pojavljuje određeni broj puta. Računanjem njegove frekvencije i odbacivanjem vrijednosti ispod nekog praga smanjuje se ukupni skup pojmova.

Drugi, prilično sličan način je određivanje broja dokumenata u kojima se taj pojam pojavio. Kao i u prvom slučaju, ako je taj broj ispod definiranog praga, pojam se odbacuje. Ova metoda polazi od pretpostavke da rijetki pojmovi ili nisu bitni za konkretnu temu ili ne utječu bitno na ukupan proces.

Frekvencija pojmova i dokumenata je najjednostavnija metoda za odabir bitnih pojmova i može se primjenjivati u slučajevima velikih količina podataka. Unatoč tome njezina preciznost je usporediva s ostalim metodama [13].

4.3.2. χ^2 test

Sljedeće se dvije metode, χ^2 test i omjer log-vjerodostojnosti (eng. *Log-likelihood ratio*), baziraju na testiranju hipoteza. Ove su metode primjenjive i u drugačijim slučajevima, ali između ostaloga pomažu u određivanju značajnosti pojam za određeni tekst. Radi se zapravo o usporedbi dva korpusa, koje može predstavljati i pojedinačni dokument. Jedan korpus predstavlja nultu hipotezu i referentan korpus, a za drugi korpus se u odnosu na prvi određuje da li se riječ pojavila određeni broj puta slučajno ili ne.

Osim χ^2 testa i metode log-vjerodostojnosti može se koristiti i t-test. On pretpostavlja normalnu razdiobu vjerojatnosti pojavljivanja pojmova i daje nešto lošije rezultate. Postoje i druge mogućnosti, kao što su Kolmogorov-Smirnov test (KS-test) i P-test [14].

Dobro svojstvo χ^2 testa je što ne pretpostavlja normalnu razdiobu vjerojatnosti, koja se rijetko pojavljuje u stvarnosti.

U metodi χ^2 test se radi o usporedbi očekivanih frekvencija iz referentnog korpusa i promatranih frekvencija korpusa čiji se pojmovi odabiru. Ako je razlika velika, nulta hipoteza o jednakosti frekvencija se odbacuje. Veća frekvencija pojavljivanja pojma u novom korpusu, odnosno dokumentu određuje taj pojam kao karakterističan za taj dokument ili korpus.

Za provođenje χ^2 testa gradi se tablica 1.

Tablica 1. χ^2 test

	Korpus 1	Korpus 2	Ukupno
Frekvencija pojma	a	b	a+b
Frekvencija ostalih pojmova	c	d	c+d
Ukupno (veličina korpusa)	a+c	b+d	a+b+c+d

Veličine *a* i *b* određuju promatrane vrijednosti i u sljedećoj formuli označavaju se s O_i (O_1 = a, O_2 = b). N_i označava veličine korpusa, a E_i očekivane vrijednosti, koje se računaju po formuli:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i},$$

odnosno prema prethodnoj tablici vrijedi:

$$\mathrm{E}_1 = (a + \mathrm{c}) \cdot \frac{a+b}{a+c+b+d} \qquad \text{i} \qquad \mathrm{E}_2 = (b + \mathrm{d}) \cdot \frac{a+b}{a+c+b+d}.$$

χ^2 se računa [1], [15] pomoću sljedeće formule i uspoređuje s kritičnim vrijednostima opisanim u sljedećem poglavlju:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i},$$

ili:

$$\chi^2 = \frac{N(ad - bc)^2}{(a+b)(c+d)(a+c)(b+d)}.$$

4.3.3. Omjer log-vjerodostojnosti (eng. *Log-Likelihood Ratio*)

Omjer vjerodostojnosti (eng. *Likelihood ratio*) nešto je lakše interpretirati nego χ^2 test. To je broj koji određuje koliko je jedna hipoteza vjerojatnija od druge. U većem broju slučajeva koristi se log-vjerodostojnost. Ova metoda se može primijeniti i u drugim situacijama, ali jedna od mogućnosti je usporedba korpusa.

Da bi se izračunala log–vjerodostojnost potrebno je formirati tablicu 2.

Tablica 2. Omjer log-vjerodostojnosti

	Korpus 1	Korpus 2	Ukupno
Frekvencija pojma	a	b	a+b
Frekvencija ostalih pojmova	c-a	d-b	c+d-a-b
Ukupno (veličina korpusa)	c	d	c+d

Vrijednost *c* određuje veličinu prvog, a vrijednost *d* veličinu drugog korpusa. U sljedećim formulama to su *N* vrijednosti. *a* i *b* su vrijednosti promatranog pojma, a u formulama su zajedno označene kao *O*. Očekivane vrijednosti računaju se kao kod χ^2 testa :

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

Iz tablice se može vidjeti da je $N_1=c$ i $N_2=d$, pa vrijedi:

$$\mathrm{E}_1 = \mathrm{c} \cdot \frac{a+b}{c+d} \qquad \text{i} \qquad \mathrm{E}_2 = \mathrm{d} \cdot \frac{a+b}{c+d}.$$

Izračun očekivanih vrijednosti već uzima u obzir veličine korpusa, pa dodatna normalizacija prije primjene formule nije potrebna. Pošto $-2\ln\lambda$ odgovara λ^2 distribuciji, log-vjerodostojnost se računa prema sljedećoj formuli:

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right).$$

Prema vrijednostima iz tablice 2 slijedi:

$$LL = 2\cdot\left(a\cdot\ln\left(\frac{a}{E_1}\right)+b\cdot\ln\left(\frac{b}{E_2}\right)\right)$$

Za određivanje kritičnih vrijednosti potrebno je odrediti proporciju *p* i broj stupnjeva slobode. Parametar *p* određuje postotak koliko jedna hipoteza mora biti vjerojatnija od druge, a broj stupnjeva slobode računa se prema formuli:

$$d.f. = (r-1)(c-1),$$

gdje *c* predstavlja broj stupaca, a *r* broj redaka u tablici.

Na temelju ovih parametara, kritične vrijednosti se očitavaju iz tablice χ^2 kritičnih vrijednosti prikazane u dodatku [1], [15]. Tablica se na isti način koristi i kod χ^2 testa.

4.3.4. Uzajamna informacija (eng. *Mutual Information* – MI)

Uzajamna informacija (eng. *mutual information*) određuje količinu informacije koju daje pojavljivanje jednog događaja o pojavljivanju drugog događaja. [1] Kod kategorizacije i automatskog indeksiranja prvi događaj se odnosi na pojmove u tekstu, a drugi na kategorije ili deskriptore [13].

Uzajamna informacija može se definirati i uz pomoć omjera vjerodostojnosti. Ona predstavlja omjer log-vjerodostojnosti pojavljivanja pojma i deskriptora (kategorije) zajedno i produkta pojavljivanja pojma i deskriptora pojedinačno [1].

Za računanje uzajamne informacije također se formira tablica 3.

Tablica 3. Uzajamna informacija

	Pridružen deskriptor / kategorija	Nema deskriptora / nije ta kategorija
Prisutnost pojma	A	B
Odsutnost pojma	C	D

Vrijednost *A* određuje broj dokumenata koji sadrže traženi pojam i opisani su određenim deskriptorom ili svrstani u određenu kategoriju. Vrijednost *B* određuje broj dokumenata u kojima se pojavljuje pojam bez deskriptora ili kategorije, a *C* odsutnost pojma uz pojavu deskriptora ili kategorije. Vrijednost *D* nije bitna za računanje uzajamne informacije. Ukupni broj dokumenata označavat će se sa *N*.

Uzajamna informacija se za događaje *x* i *y* računa kao:

$$I(x,y) = \log_2 \frac{P(xy)}{P(x)P(y)} = \log_2 \frac{P(x \mid y)}{P(x)} = \log_2 \frac{P(y \mid x)}{P(y)}.$$

U spomenutom slučaju kada su događaji pojmovi u tekstu i kategorije, događaj *x* se odnosi na riječ, a *y* na kategoriju. Koristeći gornju tablicu uzajamna informacija se može aproksimirati kao:

$$I(x,y) \approx \log_2 \frac{A \cdot N}{(A+C) \cdot (A+B)}.$$

U slučaju da su događaji x i y nezavisni, odnosno da pojam uopće nije bitan za određivanje dotičnog deskriptora ili kategorije vrijednost uzajamne informacije biti će 0.

Da bi se odredili pojmovi bitni za daljnji proces, primjenjuju se sljedeće mjere:

$$I_{avg}(x) = \sum_{i=1}^{m} P(y_i) I(x, y_i)$$

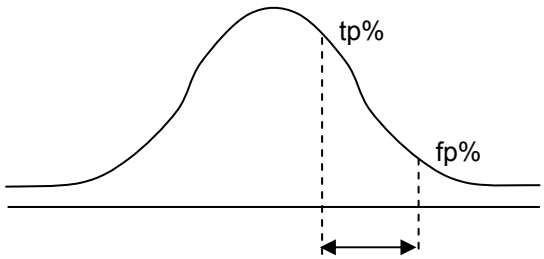
ili

$$I_{\max}(x) = \max_{i=1}^{m} \{I(x, y_i)\}.$$

Osnovni nedostatak ove metode je u tome što nije primjenjiva na različite pojmove s velikim odstupanjima u frekvencijama. Uzajamna informacija je dobar pokazatelj nezavisnosti događaja, ali kod zavisnih događaja pod utjecajem frekvencije pojavljivanja prvog događaja, odnosno pojma u tekstu [1], [13].

4.3.5. Bi-normalna separacija (eng. *Bi-normal separation*)

Bi-normalna separacija (BNS) se može promatrati na dva načina. Prvi je pomoću separacijskog praga, gdje se pretpostavlja se da je pojavljivanje svakog pojma u dokumentu modelirano pomoću slučajne varijable po normalnoj distribuciji, gdje krivulja prelazi određeni prag. Učestalost pojma odgovara području ispod krivulje, gdje ona prelazi prag. Ako se pojam češće nalazi u skupu pozitivnih primjera (u dokumentima koji pripadaju određenoj kategoriji), njegov će se prag nalaziti dalje od repa distribucije nego za negativne primjere (u dokumentima koji ne pripadaju toj kategoriji). BNS je mjera udaljenosti ta dva praga.



Slika 1. BNS mjera pomoću separacijskog praga

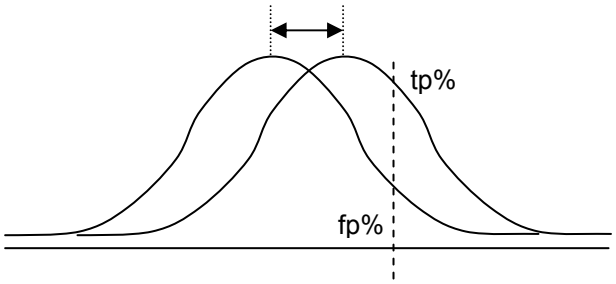
Za drugi pristup bi-normalnoj separaciji potrebno je uvesti dva nova pojma. To su mjera uzorka točno pozitivnih *tpr (*eng. *sample true positive rate*) i mjera uzorka netočno pozitivnih primjera *tfr (*eng. *sample false positive rate*):

$$tpr = \frac{tp}{pos} \qquad \text{i} \qquad tfr = \frac{fp}{neg},$$

gdje *pos* označava sve pozitivne primjere (dokumente u nekoj kategoriji), *neg* sve negativne (dokumenti koji nisu u toj kategoriji), *tp* broj pozitivnih primjera koji sadrže pojam, a *tf* broj negativnih koji sadrže pojam.

Ova mjera određuje horizontalnu udaljenost između dvije krivulje po normalnoj razdiobi, čija je udaljenost određena prema *tpr* i *tfr* vrijednostima.

BNS mjera je proporcionalna području ispod ROC krivulje, koja nastaje kao rezultat preklapanja normalnih razdioba.



Slika 2. BNS mjera pomoću ROC krivulje

Formula za računanje BNS je dana sa:

$$\mathrm{F}^{-1}(\mathrm{tpr}) - \mathrm{F}^{-1}(\mathrm{fpr})$$

ili

$$BNS(x_i, y) = \left| F^{-1}\left(\frac{p(x_i = 1, y = 1)}{p(y = 1)} \right) - F^{-1}\left(\frac{p(x_i = 1, y = 0)}{p(y = 0)} \right) \right|,$$

gdje je F^{-1} inverz funkcije normalne razdiobe [16], [17].

4.3.6. Omjer izgleda (eng. *Odds Ratio - OR*)

Ova mjera je način provjere da li je vjerojatnost jednaka za dva skupa. Omjer izgleda odražava vjerojatnost pojavljivanja pojma u pozitivnom skupu, a normalizirana je vjerojatnošću u negativnom.

Za računanje gradi se tablica 4.

Tablica 4. Omjer izgleda

	Nema deskriptora / nije ta kategorija	Pridružen deskriptor / kategorija
Odsutnost pojma	a	b
Prisutnost pojma	c	d

Vjerojatnost za odsutnost pojma je *a/b*, a za prisutnost pojma *c/d*. Iz toga se može dobiti formula za omjer izgleda [18]:

$$OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot b}{c \cdot d}.$$

4.3.7. Informacijska dobit (eng. *Information Gain* - IG)

Informacijska dobit mjeri količinu informacije za predviđanje kategorije koja se dobiva na temelju prisutnosti ili odsutnosti pojma u dokumentu. Konkretno, računaju se razlike u entropijama, a osnovna prednost metode je u tome što se maksimiziranjem informacijske dobiti minimizira nesigurnost u konačnim rezultatima. $\{c_i\}_m^{i=1}$ predstavlja skup kategorija, pa se informacijska dobit za pojam *t* definira kao:

$$G(t) = -\sum_{i=1}^{m} p(c_i)\log p(c_i) + p(t)\sum_{i=1}^{m} p(c_i \mid t)\log p(c_i \mid t) + p(\bar{t})\sum_{i=1}^{m} p(c_i \mid \bar{t})\log p(c_i \mid \bar{t})$$

Za svaki pojam iz skupa dokumenata za učenje računa se informacijska dobit i odbacuju oni pojmovi čija je vrijednost ispod određenog praga [13], [1].

4.3.8. Povezanost riječi (eng. *Term Strength* - TS)

Metoda povezanosti riječi procjenjuje važnost pojma na temelju toga koliko je vjerojatno da će se pojam pojaviti u usko povezanim dokumentima. Skup za učenje koristi se za dobivanje parova dokumenata čija sličnost prelazi određeni prag. Sličnost se računa na temelju kosinusa dva vektora koji predstavljaju dokumente. Povezanost riječi se računa procjenom uvjetne vjerojatnosti da se pojam nalazi u drugoj polovici para sličnih dokumenata uz uvjet da se nalazi u prvoj polovici para.

Ako su *x* i *y* slični dokumenti koji tvore jedan par, a *t* pojam, onda se povezanost riječi računa pomoću:

$$s(t) = p(t \in y \mid t \in x) .$$

Pretpostavlja se da dokumenti koji sadrže mnoge slične riječi su i sami slični, a te riječi sadrže velik dio informacije. Ova metoda ne uzima u obzir povezanost između pojma i kategorije. Po tome je slična metodi frekvencije dokumenata, a znatno se razlikuje od, na primjer, informacijske dobiti ili uzajamne informacije [13].

4.4. Matrica pojam-dokument

Najčešća metoda prikaza tekstualnih dokumenta za procesiranje računalom je vektorska reprezentacija (eng. *vector space model*). Ta se struktura naziva vreća riječi (eng. *bag-of-words*). Uzimaju se sve riječi iz dokumenta bez da se koristi ikakav redoslijed ili struktura. U skupu dokumenata svaki je dokument predstavljen kao vreća riječi sadržavajući sve riječi koje se u njemu javljaju. Vreća riječi se može proširiti dodatnim svojstvom da može sadržavati nizove riječi (n-grame) umjesto samih pojmova, što može znatno poboljšati performanse sustava za obradu teksta [19].

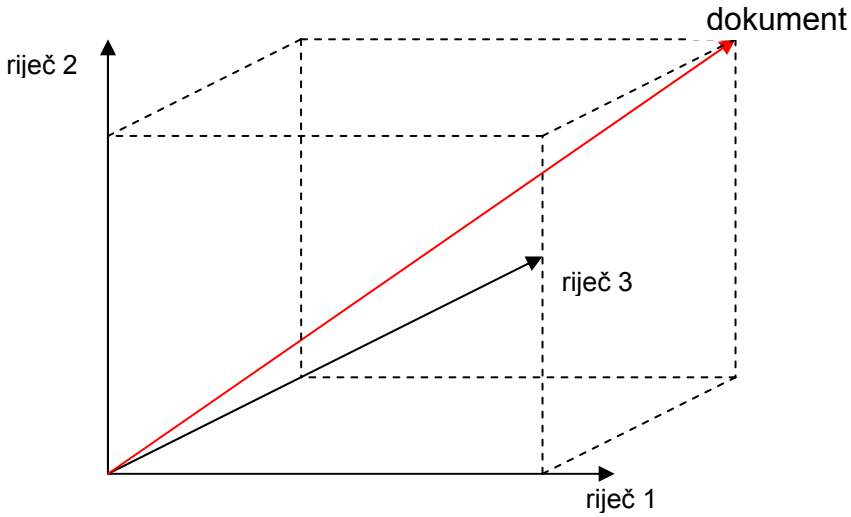
Vektori redaka i stupaca matrice pojmova i dokumenata A = [a_{ij}] na slici 3 predstavljaju sve pojmove i dokumente u korpusu.

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \dots & a_{ij} & \dots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} \begin{matrix} w_1 \\ \dots \\ w_i \\ \dots \\ w_m \end{matrix}$$

$$d_1 \quad \dots \quad d_j \quad \dots \quad d_n$$

Slika 3. Matrica pojam-dokument

Element a_{ij} predstavlja broj pojavljivanja pojma w_i u dokumentu d_i. Osim same frekvencije moguće je za računanje elemenata matrice koristiti i neku od težinskih funkcija opisanih u sljedećem poglavlju. Semantička povezanost između dokumenata se procjenjuje, na primjer, računanjem kosinusa kuta između dva vektora dokumenta [20]. Dokumenti su prikazani u n-dimenzionalnom prostoru, gdje svaka riječ odgovara jednoj dimenziji, kao na slici 4.



Slika 4. Dokument u vektorskom prostoru

4.5. Težina riječi

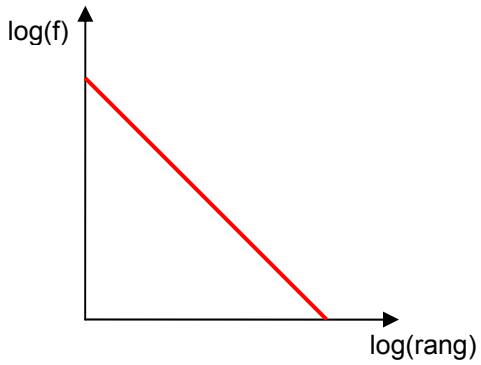
4.5.1. Zipfovo pravilo

Zipfovo pravilo se bazira na ideji da su riječi koje se javljaju često u tekstu vrlo uobičajene i ne nose previše značenja. Ista je situacija i sa rijetkim riječima, jer ukoliko se pojavljuju svega jedan ili dva puta u dugačkom tekstu, ne mogu se smatrati bitnima za sadržaj.

Svakoj riječi pridijeljen je neki rang, na način da najfrekventnija riječ ima rang 1, sljedeća 2 itd. Prema Zipfovom pravilu produkt ranga i logaritma frekvencije je konstantan.

$$\log(\,frekvencija) \cdot rang = konst.$$

Na Zipfovu krivulju na slici 5 može se primijeniti Luhnov koncept, da se najvažnije riječi nalaze na sredini krivulje i imaju srednje vrijednosti frekvencija. Na tome se baziraju mnoge težinske funkcije.



Slika 5. Zipfova krivulja

4.5.2. Produkt frekvencije pojma i inverzne frekvencije dokumenata

Normalno je za pretpostaviti da će riječi koje su bitne za dokument autor češće spominjati u tekstu. Zato se kao jednu od mjera težine pojmova uzima njihova frekvencija (*tf*), koja zapravo označava broj pojavljivanja određenog pojma u tekstu. Prilično je očito da će rijedak pojam imati znatno veći utjecaj na kraće tekstove. Zato se radi smanjenja utjecaja frekvencije pojmova kod tekstova s velikim razlikama u duljini često koristi običan ili prirodni logaritam. Frekvencija pojmova je znatno učinkovitija mjera težine kod dužih tekstova, jer kod onih kraćih se sve riječi pojavljuju samo nekoliko puta i ta informacija može navesti na krive zaključke.

Čak i nakon eliminacije stop riječi tekst uvijek sadrži mnoge uobičajene riječi koje ne nose nikakvo značenje za sadržaj. U što se više tekstova neki pojam javi to je njegovo značenje manje važno. Ako se pojam javlja u jako malom broju tekstova, on može vrlo dobro odvajati te tekstove od ostalih. Zato težina pojma treba obrnuto proporcionalno ovisiti o broju tekstova u kojima se pojam javlja, a naziva se inverzna frekvencija dokumenata (*idf*). Ona se obično računa kao:

$$idf_i = \log\left(\frac{N}{n_i}\right),$$

gdje N označava ukupan broj dokumenata, a n_i broj dokumenata koji sadrže pojam i. Umjesto običnog može se koristiti i prirodni logaritam.

Produkt frekvencije pojma i inverzne frekvencije dokumenata je mjera težine koja uzima u obzir oba ranije navedena kriterija. Najbolji pojmovi su oni koji se pojavljuju često unutar teksta, ali rijetko u drugim dokumentima. Ova mjera računa se najčešće pomoću:

$$tf \times idf_i = tf_i \cdot \log\left(\frac{N}{n_i}\right).$$

4.5.3. Normalizacija duljine

Tekstovi u dokumentima mogu biti različitih dužina. U dužim tekstovima su frekvencije pojmova znatno veće, zbog čega je nemoguće odrediti stvarnu važnost pojma. Zato je često potrebno provesti normalizaciju duljine. Normalizacija duljine je često uključena u težinske funkcije, a najčešće se normalizira upravo faktor frekvencije pojmova.

Frekvencija pojma može se normalizirati tako da se frekvencija pojma *i* podijeli s maksimalnom frekvencijom nekog pojma *j* nađenog u tekstu, što se može vidjeti iz formule:

$$\frac{tf_i}{\max tf_j}.$$

Rezultat ove normalizacije je između 0 i 1. Kako bi se smanjila razlika između čestih i rijetkih pojmova, dobivenu je vrijednost moguće pomnožiti sa 0.5. Rezultatu se dodaje 0.5 i konačni iznos poprima vrijednost između 0.5 i 1:

$$0.5 + 0.5\left(\frac{tf_i}{\max tf_i}\right).$$

Još jedan uobičajeni način normalizacije je kosinusna normalizacija dana formulom:

$$\frac{tf_i}{\sqrt{\sum (tf_j)^2}}.$$

Osim na frekvenciju pojma, kosinusna normalizacija može se primijeniti i na produkt frekvencije pojma i inverzne frekvencije dokumenta:

$$\frac{tf_i \cdot \log\left(\frac{N}{n_i}\right)}{\sqrt{\sum\left(tf_j \cdot \log\left(\frac{N}{n_i}\right)\right)^2}}.$$

4.5.4. Vrijednost diskriminacije pojma

Model diskriminacije pojmova pretpostavlja da su najbolji pojmovi za identifikaciju sadržaja oni koji razlikuju jedan dokument od drugoga. Vrijednost diskriminacije pojma mjeri koliko neki pojam pomaže u razlikovanju tog dokumenta od ostalih. Loši pojmovi povećavaju povezanost između tekstova, dok ju dobri smanjuju. Vrijednost diskriminacije pojma računa se kao razlika u povezanosti tekstova prije i poslije dodavanja tog pojma.

Vrijednost diskriminacije pojma može zamijeniti inverznu frekvenciju dokumenta ili produkt frekvencije pojma i inverzne frekvencije dokumenta, ali ipak ima i mnoge zamjerke.

4.5.5. Težine važnosti pojma

Težina važnosti pojma se određuje na temelju vjerojatnosti njegovog pojavljivanja u bitnim i nebitnim dokumentima. Bitni dokumenti predstavljaju one koji pripadaju nekoj kategoriji ili su opisani nekim deskriptorom, dok su nebitni svi ostali. Postoje mnoge funkcije za računanje težine važnosti pojma, a jedna od njih je sljedeća:

$$\log \frac{\left(\frac{r_i}{R - r_i} \right)}{\left(\frac{n_i - r_i}{N - n_i - R + r_i} \right)},$$

gdje je *N* ukupan broj tekstova, *R* broj bitnih tekstova, n_i broj tekstova koji sadrže pojam *i*, a r_i broj bitnih tekstova koji sadrže pojam *i* [7].

4.6. Sličnosti vektora

Uobičajena mjera sličnosti među vektorima je već ranije spomenuti kosinus kuta među njima, no moguće je i korištenje raznih drugih metoda, kao i njihovih kombinacija.

Za usporedbu podataka mogu se koristiti binarni vektori, gdje svaka komponenta može poprimiti vrijednost 0 ili 1. U tablici 5. su dane mjere sličnosti za takve vektore.

Tablica 5. Sličnosti binarnih vektora

Mjera sličnosti	**Definicija**
Koeficijent podudaranja	$\lvert X \cap Y \rvert$
Dice koeficijent	$\frac{2\lvert X \cap Y \rvert}{\lvert X \rvert + \lvert Y \rvert}$
Jaccard (ili Tanimoto) koeficijent	$\frac{\lvert X \cap Y \rvert}{\lvert X \cup Y \rvert}$
Koeficijent preklapanja	$\frac{\lvert X \cap Y \rvert}{\min(\lvert X \rvert, \lvert Y \rvert)}$
Kosinus	$\frac{\lvert X \cap Y \rvert}{\sqrt{\lvert X \rvert \times \lvert Y \rvert}}$

Dokumenti se znatno bolje predstavljaju pomoću vektorskog prostora, kojeg koriste i sljedeće mjere sličnosti.

4.6.1. Skalarni produkt

Skalarni produkt vektora $\vec{x}$ i $\vec{y}$ definira se pomoću sljedeće formule:

$$\vec{x} \cdot \vec{y} = \sum_{i=1}^{n} x_i y_i \ .$$

4.6.2. Kosinus

Kosinus zapravo određuje kosinus kuta između dva vektora. On poprima vrijednosti od 1 za vektore u istom smjeru, 0 za ortogonalne vektore i –1 za vektore u suprotnim smjerovima.

Duljina vektora definira se kao:

$$\left|\vec{x}\right| = \sqrt{\sum_{i=1}^{n} x_i^2} \ ,$$

a kosinusna mjera n-dimenzionalnih vektora je:

$$\cos\left(\vec{x}, \vec{y}\right) = \frac{\vec{x} \cdot \vec{y}}{\left|\vec{x}\right|\left|\vec{y}\right|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \ .$$

4.6.3. Okapi

Računanje sličnosti ovom metodom bazira se na Robertson-Spark Jones težini:

$$w^{(1)} = \log \frac{\left(\frac{r + 0.5}{R - r + 0.5} \right)}{\left(\frac{n - r + 0.5}{N - n - R + r + 0.5} \right)},$$

gdje je *N* ukupni broj dokumenata, *n* broj dokumenata koji sadrže određeni pojam, *R* broj dokumenata u jednoj kategoriji i *r* broj takvih dokumenata koji sadrže određeni pojam.

Okapi formula glasi:

$$Okapi_{t,d} = \sum_{l \in t \cap d} w^{(1)} \frac{(k_1 + 1)TF_{l,d}}{K + TF_{l,d}} \frac{(k_3 + 1)TF_{l,t}}{k_3 + TF_{l,t}},$$

gdje je *K* dan sa:

$$K = k_1 \left((1 - b) + b \frac{|d|}{M} \right),$$

a k_1, k_3 i *b* predstavljaju konstante, tako da je k_1 između 1.0 i 2.0, *b* obično 0.75, a k_3 između 0 i 1000. $TF_{l,d}$ je frekvencija pojma u opisu deskriptora, a $TF_{l,t}$ frekvencija pojma u tekstu novog dokumenta. $|d|$ je broj asocijata u listi za pojedini deskriptor, a *M* prosječna dužina liste asocijata.

Za potrebe indeksiranja okapi formula se može aproksimirati sa:

$$Okapi_{t,d} = \sum_{l \in t \cap d} \left(\log \frac{N - DF_l + 0.5}{DF_l + 0.5} \frac{(k1+1)TF_{l,d}}{k_1\left((1-b) + b\frac{|d|}{M}\right) + TF_{l,d}} \frac{(k_3+1)TF_{l,t}}{k_3 + TF_{l,t}} \right),$$

gdje je DF_l broj deskriptora kojima je lema *l* pridijeljena kao asocijat, a *N* ukupni broj deskriptora [21], [22], [23], [24].

5. Mjere efikasnosti i usporedbe

Za odrađivanje efikasnosti potrebno je odrediti odnos stvarnih kategorija i onih dobivenih radom sustava, kao u tablici 6. *TP* predstavlja točno pozitivno, *TN* točno negativne, *NP* netočno pozitivno, a *NN* netočno negativno klasificirane primjere [25].

Tablica 6. Kontigencijska tablica (eng. *Confusion matrix*)

		klasa	
		pozitivno	negativno
predviđanje	pozitivno	TP	NP
	negativno	NN	TN

Na temelju ove tablice moguće je odrediti različite mjere efikasnosti, kao što su preciznost, odziv, E-mjera i F-mjera. U slučajevima kada se javlja više kategorija koristi se mikro ili makro usrednjavanje.

5.1. Odziv i preciznost

Odziv (eng. *recall, R*) je udio članova neke klase koje je sustav ispravno klasificirao u ukupnom broju članova te klase.

$$R = \frac{TP}{NN + TP}$$

Točnost (eng. *precision, P*) je udio članova neke klase koji su joj dodijeljeni sustavom u ukupnom broju članova koji joj pripadaju.

$$P = \frac{TP}{NP + TP}$$

5.2. E-mjera

Za usporedbu više sustava potrebno je uzeti u obzir oba spomenuta faktora, što je omogućeno pomoću E-mjere:

$$E_\beta = 1 - \frac{(\beta^2 + 1)P \cdot R}{\beta^2 P + R},$$

gdje faktor β određuje relativnu važnost između preciznosti i odziva.

5.3. F-mjera

Drugi način povezivanja preciznosti i odziva u jednu mjeru je F-mjera. Ona se može izraziti pomoću E-mjere:

$$F_\beta = 1 - E_\beta,$$

ili pomoću preciznosti i odziva:

$$F_\beta = \frac{(\beta^2 + 1)P \cdot R}{\beta^2 P + R},$$

gdje faktor β ponovno određuje relativnu važnost između preciznosti i odziva. Ako se njegova vrijednost postavi na 1, što znači da se preciznost i odziv uzimaju u obzir s jednakim omjerima, dobiva se F_1 mjera [7]:

$$F_1 = \frac{2P \cdot R}{P + R}.$$

5.4. Makro i mikro usrednjavanje (eng. Micro and macro averaging)

Ukoliko je potrebno odrediti efikasnost kategorizacije u slučajevima kada se javljaju više od dvije kategorije mogu se koristiti dva različita pristupa. Prvi je makro usrednjavanje, kada se računa efikasnost prema određenoj mjeri za svaku kategoriju posebno. Ukupna efikasnost se dobiva kao aritmetička sredina pojedinačnih.

Drugi pristup naziva se mikro usrednjavanje. Ovom metodom računaju se skupovi *TP*, *TN*, *NP*, *NN* za sve kategorije zajedno i na temelju toga određuje ukupna efikasnost. Osnovna razlika ova dva pristupa je u tome da makro usrednjavanje daje jednaku težinu svim kategorijama, dok mikro usrednjavanje izjednačuje sve objekte. Mikro usrednjavanje daje bolje rezultate za velike kategorije, dok je makro usrednjavanje bolje za sve kategorije zajedno [1].

6. Primjeri sustava za automatsko indeksiranje

U ovom poglavlju je opisano sedam različitih sustava za automatsko indeksiranje. Oni koriste različite metode indeksiranja, ali im je svima zajedničko da se deskriptori dodjeljuju na temelju kontroliranog rječnika. Uz sheme sustava i opis načina rada, dani su i rezultati provedenih testova.

6.1. AUTINDEX

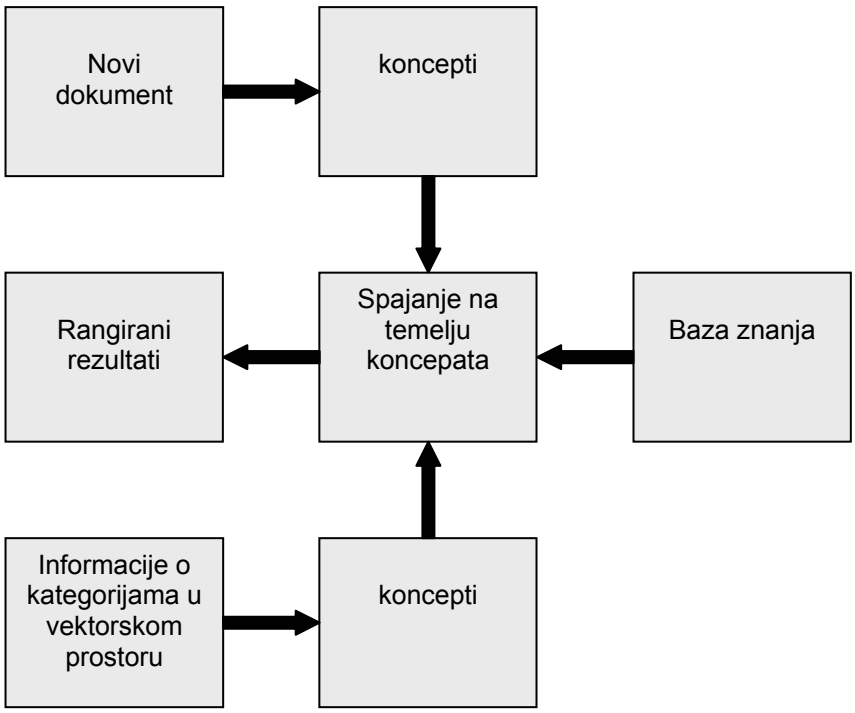
AUTINDEX je sustav za automatsko indeksiranje i klasifikaciju koji se provodi u dva osnovna koraka. U prvom se obavlja indeksiranje pomoću nekontroliranog rječnika uz upotrebu metoda za proseciranje prirodnog jezika, a rezultat je lista ključnih riječi. Ta lista se prerađuje u drugom koraku pomoću kontroliranog rječnika u obliku tezaurusa.

Trenutačna verzija AUTINDEX sustava radi za engleski i njemački jezik, a osim toga i za dvojezično indeksiranje i klasifikaciju.

U prvom koraku rada sustava obavlja se lingvistička analiza, koja se sastoji od segmentacije, razrješavanja višeznačnosti i lematizacije. Za određivanje ključnih riječi koriste se statističke funkcije bazirane na frekvenciji. Tu se ne radi o frekvenciji samih pojmova, nego frekvenciji semantičkih klasa, koje se računaju za svaku riječ prilikom lingvističke analize. U obzir se uzimaju samo imenice, glagoli i pridjevi, za koje se računaju semantičke značajke. Na kraju se uzimaju najfrekventnije semantičke značajke i svaka riječ ili fraza u tekstu koja ju sadrži ulazi u tzv. skup ključnih riječi (eng. *keyword set*).

U drugom koraku lista ključnih riječi ulazi u proces indeksiranja pomoću tezaurusa, sličan indeksiranju temeljenom na konceptima (eng.

concept-based indexing). Shematski prikaz indeksiranja temeljenog na konceptima dan je na slici 6 [26].



Slika 6. Shema indeksiranja temeljenog na konceptima

Rezultat rada AUTINDEX sustava je strukturirana lista deskriptora, ključnih riječi, informacija o klasifikaciji i raznih drugih parametara. Listu deskriptora korisnik može i ručno mijenjati, kao i dodavati pojmove tezaurusu.

Sustav je korišten za indeksiranje 500 sažetaka na njemačkom jeziku. Omjer automatski i ručno dodijeljenih deskriptora je iznad 70%, a vrijeme indeksiranja po dokumentu iznosilo je 25 sekundi, što je znatno manje nego u slučaju ručnog indeksiranja [27].

6.2. AIR/X

AIR/X je sustav za indeksiranje temeljen na pravilima i koristi pojmove iz određenog rječnika (deskriptore). Za obavljanje indeksiranja potreban je rječnik koji se sastoji od pravila koja preslikavaju pojmove u deskriptore, a gradi se automatski iz ručno indeksiranih dokumenata. Sustav radi sa sažecima tekstova na engleskom jeziku.

Za stvaranje rječnika za indeksiranje koristi se faktor dodjeljivanja *z(t,s)* (eng. *association factor*) za pojam *t* i deskriptor *s*, definiran kao:

$$z(t,s) = \frac{h(t,s)}{f(t)},$$

gdje *f(t)* označava broj dokumenata koji sadrže pojam *t*, a *h(t,s)* broj onih od *f(t)* dokumenata kojima je deskriptor ručno pridružen. Ako *z(t,s)* prelazi određeni prag, pravilo oblika $t \rightarrow s$ se sprema u rječnik.

Postupak indeksiranja novih dokumenata se bazira na Darmstadtovom pristupu indeksiranju (eng. *Darmstadt Indexing Approach* – DIA) [28] i dijeli se na dva koraka: opis i odluku. U prvom koraku se sakupljaju informacije o povezanosti pojmova i deskriptora. Ti podaci tvore bazu za odlučivanje koja se koristi u sljedećem koraku. Drugi korak započinje identifikacijom pojmova. Pošto je taj zadatak nemoguće savršeno izvesti, svaki se pojam dodatno identificira s posebnim oblikom pojavljivanja *v* (eng. *form of occurence,* FOC), gdje različiti FOC-ovi označavaju različite nivoe sigurnosti. FOC-ovi se definiraju na temelju sigurnosti da je neki pojam stvarno pronađen (za pojmove koji se sastoje od više riječi) i važnosti pojma u dokumentu (npr. frekvencija, položaj u dokumentu...). Ako je pojam *t* pronađen u dokumentu *d*, a u rječniku se nalazi pravilo oblika $t \rightarrow s$, stvara se deskriptorska indikacija (eng. *descriptor indication*) od pojma *t* prema deskriptoru *s*. Skup svih

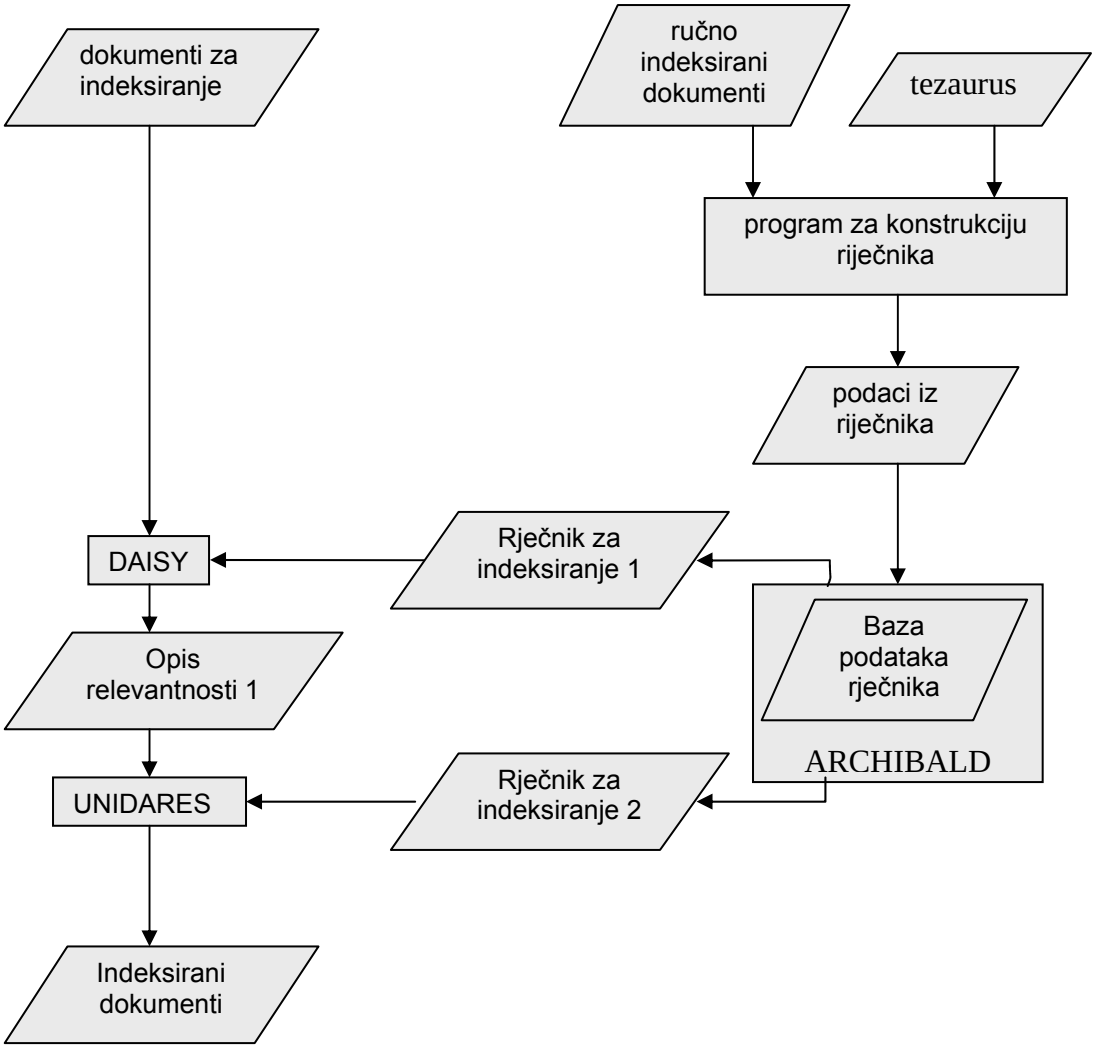
deskriptorskih indikacija koje vode od dokumenta *d* prema deskriptoru *s* naziva se opis relevantnosti *RD x(s,d)* (eng. *relevance description*).

Dobiveni opisi relevantnosti koriste se u drugom koraku, kako bi se odredilo koliko je dodjela deskriptora *s* dokumentu *d* ispravna. Za tu procjenu koristi se funkcija indeksiranja *a(x)* (eng. *indexing function*). Korištene su različite funkcije indeksiranja, a jedna od uspješnijih, korištena u AIR/PHYS sustavu, opisanom kasnije, zove se polinomi najmanjih kvadrata (eng. *Least square polynomials*). Funkcija indeksiranja je oblika:

$$a(\vec{x}) = \vec{b}^{T} \cdot \vec{x} ,$$

gdje je $\vec{b}$ vektor koeficijenata koji minimiziraju očekivanje kvadratne pogreške.

Dva spomenuta koraka se prema Darmstadtovom pristupu indeksiranju ponavljaju u nekoliko faza indeksiranja. Korištenjem više faza, svaka postaje jednostavnija i spretnija za održavanje. AIR/X sustav koristi dvije faze i prikazan je na slici 7.



Slika 7. Shema AIR/X sustava

Podsustav ARCHIBALD zadužen je za rad s bazom podataka rječnika. Osim toga, on omogućava ekstrakciju rječnika za indeksiranje iz baze za dvije faze rada cijelog sustava. Podsustav DAISY identificira pojmove i gradi prvi RD skup. Prvo odlučivanje, konstrukcija drugog RD skupa i drugo odlučivanje obavljaju se u UNIDARES podsustavu.

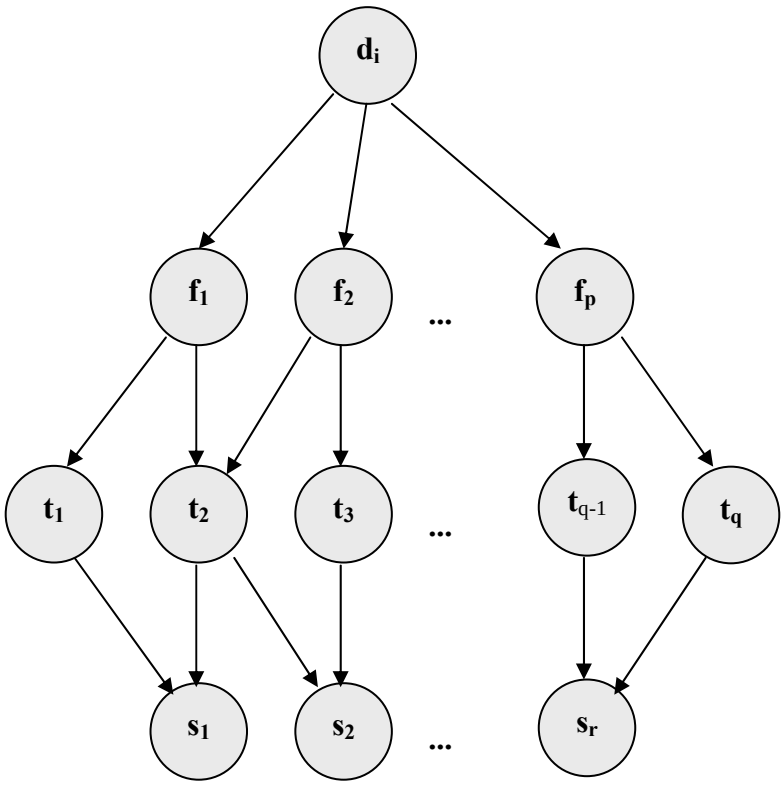
Cijeli AIR/X sustav je temeljen na pravilima, pa se promjenom baze pravila može primijeniti u različitim područjima. Jedna bitnija AIR/X aplikacija je AIR/PHYS sustav izrađen za rad s velikom bazom PHYS s područja fizike u Fachinformationszentrum Karlsruhe u Njemačkoj. 1983. godine izrađen je

eksperimentalni prototip i na temelju dobrih rezultata 1985. je započeo razvoj sustava. Baza PHYS je već 1992. godine sadržavala preko milijun dokumenata uz velik godišnji porast [29], [30].

6.3. Bayesove mreže

Indeksiranje na temelju Bayesovih mreža nastalo je kao alternativa Darmstadtovom pristupu indeksiranju. Preuzeti su mnogi koncepti iz prethodne metode, kao što su identifikacija pojmova i izgradnja rječnika za indeksiranje.

Mreža koja se koristi za automatsko indeksiranje prikazana je na slici 8.



Slika 8. Mreža za automatsko indeksiranje

Dokument koji je potrebno indeksirati označen je sa d_i. Sa $t_1, ..., t_q$ su označeni pojmovi identificirani u dokumentu, a sa $f_1, ..., f_p$ FOC-ovi pridruženi pojmovima. Pojmovi $t_1, ..., t_q$ se preslikavaju u deskriptore $s_1, ..., s_r$ pomoću rječnika za indeksiranje. Ako je pojam *t* pronađen u tekstu i u rječniku postoji relacija *(t,s)*, tada se stvara veza na grafu između pojma *t* i deskriptora *s*.

Svaki od čvorova sadrži vrijednost koja opisuje njegovu zavisnost o roditeljskom čvoru. Na taj način su uzete u obzir različite nesigurnosti, kao što su identifikacija pojma ili pridruživanje deskriptora pojmu na temelju ručno indeksiranih dokumenata. Pomoću tih vrijednosti može se odrediti vjerovanje (eng. *belief*) za svaki čvor. Vjerovanje pridruženo *s* čvorovima određuje vjerojatnost da deskriptor s_j ispravno indeksira dokument d_i. Na samom kraju postupka odabiru se samo oni deskriptori koji prelaze određeni prag.

Rađeni su eksperimenti s istom bazom PHYS, kao i kod Darmstadtovog pristupa indeksiranju. Pokusi pokazuju da je ova metoda primjenjiva za indeksiranje, ali su rezultati ipak bolji u prethodnom slučaju [31].

6.4. Indeksiranje prema BI-RADS leksikonu

Rezultati mamografije su se uglavnom pohranjivali u obliku slobodnog teksta, koji je ručno bilo teško pretraživati i klasificirati. S porastom broja izvještaja javila se potreba za automatskim indeksiranjem. Američki radiolozi odredili su 5 osnovnih tipova izvještaja, koji se također javljaju u BI-RADS leksikonu. Unutar tih 5 kategorija on sadrži 43 deskriptora.

Algoritam se dijeli na fazu učenja i indeksiranja. Koristi se matrica preslikavanja *W*, koja predstavlja linearno preslikavanje između frekvencija pojmova u izvještajima i njima pridruženih BI-RADS pojmova. Matrica *A* predstavlja izvještaje (matrica pojmova i dokumenata), gdje se stupci se odnose na izvještaje, a retci na pojmove u tim dokumentima. Svaki element matrice je cijeli broj koji pokazuje koliko puta se određeni pojam pojavio u nekom izvještaju. BI-RADS pojmovi su predstavljeni pomoću matrice *B*, koja se gradi na temelju prethodno indeksiranih dokumenata tako da stupci predstavljaju dokumente, a retci deskriptore. Ako je neki deskriptor pridružen određenom dokumentu, odgovarajući element matrice će imati vrijednost 1, a inače 0. Matrica *W* računa se jednom za cijeli postupak, pomoću sljedeće linearne jednadžbe:

$$WA=B$$

Elementi matrice *W* računaju se metodom najmanjih kvadrata (eng. *Linear Least Sqares Fit,* LLSF) uz pomoć metode dekompozicije na singularne vrijednosti (eng. S*ingular Value Decomposition* - SVD).

Novi izvještaj može se predstaviti vektorom $\vec{a}$ u kojem je svaki element frekvencija određenog pojma (riječi ili fraze). Izvještaj se kodira računajući vektor $\vec{b}$, koji sadrži BI-RADS pojmove dodijeljene izvještaju, pomoću:

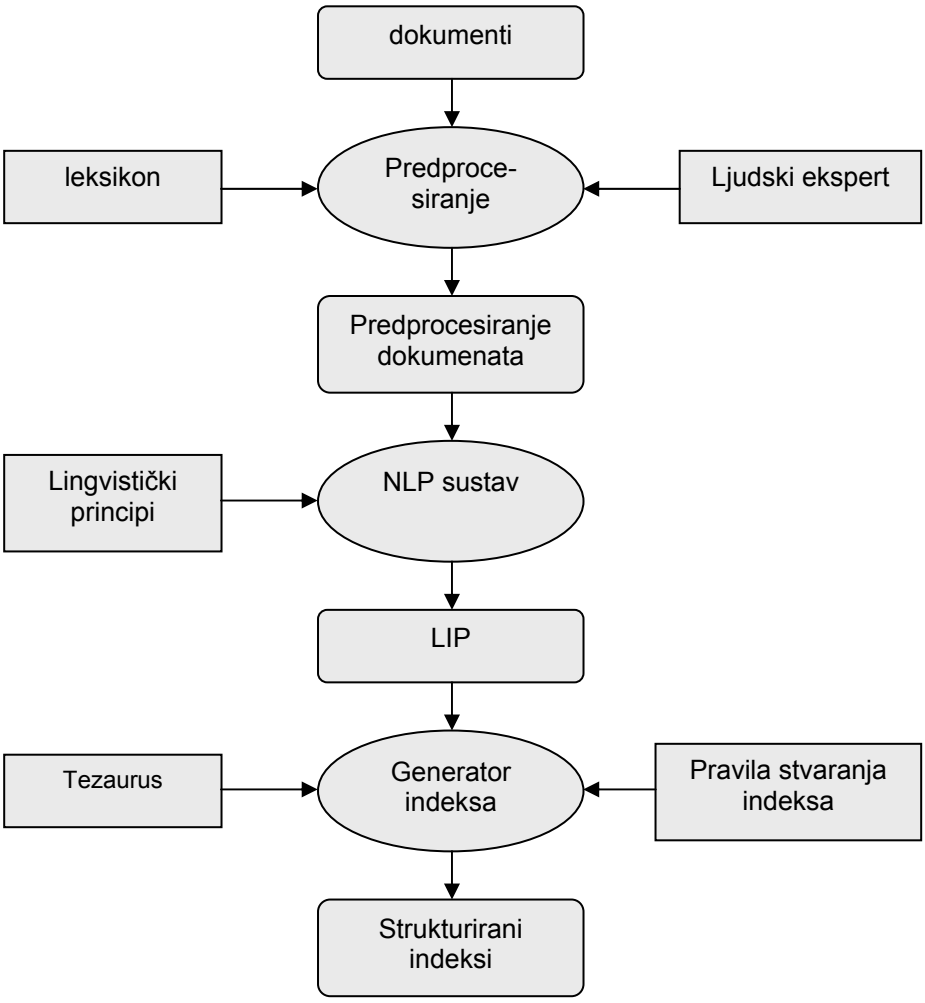
$$\vec{b} = W\vec{a} .$$

Izračunati vektor sadrži realne vrijednosti, ali određivanjem neke granice, on poprima binarne vrijednosti. Vrijednosti iznad praga postaju jednake 1, a one ispod 0. Vrijednosti postavljene u 1 određuju deskriptore koje je potrebno pridružiti izvještaju.

Ovaj sustav radi sa znatno manjim skupom deskriptora od skupova u slučajevima ostalih sustava, pa nije direktno usporediv sa ostalima. Procijenjeno je da je optimalna vrijednost ranije spomenutog praga 0.6. U tom slučaju prosječna preciznost iznosi 83.4% $\pm$ 5.3%, a odziv 35.4% $\pm$ 5.6% [32], [33].

6.5. CONDORCET

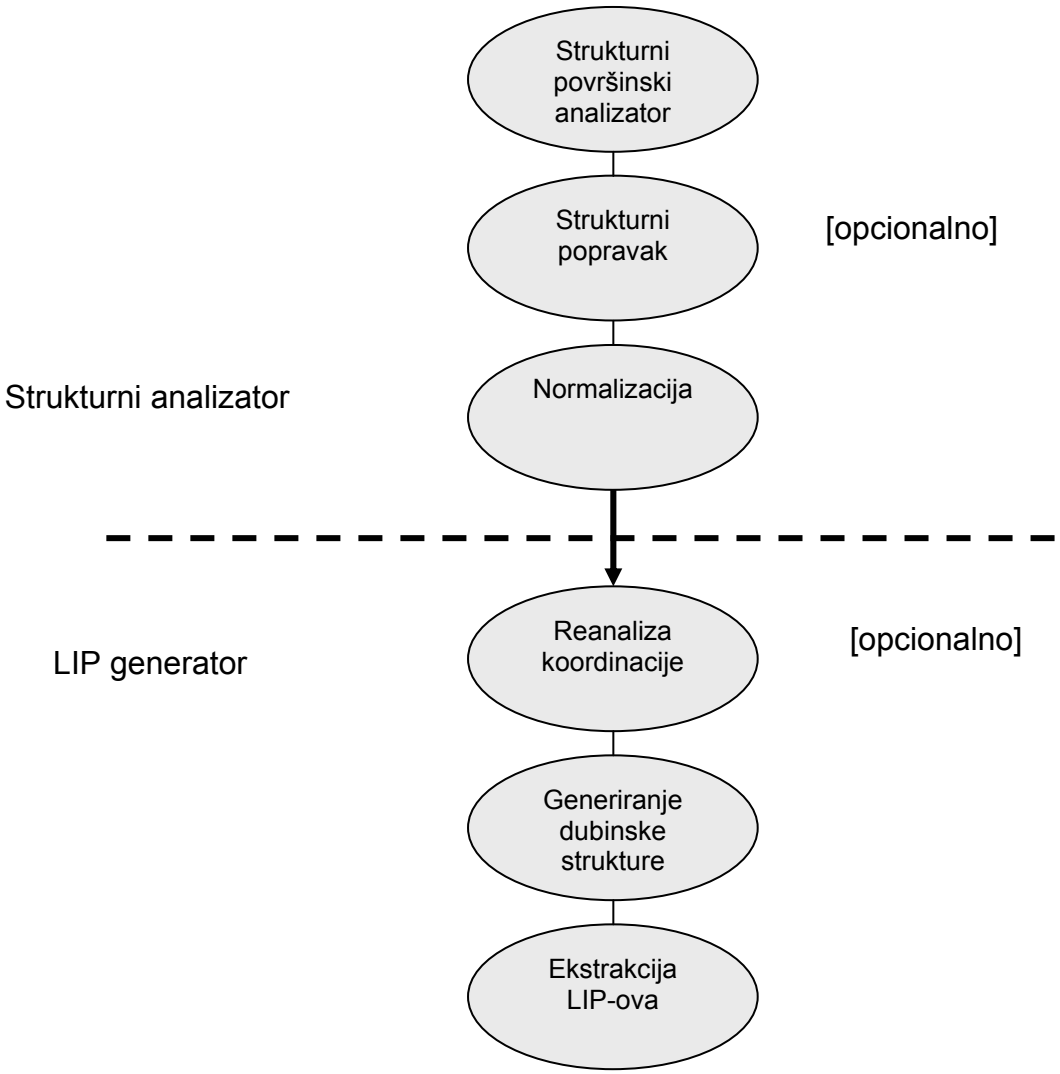
CONDORCET sustav se sastoji od više različitih faza, što omogućuje lakše održavanje i prilagodbu različitim jezicima i tipovima dokumenata. Tri osnovna dijela sustava su predprocesiranje, obrada prirodnog jezika i generiranje indeksa. Sustav je prikazan na slici 9.



Slika 9. Shema CONDORCET sustava

Podsustav za indeksiranje se sastoji od dvije faze. U prvoj se izvornim dokumentima dodaju SGML oznake, koje određuju dijelove dokumenta, kao što su na primjer naslov ili formule. Nakon toga, u drugoj fazi predprocesiranja se na temelju leksikona određuju vrste pronađenih riječi i sve dodatne informacije. Ovdje je omogućena i komunikacija s korisnikom, koji može sam upisati informacije o nepoznatim riječima.

Sustav za procesiranje prirodnog jezika (NLP sustav – eng. *Natural Language Processing*) također se može razdvojiti u dva bitna dijela. Prvi radi strukturnu analizu, dok drugi generira LIP-ove. LIP (eng. *Linguistic Information Package*) je struktura koja se sastoji od tri polja s bitnim lingvističkim informacijama. NLP sustav prikazan je detaljnije na slici 10.



Slika 10. Prikaz NLP sustava

Osnovna zadaća strukturnog površinskog analizatora je formiranje strukturnih fraza kombiniranjem oznaka za dijelove jezika, kao i spajanje tih fraza u strukture rečenica. To se obavlja pretvorbom ulaznih podataka u kanonsko stablo. U nekim slučajevima struktura ulaznih podataka onemogućuje ovakvu obradu, pa se primjenjuje strukturni popravak. U zadnjem koraku strukturne analize (normalizacija) provjerava se da li je stvarno generirana kanonska struktura, pogodna za daljnju obradu.

LIP-generator započinje svoju obradu dodajući nove implicitne veze strukturama, kako bi se poboljšala obrada. Nakon toga stvaraju se duboke strukture kako bi se izjednačile različite sintaksne varijacije sa istim značenjem. Napokon se u procesu ekstrakcije LIP-ova dobivaju strukture pogodne za generiranje indeksa.

Generator indeksa prima LIP-ove iz prethodnog procesa i pretraživanjem tezaurusa dodjeljuje pripadne deskriptore. On ima i dodatna svojstva kao što je uzimanje samo dijela LIP-a ili odbacivanje specifičnijih pojmova.

Testovi su pokazali veliku uspješnost sustava. Efektivnost sustava za jednostavne koncepte iznosi preko 90%, ali se ta vrijednost za strukturirane koncepte znatno smanjuje [34], [35].

6.6. Indeksiranje temeljeno na asocijacijama (Berkeley)

Na temelju INSPEC baze u MELVYL katalogu Sveučilišta u Kaliforniji stvoren je rječnik asocijacija leksičkih pojmova pronađenih u naslovima, imenima autora i sažecima sa ručno dodijeljenim kategorijama (indeksima, deskriptorima) INSPEC tezaurusa. Drugu fazu postupka čini indeksiranje novih dokumenata na temelju stvorenog rječnika.

Za testiranje su korišteni naslovi, imena autora i sažeci u različitim kombinacijama. Nešto bolje rezultate za preciznost i odziv daju testovi u koje su uključeni sažeci, dok imena autora nemaju većeg utjecaja. U najboljem slučaju preciznost iznosi 0.60, a odziv 0.20.

Veza između pojma u dokumentu i pridruženog deskriptora određuje se na temelju vjerojatnosti da se zajedno nađu u jednom dokumentu. Ako je vjerojatnost da se pojave zajedno veća od slučajnosti, znači da su međusobno povezani. Ta se zavisnost određuje pomoću metode omjera vjerodostojnosti, odnosno formulom:

$$\lambda = \frac{\max_{\omega} H(\omega,k)}{\max_{\Omega} H(\omega,k)} = \frac{\max_{p} H(p,p;k_1,n_1,k_2,n_2)}{\max_{p1p2} H(p_1,p_2;k_1,n_1,k_2,n_2)},$$

gdje je

$$H(p;k,n) = \binom{n}{k} p^k (1-p)^{n-k},$$

odnosno:

$$H(p_1,p_2;k_1,n_1,k_2,n_2) = \binom{n_1}{k_1} {p_1}^{k_1} (1-p_1)^{n_1-k_1} \binom{n_2}{k_2} {p_2}^{k_2} (1-p_2)^{n_2-k_2}$$

Hipoteza koja se testira označena je sa ω, a cijeli prostor parametara sa Ω. Broj pozitivnih primjera je *k*, a ukupan broj *n*. Očekivana vrijednost rezultata je *p*. Vrijednosti *k* i *n* računaju se na temelju kontigencijske tablice:

AB	A¬B
¬AB	¬A¬B

A označava pojam u tekstu, a *B* deskriptor, dok *¬A* i *¬B* označavaju odsutnost pojma, odnosno deskriptora. Za vrijednosti *k* i *n* vrijedi:

$$k1 = AB,$$

$$n1 = AB + \neg AB,$$

$$k2 = A\neg B,$$

$$n2 = A\neg B + \neg A\neg B.$$

Svaki dokument u skupu za učenje predstavljen je pomoću *m* pojmova *a* i *n* deskriptora *b*, $D_i = (\{a_{i1}, ..., a_{im}\}; \{b_{i1}, ..., b_{in}\})$. Na temelju svakog para $a_{im} \times b_{in}$ stvara se ili obnavlja kontigencijska tablica, a svaki par utječe na svaku tablicu. Nakon formiranja svih tablica računaju se međusobne zavisnosti pojmova i deskriptora. Dobiveni podaci predstavljaju rječnik asocijacija.

Za indeksiranje novog dokumenta, potrebno je za sve pojmove koji se u njemu javljaju, pronaći adekvatna pravila u rječniku asocijacija. Na taj način pronalaze se svi potencijalni deskriptori. Dokument se pretvara u vektor $D_i=(x_{i1}, ..., x_{in})$, gdje su x_{ij} težine deskriptora *j* za dokument *i*. Ako ima više pojmova za koje postoji veza u rječniku s deskriptorom j, njihove se težine zbrajaju i tako dobiva x_{ij}. Za konačne deskriptore uzima se određeni broj s najvećim težinama [36], [37].

6.7. NASA MAI

Sustav MAI (eng. *Machine-Aided Indexing*) je izrađen kako bi se ubrzao postupak indeksiranja znanstvenih i stručnih izvještaja, ali i smanjili troškovi. MAI je sustav namijenjen indeksatorima i svi rezultati se moraju ručno pregledavati.

MAI sustav se sastoji od tri osnovna dijela:

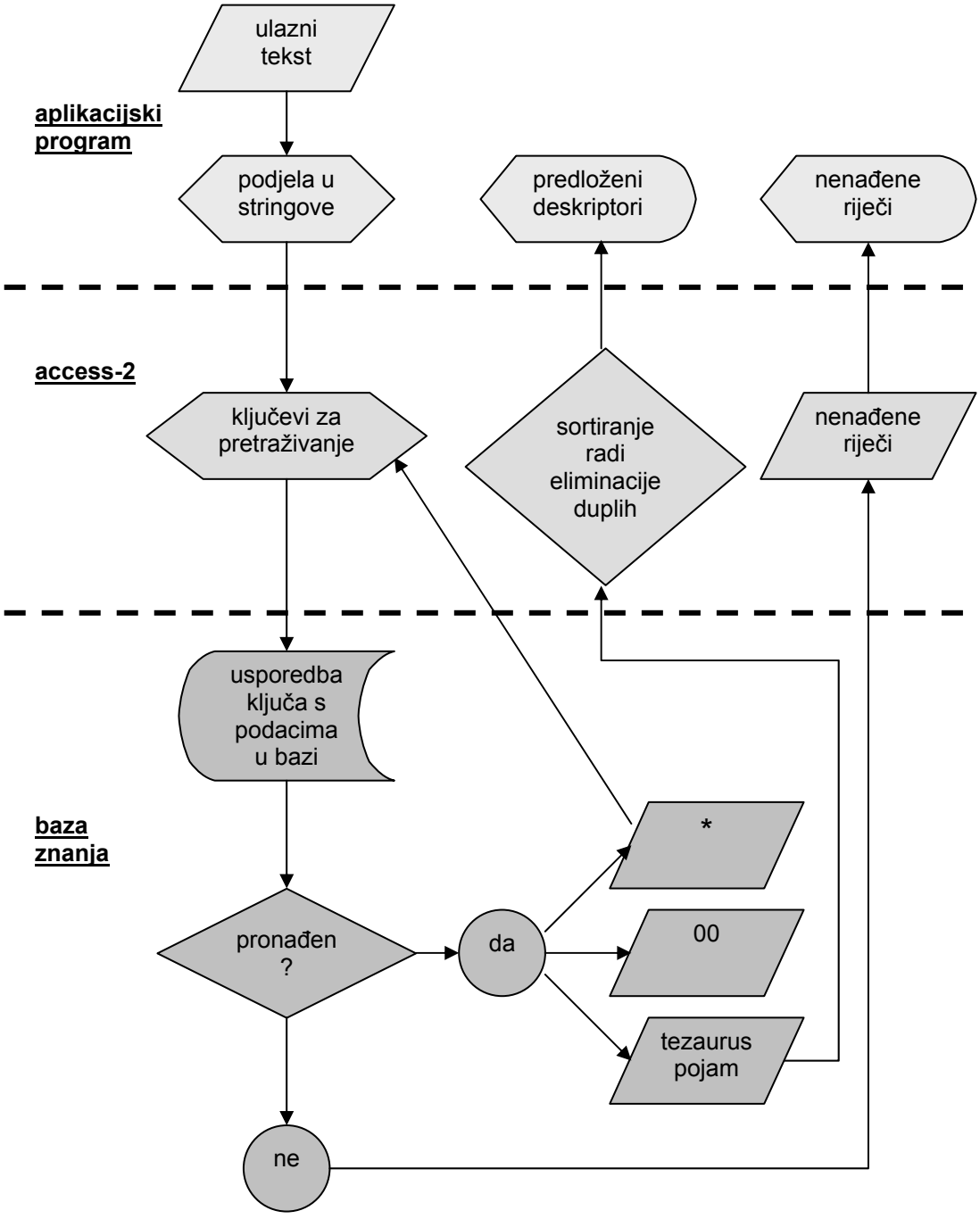
1. baza znanja
2. aplikacijski programi
3. *Access-2*, program koji slaže ključeve za pretraživanje spajanjem riječi, traži te ključeve u bazi znanja i vraća kandidate iz NASA tezaurusa.

Bazu znanja predstavlja skup podataka koji daje prijevod pojmova iz prirodnog jezika u kontrolirani rječnik. Dva osnovna polja su ključ i polje za pojmove. Polje za pojmove može sadržavati jedan ili više pojmova sa značenjem ekvivalentnim ključu, 00 ili *. Ako su u tom polju pronađeni pojmovi iz kontroliranog rječnika, oni će biti predloženi kao deskriptori. Ako je nađen znak *, ključu se dodaje sljedeća riječ i nastavlja pretraživanje. U slučaju da su pronađene dvije nule, ta riječ se ne prevodi. Sva su pravila organizirana na način da su ključevi sortirani tim redoslijedom koji osigurava da složeni izrazi imaju prednost pred pojedinačnim pojmovima.

Zadatak aplikacijskog programa je odrediti izvor teksta za procesiranje, odrediti pojmove u tekstu, pozvati *Access-2* sloj, primiti rezultat od sustava i prezentirati ga, odnosno stvoriti izvještaj.

Access-2 sloj prima riječi od aplikacijskog programa i stvara složene semantičke jedinice od više pojmova. Te fraze ili riječi pretražuju se u bazi znanja. Ako određena riječ nije nađena, takva poruka se šalje korisniku. U slučaju pronađenog pravila, *Access-2* vraća sortirane predložene deskriptore.

Rad sustava može se vidjeti na slici 11.



Slika 11. Shema NASA MAI sustava

U najboljem slučaju oko 50% deskriptora koje je preporučio MAI sustav izabrali su također i indeksatori (preciznost). Druga mjera evaluacije odnosi se na udio deskriptora koje je izabrao MAI među onima koje su dodijelili indeksatori (odziv). Slične vrijednosti su postignute i za ovu mjeru [38].

7. Indeksiranje pomoću Eurovoca

Kako bi se ubrzao postupak indeksiranja dokumenata i omogućilo višejezično pretraživanje razvijen je sustav za automatsko indeksiranje pomoću Eurovoca. Cijeli projekt pokrenuo je *Joint Research Centre* u Italiji.

Ovdje će biti opisan isti postupak uz promjenu određenih parametara kako bi se prilagodio hrvatskom jeziku.

7.1. Eurovoc

Dokumentacijski centri i knjižnice klasificiraju i indeksiraju svoju građu kako bi pomogli korisnicima u pretraživanju i pronalaženju željenih dokumenata i publikacija. Mnoge parlamentarne institucije u Europi koriste isti višejezični tezaurus Eurovoc.

Eurovoc je organiziran hijerarhijski i sadrži oko 6000 pojmova, a dostupan je u 21 različitih jezika. Podijeljen je na 21 područje, a na sljedećem nivou na 127 mikro tezaurusa. Podjela se nastavlja do osme razine. Eurovoc je konceptualni tezaurus, što znači da su pojmovi prilično općeniti i opisuju osnovne koncepte iz danog područja, koje obuhvaća sve djelatnosti Europske unije i sve važne pojmove koji se javljaju u najrazličitijim dokumentima EU-a. Za potrebe jednoznačne i precizne sadržajne obrade službene dokumentacije, Eurovoc je također preveden na hrvatski jezik.

Pojmovi, odnosno deskriptori, mogu biti povezani na dva načina: kao širi i uži pojam ili kao srodni pojmovi, koji povezuju hijerarhijski nezavisne deskriptore. Ovaj tezaurus sadrži i nedeskriptore, koji pomažu indeksatorima u pronalaženju odgovarajućih deskriptora [4].

Hijerarhijski prikaz Eurovoca dan je na slici 12. [39]

04 politika

0406 politički okvir

0411 politička stranka

0416 izborni postupak i glasovanje

0421 parlament

0426 rad parlamenta

0431 politika i javna sigurnost

0436 izvršna vlast i javne službe

0426 rad parlamenta

parlamentarni postupak

- glasovanje u parlamentu
  - delegirano glasovanje
  - elektroničko glasovanje
  - glasačka stega
  - glasovanje za tekst u cjelini
  - kvorum
  - poimenično glasovanje
- javna rasprava
- parlamentarna sjednica
  - dnevni red
  - materijal za sjednicu
    - odborski izvještaj
  - objašnjenje glasovanja
  - parlamentarna rasprava
  - vrijeme za zastupnička pitanja
- parlamentarno zasjedanje
- poslovnik parlamenta
- prekid sjednice

zakonodavni postupak

- amandman
- donošenje zakona
  - usvajanje zakona glasovanjem
- izmjena zakona
- izrada nacrta prijedloga zakona
- mišljenje
- objava zakona
- proglašenje zakona
- zakonodavni poticaj
  - nevladin prijedlog zakona
  - vladin prijedlog zakona

Slika 12. Hijerarhijski prikaz Eurovoca

7.2. Razlozi za indeksiranje tezaurusom

Pošto se radi o višejezičnom tezaurusu, gdje je svaki deskriptor preveden na točno jedan način na sve ostale jezike, postavljanjem upita na bilo kojem od ponuđenih jezika, dobivaju se ukupni rezultati neovisni o jeziku. Drugim riječima, postavljanjem upita na hrvatskom, kao rezultat se dobivaju također i dokumenti odgovarajućeg sadržaja pisani engleskim ili nekim drugim jezikom Europske unije. Ovo svojstvo ostvareno je pomoću jedinstvenih identifikacijskih brojeva pridruženih svakom deskriptoru, čime se dobiva jezična nezavisnost.

Osim navedene prednosti da je takav sustav koji koristi indeksiranje pomoću višejezičnog tezaurusa automatski višejezičan, dodatne pogodnosti su sljedeće:

1. Dokumenti vraćeni kao rezultat pretraživanja su uvijek bitni i usko vezani za pojam po kojem se pretraživalo, jer su u postupku dodjele deskriptora odabiru samo oni najvažniji.
2. Upiti se proširuju na način da se pretraživanjem po nekom općenitom pojmu, zbog hijerarhijske organizacije pojmovnika mogu također dobiti i dokumenti vezani uz neki njemu podređen i specifičniji.
3. Popis deskriptora pridijeljenih pojedinom dokumentu može se promatrati kao neka vrsta sažetka, jer daje uvid u najbitnije teme pokrivene u dokumentu.

Ručno dodjeljivanje deskriptora izrazito je vremenski zahtjevan i skup posao. Čak i profesionalni indeksatori mogu indeksirati mali broj dokumenata dnevno, pa se javlja potreba za uvođenjem automatskih ili barem poluautomatskih rješenja [4].

8. Algoritam

Indeksiranje pomoću tezaurusa se može ostvariti pomoću dvije osnovne skupine metoda: lingvističkih metoda temeljenih na pravilima i statističkih metoda. Metode temeljene na pravilima sastoje se od desetaka tisuća pravila koja povezuju pojavljivanja ili odsustva pojedinih pojmova. Stvaranje te liste pravila vrlo je naporan i nezahvalan posao, jer se postupak mora ponavljati za svaki jezik pojedinačno. Iako će sljedeći eksperimenti biti rađeni isključivo za hrvatski jezik, Eurovoc postoji i na drugih 20 jezika za koje bi isti ili sličan postupak trebao biti primjenjiv. Kao rezultat toga, metode temeljene na pravilima su gotovo potpuno neupotrebljive, pa će dalje biti razmatrane isključivo statističke metode.

Ranije su spomenute dvije vrste određivanja ključnih riječi: ekstrakcija i dodjeljivanje ključnih riječi. Deskriptori u Eurovocu su dosta općeniti i korištena je specifična terminologija koja se rijetko javlja u samom tekstu dokumenata. Osim toga, može se dogoditi da se riječ ili niz riječi prisutnih u tekstu nalaze kao deskriptori u tezaurusu, ali s obzirom na razliku u značenju nisu tom tekstu dodijeljeni kao deskriptori.Ta činjenica znatno otežava indeksiranje Eurovocom i očito je da jedino dodjeljivanje ključnih riječi dolazi u obzir.

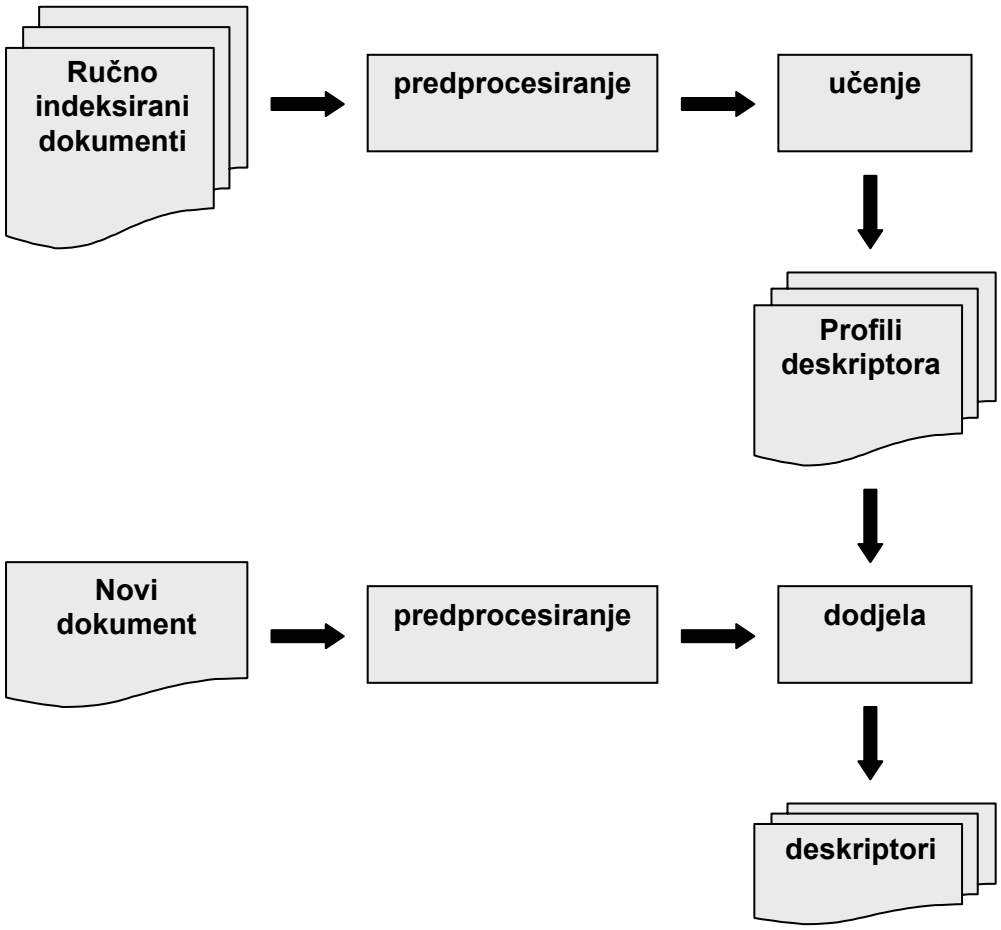
Prilikom dodjeljivanja deskriptora Eurovoca svakom dokumentu može se pridružiti bilo koji od 6075 deskriptora. Radi se zapravo o višestrukoj kategorizaciji, jer svaki dokument može biti opisan proizvoljnim brojem deskriptora. Radi eliminacije nedovoljno značajnih tema vezanih uz pojedini dokument i održavanja liste pripadnih deskriptora donekle sažetom, broj dodijeljenih deskriptora se ipak nastoji smanjiti.

Pošto se radi o problemu klasifikacije, mogle bi biti primijenjene mnoge od uobičajenih metoda. Ipak, mora se uzeti u obzir količina podataka koja mora biti obrađena, a posao se treba obaviti u nekom realnom vremenu.

Zato je potrebno koristiti što jednostavniji i računarski ne naročito složen postupak za dodjelu deskriptora. On se sastoji od tri bitne faze. Nakon uobičajenog predprocesiranja tekstova, stvaraju se profili deskriptora. Oni predstavljaju listu, odnosno vektor riječi iz samog teksta koji najbolje opisuju pojedini deskriptor. Svaki novi dokument koji je potrebno indeksirati prvo se prikazuje na isti način kao i deskriptori, a nakon toga računa sličnost sa svim profilima. Na temelju vrijednosti izračunatih sličnosti vektora, određuju se prikladni deskriptori [4].

Ovdje se radi o sustavu za nadzirano strojno učenje, što znači da je potreban skup ručno indeksiranih dokumenata na temelju kojih sustav indeksira ostale. Taj skup se još naziva i skup za učenje. U tom se skupu mora nalaziti barem nekoliko dokumenata koji su indeksirani određenim deskriptorom, kako bi mogli biti stvoreni odgovarajući profili. Ako neki deskriptor nije dovoljno zastupljen u skupu za učenje, on se također neće niti moći dodjeljivati novim dokumentima. Skup ručno indeksiranih dokumenata, koji služe za provjeru uspješnosti rada sustava zove se skup za testiranje.

Vizualni prikaz cijelog procesa dan je na slici 13. [40]



Slika 13. Vizualni prikaz indeksiranja Eurovocom

8.1. Predprocesiranje

U većini slučajeva kao najuspješnije metode predprocesiranja pokazali su se eliminacija stop riječi, morfološka normalizacija i formiranje n-grama. Radi jednostavnosti postupka, u sljedećim eksperimentima će se koristiti jedino lista stop riječi. Ona je formirana na temelju hrvatskog jezika, a ne samo skupa dokumenata s kojima se provode eksperimenti.

8.2. Odabir pojmova za asocijate

Kod odabira pojmova koji predstavljaju asocijate za izgradnju profila deskriptora, ukupni skup pojmova koji se nalazi u dokumentu potrebno je reducirati i odrediti samo one pojmove koji su karakteristični za taj tekst. Za odabir atributa koristi se metoda omjera log-vjerodostojnosti ili sama frekvencija dokumenata.

Ako se koristi omjer log-vjerodostojnosti, uspoređuju se dva korpusa. Kao korpus za usporedbu uzima se cijeli skup dokumenata za učenje, a korpus koji se uspoređuje predstavlja sam dokument čije je asocijate potrebo odrediti. Za oba korpusa se računaju frekvencije pojmova.

Za svaki pojam se određuje omjer log-vjerodostojnosti. Granica prihvatljivih pojmova definirana je pomoću proporcije *p* i stupnjeva slobode. Pronalaženjem tih vrijednosti u tablici χ^2 kritičnih vrijednosti [41] određuje se donja granica za omjer log-vjerodostojnosti. Pojmovi s vrijednostima iznad te granice se uzimaju kao asocijati, dok se drugi odbacuju.

Odabir pojmova se može ostvariti i samo na temalju njihove frekvencije. Jedna mogućnost je promatrati frekvenciju pojma u skupu

dokumenata koji su indeksirani određenim deskriptorom, a druga je da se uzme u obzir ukupna frekvencija nekog pojma u cijelom korpusu.

Moguće je odrediti i dodatni uvjet koji osigurava kada neki pojam iz teksta može postati asocijat dodijeljenog deskriptora. To je minimalan broj tekstova indeksiranih odgovarajućom deskriptorom koji sadrže pojam u tekstu.

Primjer odabira pojmova za asocijate

Može se uzeti da tablica potrebna za određivanje omjera log-vjerodostojnosti izgleda kao tablica 7, uz napomenu da vrijednosti nisu realne, nego su dane samo kao primjer.

Tablica 7. Primjer odabira asocijata

	Korpus 1	Korpus 2	Ukupno
Frekvencija pojma	a=10	b=90	a+b=100
Frekvencija ostalih pojmova	c-a=300	d-b=1000	c+d-a-b=1300
Ukupno	c=310	d=1090	c+d=1400

Očekivane vrijednost računaju se prema formuli:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

Iz toga se dobiva da je :

$$E_1 = \frac{N_1 \sum_i O_i}{\sum_i N_i} = \mathrm{c} \cdot \frac{a+b}{c+d} = 310 \cdot \frac{10+90}{310+1090} = 22.14$$

$$E_2 = \frac{N_2 \sum_i O_i}{\sum_i N_i} = \mathrm{d} \cdot \frac{a+b}{c+d} = 1090 \cdot \frac{10+90}{310+1090} = 77.86$$

Omjer log-vjerodostojnosti se računa na sljedeći način:

$$LL = 2 \cdot \sum_i O_i \ln\left(\frac{O_i}{E_i}\right) = 2 \cdot \left(a \cdot \ln\left(\frac{a}{E_1}\right) + b \cdot \ln\left(\frac{b}{E_2}\right)\right) =$$

$$= 2 \cdot \left(10 \cdot \ln\left(\frac{10}{22.14}\right) + 90 \cdot \ln\left(\frac{90}{77.86}\right)\right) = 10.18$$

Proporcija *p* može se postaviti na vrijednost 0.15, a broj stupnjeva slobode određen je formulom:

$$d.f. = (r-1)(c-1) = (2-1)(2-1) = 1.$$

Iz tablice χ^2 kritičnih vrijednosti može se očitati da kritična vrijednost iznosi 2.07. Pošto vrijedi da je 10.18 > 2.07 promatrani pojam može se uzeti kao asocijat.

8.3. Profili deskriptora

Za svaki deskriptor odabiru se dokumenti kojima je on dodijeljen i pronalaze riječi koje se smatraju karakterističnе za taj podskup dokumenata. Učenje se ne može obaviti na samo jednom ili vrlo malom broju primjera, pa je potrebno odrediti minimalan broj dokumenata po deskriptoru, da bi se proces učenja uopće isplatio. U slučaju manjeg broja dokumenata od minimalnog, sustav bi se specijalizirao samo za dokumente vrlo slične tim postojećim i profil tog deskriptora zapravo bi postao profil dotičnog dokumenta. Korištenjem tog minimalnog praga, znatno se poboljšava postupak učenja, ali su zato mnogi deskriptori zanemareni. U fazi dodjeljivanja deskriptora moći će se koristiti samo oni čiji profili su uspješno stvoreni. Za povećanje skupa stvorenih profila bilo bi potrebno proširiti skup dokumenata za učenje, ali i provesti kvalitetnije ručno indeksiranje, kako bi svi deskriptori bili dovoljno zastupljeni.

Profil svakog deskriptora predstavljen je pomoću vektora. Taj vektor se sastoji od karakterističnih pojmova, asocijata i pridijeljenih težina. Težine se računaju na sljedeći način:

$$Weight_{l,d} = W_{l,d} \cdot IDF_l ,$$

gdje $Weight_{l,d}$ predstavlja težinu leme l kao asocijata deskriptora d. $W_{l,d}$ definiran je kao:

$$W_{l,d} = \sum_{t \in T_{l,d}} \frac{1}{Nd_t} ,$$

gdje t predstavlja tekst, Nd_t broj ručno dodijeljenih deskriptora za tekst t, a $T_{l,d}$ skup tekstova koji su ručno indeksirani pomoću deskriptora d i sadrže lemu l. Ovdje se radi o normalizaciji težine, a ona se obavlja prema broju drugih

deskriptora pridruženih istom dokumentu. Cilj je smanjenje težina onim deskriptorima koji su samo jedni u nizu deskriptora pridruženih nekom dokumentu u odnosu na one koji su na primjer jedini pridruženi.

IDF_l predstavlja inverznu frekvenciju deskriptora i računa se kao:

$$IDF_l = \log\left(\frac{Max_{DF_l}}{\beta \cdot DF_l} + 1\right),$$

gdje DF_l označava frekvenciju deskriptora, odnosno broj deskriptora u kojima se lema *l* pojavljuje kao asocijat. $MaxDF_l$ je maksimalna vrijednost frekvencije deskriptora za sve leme. Dijeljenjem sa frekvencijom deskriptora smanjuje se utjecaj onih lema koje su asocijati mnogim deskriptorima. Njihov je utjecaj još više oslabljen korištenjem parametra β. Postavljenjem njegove vrijednosti na 10 nestaje utjecaj lema koje se pojavljuju barem 10% onoliko često kao najučestalija ($MaxDF_l$).

Konačna formula za računanje težine glasi:

$$Weight_{l,d} = \left(\sum_{t \in T_{l,d}} \frac{1}{Nd_t}\right) \cdot \log\left(\frac{Max_{DF_l}}{\beta \cdot DF_l} + 1\right).$$

Primjer stvaranja profila

Ako se, na primjer, želi stvoriti profil za deskriptor "buka", potrebno je najprije naći sve dokumente koji su indeksirani tim deskriptorom. Radi jednostavnosti primjer će biti prikazan samo na temelju riječi koje se javljaju u naslovu dotičnih dokumenata, a analogan postupak se može provesti i nad cijelim tekstom dokumenta. Dokumenti indeksirani deskriptorom "buka" su sljedeći:

1.	Zakon o zaštiti od buke
2.	Naredba o homologaciji mopeda s dva kotača u pogledu emisije buke
3.	Naredba o homologaciji motornih *vozila* s tri kotača u pogledu emisije buke
4.	Ispravak Naredbe o homologaciji motornih *vozila* s tri kotača u pogledu emisije buke
5.	Naredba o homologaciji zamjenskih sustava za smanjenje buke
6.	Pravilnik o najvišim dopuštenim razinama buke u sredini u kojoj ljudi rade i borave
	…

Za svaki dokument kreira se lista asocijata i oni se pridjeljuju deskriptoru. Na primjer:

	Dokument 1.
1.	Zakon
2.	Zaštiti
3.	Buke

	Dokument 2.
1.	naredba
2.	homologaciji
3.	mopeda
4.	kotača
5.	pogledu
6.	emisije
7.	buke

	Dokument 5.
1.	naredba
2.	homologaciji
3.	zamjenskih
4.	sustava
5.	smanjenje
6.	buke

Na isti način se obrađuju i ostali dokumenti i kreira lista asocijata za deskriptor:

	Deskriptor: "buka"
1.	zakon
2.	zaštiti
3.	buke
4.	naredba
5.	homologaciji
6.	mopeda
7.	kotača
8.	pogledu
9.	emisije
10.	zamjenskih
11.	sustava
12.	smanjenje
	...
	vozila
	...

Prema prethodno spomenutoj formuli računaju se težine pojedinih asocijata. Ovdje će biti prikazano računanje težine za asocijat "vozila". U primjeru se radi sa tekstovima koji nisu prošli proces morfološke normalizacije.

Pronađena su samo dva dokumenta opisana deskriptorom "buka", koji sadrže pojam vozila. Oba dokumenta imaju po tri pridijeljena deskriptora, pa *Ndt* za svaki dokument poprima vrijednost 3:

$$W_{vozila,buka} = \sum_{t \in T_{vozilabuka}} \frac{1}{Nd_t} = \frac{1}{3} + \frac{1}{3} = 0.667$$

Pojam vozila je asocijat za čak 93 deskriptora, dok maksimalan broj pojavljivanja nekog pojma kao asocijata iznosi 329:

$$IDF_{vozila} = \log\left(\frac{Max_{DF_l}}{\beta \cdot DF_{vozila}} + 1\right) = \log\left(\frac{329}{10 \cdot 93} + 1\right) = 0.1315$$

Pomoću tih vrijednosti se dobiva konačna težina:

$$Weight_{vozila,buka} = W_{vozila,buka} \cdot IDF_{vozila} = 0.667 \cdot 0.1315 = 0.08770$$

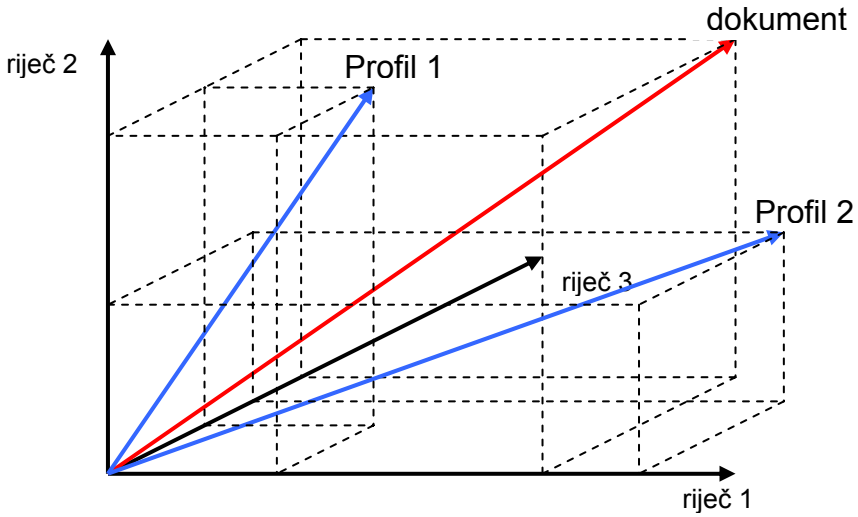
Kao rezultat se dobiva vektor koji sačinjavaju elementi sljedeće tablice:

	asocijat	težina
1.	buke	3.34471
2.	emisije	3.11539
3.	pogledu	2.56344
4.	smanjenje	1.24106
5.	kotača	0.914471
6.	motocikla	0.827375
7.	mopeda	0.675171
8.	sustava	0.457235
9.	motornih	0.269636
10.	**vozila**	0.08770
11.	…	

8.4. Dodjeljivanje deskriptora

Nakon završenog procesa učenja, novi dokument koji se želi indeksirati također se prikazuje kao vektor. Tekst je potrebno predprocesirati na isti način kao skup za učenje, a nakon toga računaju se frekvencije lema. Lista lema sa pripadnim težinama koja je proporcionalna učestalosti pojavljivanja određene leme u tekstu čini vektor novog dokumenta. Računanjem sličnosti između dva vektora koji prikazuju novi dokument i profil svakog deskriptora pojedinačno, određuju se pripadni deskriptori. Nekoliko najsličnijih profila novom dokumentu su reprezentacija deskriptora koje je potrebno pridijeliti.

Vektori profila i dokumenta mogu se prikazati i usporediti u istom vektorskom prostoru kao na slici 14.



Slika 14. Vektorski prikaz profila i dokumenta

Za računanje sličnosti vektora korištene su sljedeće mjere i kombinirane u različitim omjerima:

1. Kosinus

$$Cos_{l,d} = \frac{\sum_{l \in t \cap d} TFIDF_{l,d} \cdot TF_{l,t}}{\sqrt{\left(\sum_{l \in d} TFIDF_{l,d}{}^2\right) \cdot \left(\sum_{l \in t} TF_{l,t}{}^2\right)}}$$

2. Okapi formula koja u ovom slučaju poprima jednostavniji oblik [40]:

$$Okapi_{t,d} = \sum_{l \in t \cap d} \log\left(\frac{N - DF_l}{DF_l}\right) \frac{TF_{l,d}}{TF_{l,d} + \frac{|d|}{M}}$$

3. Skalarni produkt vektora

$$Skal_{l,d} = \sum_{l \in t \cap d} TFIDF_{l,d} \cdot TF_{l,t}$$

u navedenim formulama se umjesto frekvencije pojam u dokumentu ($TF_{l,t}$) može koristiti i samo pojavljivanje riječi, čime se dobiva binarni vektor.

Primjer dodjele deskriptora

Novi dokument kojem je potrebno pridružiti deskriptore je sljedeći:

Naputak o izmjenama i dopunama Naputka za provođenje postupka homologacije vozila.

Za njega se gradi vektor prikazan sljedećom tablicom, tako da su težine određene frekvencijom pojmova:

	Pojam	Težina
1.	naputak	1
2.	izmjenama	1
3.	dopunama	1
4.	naputka	1
5.	provođenje	1
6.	postupka	1
7.	homologacije	1
8.	vozila	1

Ako se kao mjera sličnosti uzme skalarni produkt, sličnost između novog dokumenta i profila deskriptora buka će se računati ovako:

$$SkalP = \sum_{i=1}^{n} x_i y_i = 1 \cdot 0.08770 = 0.08770$$

Jedini pojam koji se pojavljuje u oba vektora je «vozila». Njegova težina kod profila je 0.08770, a kod novog dokumenta 1. Kod svih ostalih pojmova, barem je jedan od faktora 0, pa nisu niti navedeni.

Provođenjem istog postupka sa svim ostalim profilima vidi se da su najveće sličnosti dobivene sa deskriptorima:

Potvrđivanje, Vozilo i *Motorno vozilo*

Ovdje je kao kriterij odabira deskriptora uzet prag sličnosti u iznosu 3, ali moguće je koristiti i različite druge kriterije.

9. Implementacija

Početni podaci su pomoću različitih konverzija pretvoreni u oblik pogodan za daljnju obradu. Skup dokumenata podijeljen je na skupove za učenje i testiranje, koji predstavljaju ulaz u sustav opisan u ovom poglavlju.

9.1. Ulazni podaci

9.1.1. Opis podataka

Podaci potrebni za rad sustava su sami dokumenti i podaci o ručno dodijeljenim deskriptorima. Radi se o zakonima Republike Hrvatske iz različitih područja. Samo kao primjeri dani su naslovi sljedećih dokumenata:

Odluka o visini novčane naknade za nezaposlenu osobu
Deskriptori: *nezaposlena osoba, socijalna pomoć, novac*

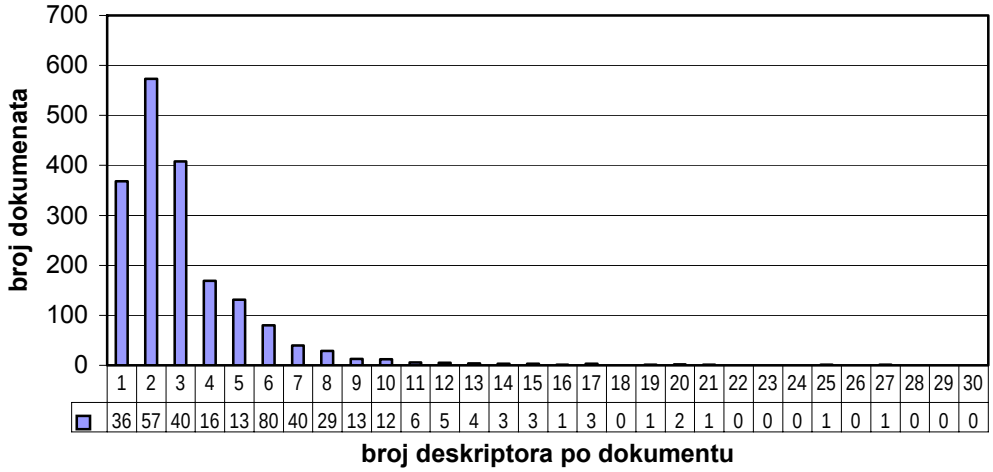
Naredba o vremenu otvorenosti aerodroma za javni zračni promet
Deskriptori: *zračna luka, zračni promet, radno vrijeme*

Pravilnik o osnovnim popisnim obrascima
Deskriptori: *popis, obrazac, popis stanovništva*

Zakon o plaćama u javnim službama
Deskriptori: *plaća, javna služba*

Odluka o gradskim porezima i prirezu poreza na dohodak
Deskriptori: *Zagreb, prirez, porez na dohodak*

Postoje velike razlike u broju dodijeljenih deskriptora, koji se kreće između 1 i 27. Na sljedećem grafu se može vidjeti raspodjela broja dodijeljenih deskriptora po dokumentima, a u prosjeku je to broj 3.



Slika 15. Broj deskriptora po dokumentima

Deskriptori koji se javljaju u najvećem broju dokumenata su sljedeći:

Porez na dohodak → 138 dokumenata
Prirez → 112 dokumenata i
Porez → 108 dokumenata.

U podacima se pojavljuje ukupno 1531 različiti deskriptor. Od toga se samo njih 322 javlja kod 5 dokumenata ili više, dok ostali nisu korišteni niti u jednom eksperimentu.

9.1.2. Format podataka

Svi podaci nalaze se u XML formatu zbog čega su bile potrebne različite konverzije.

Zapis strukture jednog dokumenta prikazan je na slici 15.

```
<RECORD>
<Naslov>Pravilnik o kalendaru rada osnovnih škola za školsku
2001./2002. godinu</Naslov>
<PRM><PRM_3>003084</PRM_3><PRM_a>osnovnoškolsko
obrazovanje</PRM_a></PRM>
<PRM><PRM_3>002970</PRM_3><PRM_a>planiranje školske
godine</PRM_a></PRM>
<PRM><PRM_3>000364</PRM_3><PRM_a>nastava</PRM_a></PRM>
<Naziv_dokumenta>2001_0349.xml</Naziv_dokumenta>
<Id_zapisa>01002398</Id_zapisa>
</RECORD>
```

Slika 16. Zapis dokumenta u XML-u

Većina oznaka je jasno iz samog imena, dok PRM_3 označava identifikacijski broj deskriptora, a PRM_a njegov puni naziv.

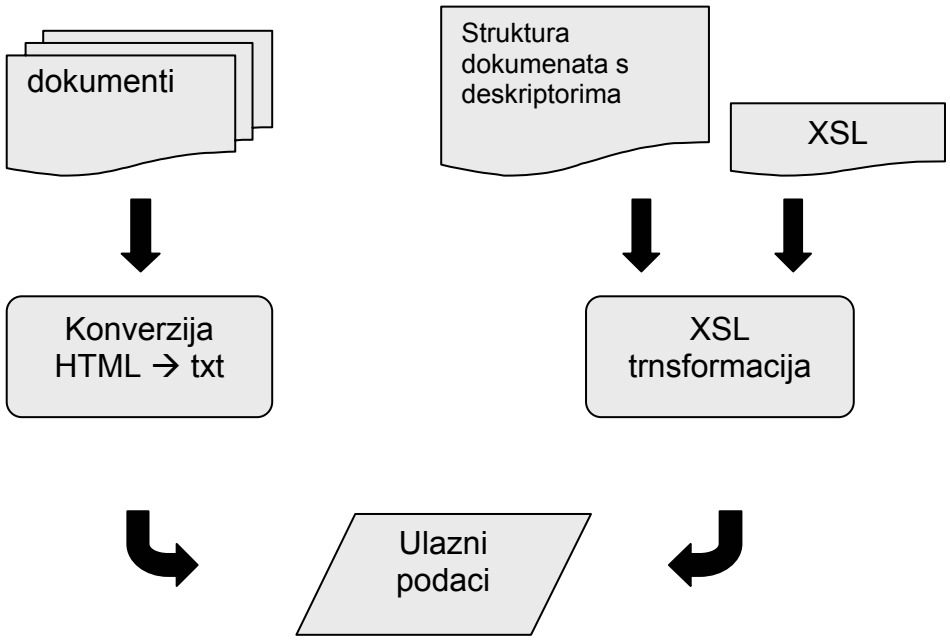
U većini eksperimenata su bili potrebni jedino nazivi dokumenata i pripadni deskriptori. Zato se početna struktura pomoću XSL transformacija prikazanih na slici 17 pretvorila u oblik:

naziv dokumenta, deskriptor1+deskriptor2+.....+deskriptorN.

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
   <xsl:template match="/">
        <xsl:for-each select="DATABASE_BIB/RECORD">
           <xsl:value-of select="Naziv_dokumenta"/>,
           <xsl:for-each select="PRM">
              <xsl:value-of select="PRM_a"/>
              <xsl:if test="not(position()=last())">+</xsl:if>
           </xsl:for-each>
           <xsl:text disable-output-escaping="yes">&#xA;</xsl:text>
        </xsl:for-each>
   </xsl:template>
</xsl:stylesheet>
```

Slika 17. XSL transformacija početne strukture

Proces pripreme ulaznih podataka prikazan je na slici 18.



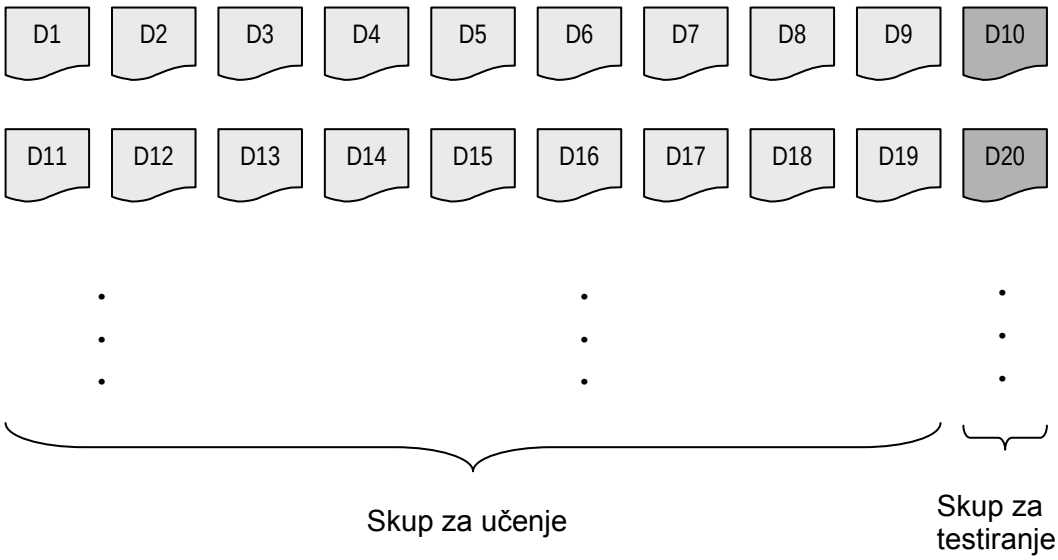
Slika 18. Priprema ulaznih podataka

9.1.3. Podjela u skupove za učenje i testiranje

Cijeli skup se sastoji od oko 1800 ručno indeksiranih dokumenata. Od toga je 90% uzeto kao skup za učenje, dok je preostalih 10% korišteno za testiranje, kao na slici 19.

Slika 19. Podjela ulaznih podataka

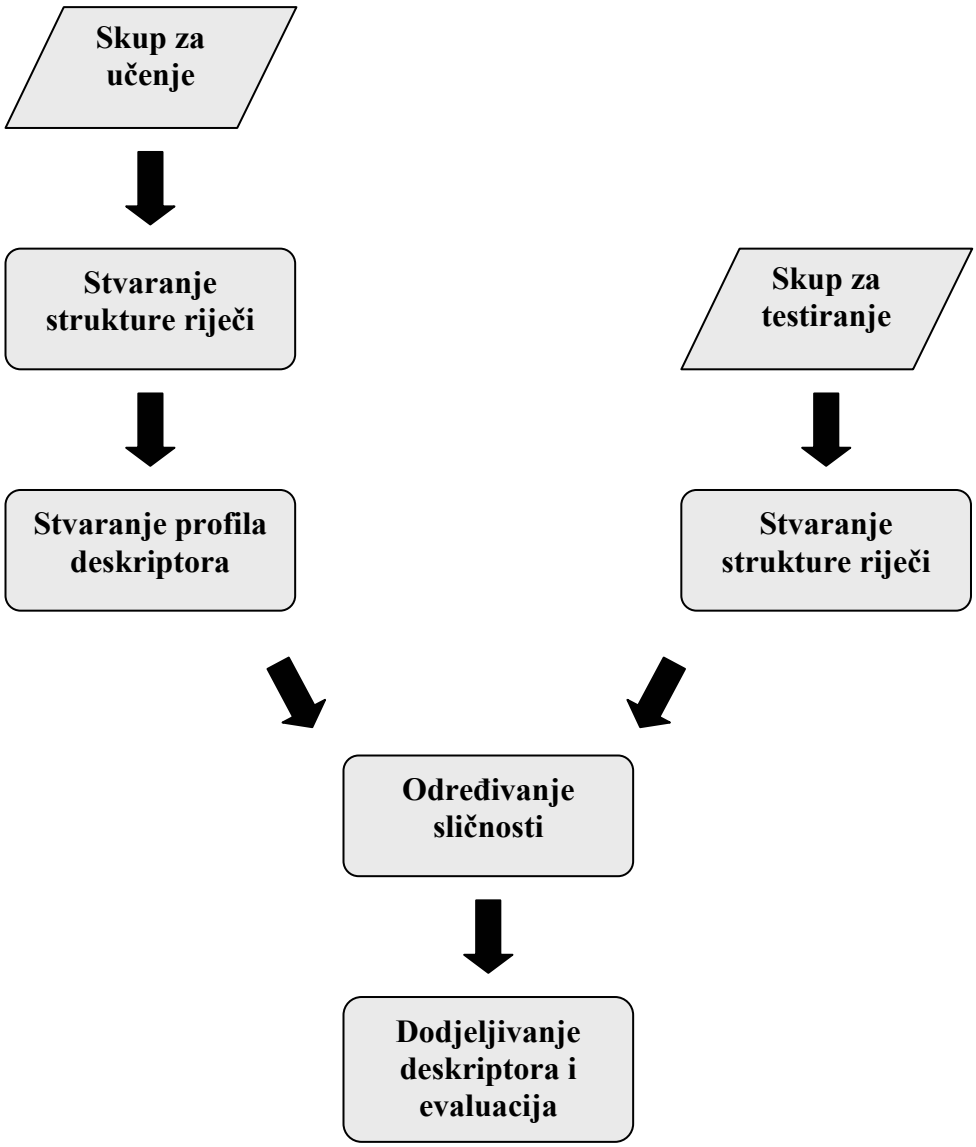
Skup za testiranje određen je na način da je izdvojen svaki deseti dokument, kao na slici 20, dok su preostali korišteni za proces učenja.



Slika 20. Odabir skupova za učenje i testiranje

9.2. Opis sustava

Sustav za indeksiranje dokumenata se sastoji od četiri osnovna dijela, kao što je i prikazano na slici 21. To su: stvaranje strukture riječi, stvaranje profila deskriptora, određivanje sličnosti i dodjeljivanje deskriptora s evaluacijom. Ulazi u sustav su već ranije opisani skup za učenje i testiranje.



Slika 21. Prikaz rada sustava

9.2.1. Stvaranje strukture riječi

U prvoj fazi postupka se iz početnih dokumenata stvaraju strukture sa potrebnim podacima. Nakon izbačenih stop riječi, za preostale se traži njihova frekvencija pojavljivanja u cijelom skupu za učenje. Osim samih riječi, izlaz iz ovog dijela programa se sastoji i od liste dokumenata. Svakom dokumentu pridružene su sve riječi, odnosno njihovi indeksi iz prethodno spomenute liste, koje su u njemu sadržane, kao i frekvencija svake pojedine riječi u samom dokumentu.

Na isti način obrađuje se i skup dokumenata za testiranje, koji je korišten za evaluaciju postupka.

9.2.2. Stvaranje profila deskriptora

U drugoj fazi se uz mogućnost podešavanja različitih parametara, kao što su minimalni broj dokumenata po deskriptoru, način odabira asocijata i minimalni broj dokumenata u kojima se asocijat mora pojaviti, stvara lista deskriptora uz pripadne asocijate. Svakom asocijatu pridružena je težina, a za potrebe okapi formule i ranije opisane DF_l i $TF_{l,d}$ vrijednosti.

9.2.3. Određivanje sličnosti

Dokumenti se, kao i profili deskriptora, pretvaraju u vektore. U ovoj fazi se na temelju jedne od ranije spomenutih metoda ili njihove kombinacije određuje sličnost svakog dokumenta za testiranje sa svim profilima.

9.2.4. Dodjeljivanje deskriptora i evaluacija

Na temelju sličnosti određene u prethodnom koraku postupka i zadanog praga određuju se deskriptori koji trebaju biti pridijeljeni određenom dokumentu.

Na temelju dobivenih rezultata može se odrediti uspješnost postupka u obliku preciznosti, odziva i f1 mjere uz mikro ili makro usrednjavanje.

10. Eksperimenti i rezultati

U sustavu opisanom u prethodnom poglavlju moguće je podešavati različite parametre. Zbog velikih vremenskih zahtjeva nisu testirane sve kombinacije, nego samo one za koje se moglo pretpostaviti da će dati bolje rezultate. Svaki eksperiment služio je za podešavanje uglavnom jednog parametra, a rezultati su prikazani pomoću grafova.

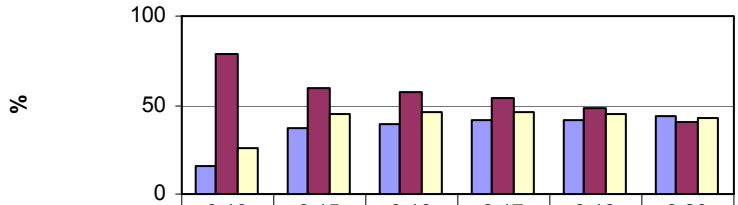
1. eksperiment

Korišteni parametri:

- Minimalni broj dokumenata po deskriptoru: 5
- Minimalni broj dokumenata u kojima se asocijat mora pojaviti: 1
- **Način odabira asocijata**: Ukupna frekvencija pojma u cijelom korpusu (10 - 300)
- Metode evaluacije: Mikro usrednjavanje
- Mjera sličnosti dokumenta i profila deskriptora: Kosinus

Na nekoliko sljedećih grafova će biti prikazani rezultati za navedeni način odabira atributa uz promjenjivu frekvenciju pojma. Također se može vidjeti ovisnost o odabiru praga sličnosti dokumenta i profila deskriptora iznad kojeg su deskriptori dodijeljeni dokumentu.

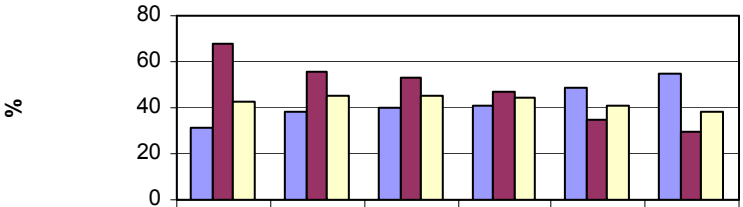
Frekvencija pojma = 10:



	0.10	0.15	0.16	0.17	0.18	0.20
preciznost	15,6	36,6	39,2	41,1	41,6	44,1
odziv	78,2	59,4	56,9	53,5	47,9	40,6
f1	26	45,3	46,4	46,5	44,5	42,3

prag sličnosti

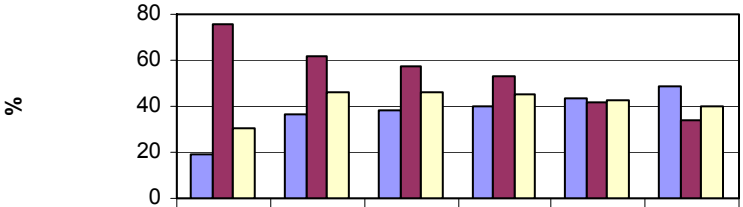
Frekvencija = 20:



	0.15	0.18	0.19	0.20	0.25	0.30
preciznost	31,1	38,1	40,3	41	48,4	54,7
odziv	67,8	55,5	52,7	47,3	34,7	29,4
f1	42,7	45,2	45,6	44	40,5	38,2

prag sličnosti

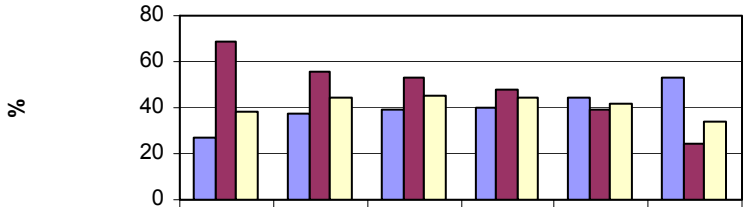
Frekvencija = 50:



	0.15	0.20	0.21	0.22	0.25	0.3
preciznost	19,4	36,4	38,3	40,1	43,1	48,6
odziv	75,4	61,6	57,1	52,9	41,7	33,9
f1	30,8	45,7	45,8	45,6	42,4	39,9

prag sličnosti

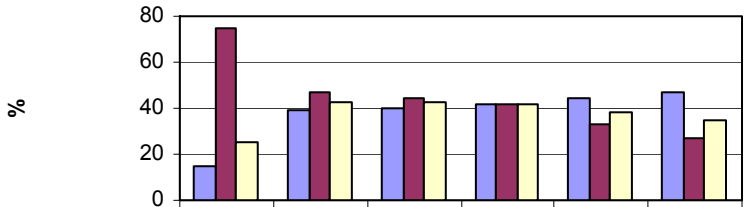
Frekvencija = 100:



	0.20	0.24	0.25	0.26	0.30	0.40
preciznost	26,8	37	38,9	40,4	44	52,7
odziv	68,3	56	52,9	48,2	39,2	24,6
f1	38,5	44,5	44,8	44	41,5	33,6

prag sličnosti

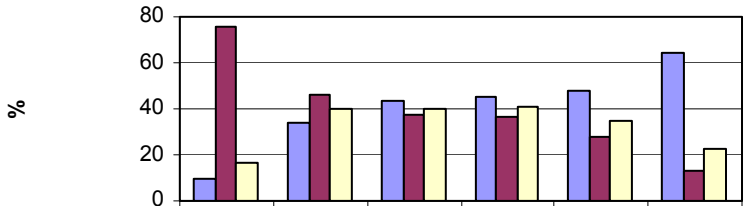
Frekvencija = 200:



	0.20	0.29	0.30	0.31	0.35	0.40
preciznost	15,1	38,9	40,3	41,6	44,2	47,1
odziv	74,5	46,9	44,5	42	33,3	27,2
f1	25,1	42,2	42,3	41,8	38	34,5

prag sličnosti

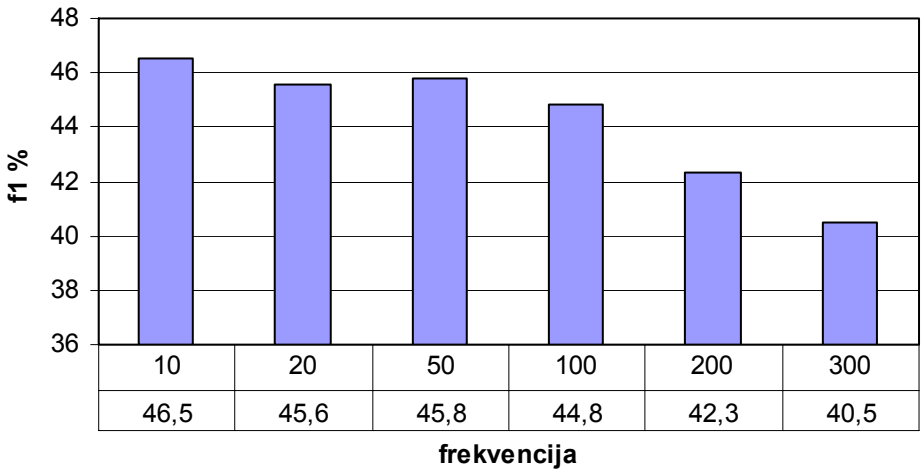
Frekvencija = 300:



	0.20	0.30	0.34	0.35	0.40	0.50
preciznost	9,5	33,6	43,3	45,6	47,6	64
odziv	75,4	45,9	37,3	36,4	27,7	13,4
f1	16,9	40,1	40,1	40,5	35	22,2

prag sličnosti

Ako se za svaku frekvenciju pojma uzme prag sličnosti koji daje najbolju f1 mjeru, rezultati su sljedeći:



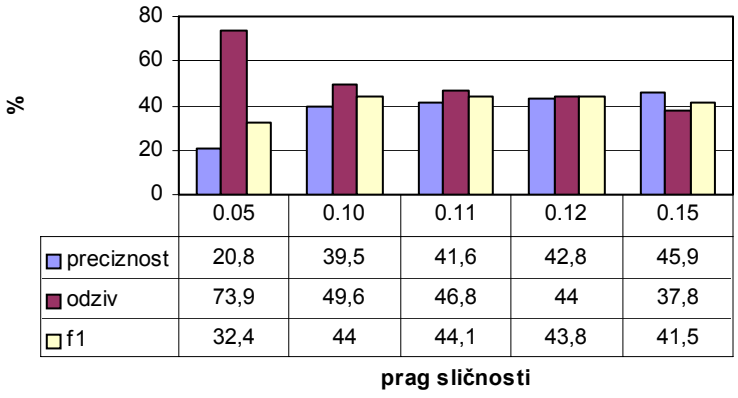
2. eksperiment

Korišteni parametri:

- Minimalni broj dokumenata po deskriptoru: 5
- Minimalni broj dokumenata u kojima se asocijat mora pojaviti: 1
- **Način odabira asocijata**: Omjer log-vjerodostojnosti (2.07 – 12.12)
- Metode evaluacije: Mikro usrednjavanje
- Mjera sličnosti dokumenta i profila deskriptora: Kosinus

Ovaj eksperiment je jednak prethodnom uz tu razliku što se ovdje koristi omjer log-vjerodostojnosti za odabir atributa, a njegova kritična vrijednost se mijenja kroz testove.

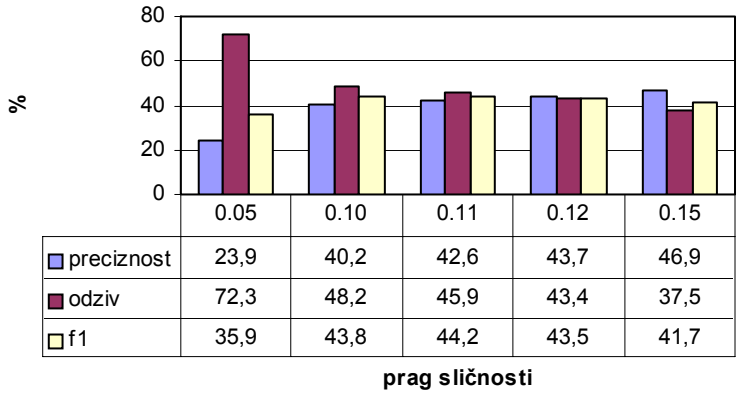
Kritična vrijednost za omjer log-vjerodostojnosti = 2.07 (p = 0.15):



	0.05	0.10	0.11	0.12	0.15
preciznost	20,8	39,5	41,6	42,8	45,9
odziv	73,9	49,6	46,8	44	37,8
f1	32,4	44	44,1	43,8	41,5

prag sličnosti

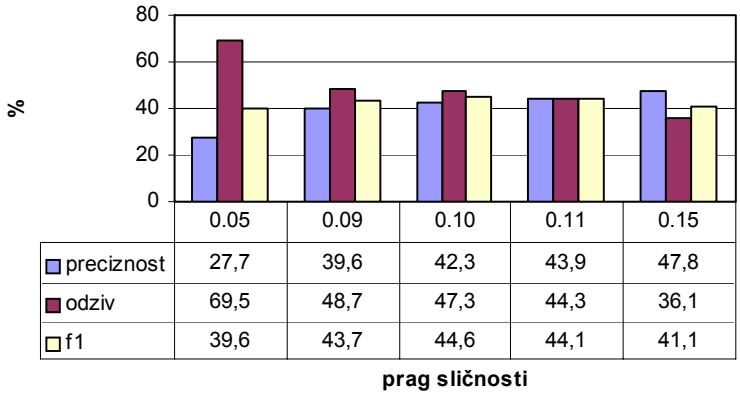
Kritična vrijednost = 2.71 (p = 0.10):



	0.05	0.10	0.11	0.12	0.15
preciznost	23,9	40,2	42,6	43,7	46,9
odziv	72,3	48,2	45,9	43,4	37,5
f1	35,9	43,8	44,2	43,5	41,7

prag sličnosti

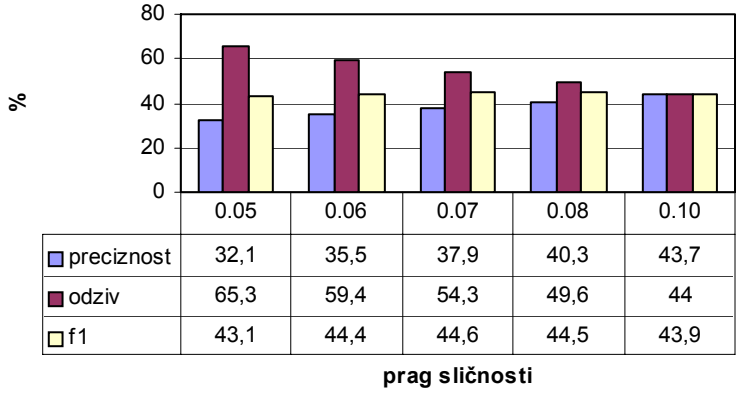
Kritična vrijednost = 3.84 (p = 0.05):



	0.05	0.09	0.10	0.11	0.15
preciznost	27,7	39,6	42,3	43,9	47,8
odziv	69,5	48,7	47,3	44,3	36,1
f1	39,6	43,7	44,6	44,1	41,1

prag sličnosti

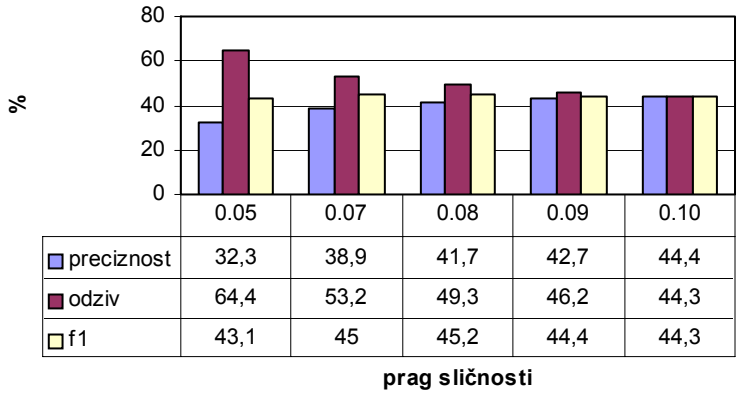
Kritična vrijednost = 6.63 (p = 0.01):



	0.05	0.06	0.07	0.08	0.10
preciznost	32,1	35,5	37,9	40,3	43,7
odziv	65,3	59,4	54,3	49,6	44
f1	43,1	44,4	44,6	44,5	43,9

prag sličnosti

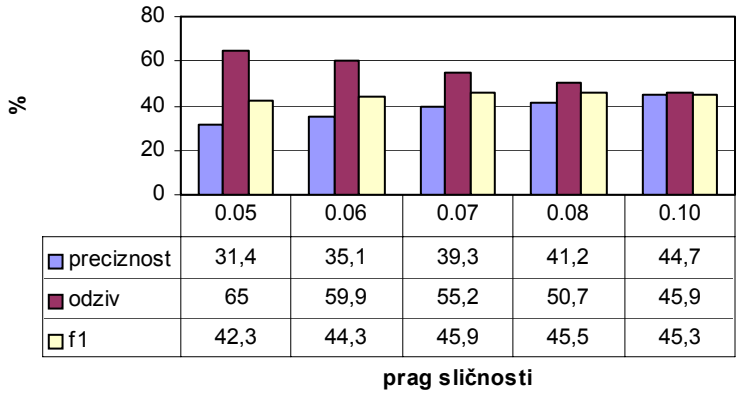
Kritična vrijednost = 7.88 (p = 0.005):



	0.05	0.07	0.08	0.09	0.10
preciznost	32,3	38,9	41,7	42,7	44,4
odziv	64,4	53,2	49,3	46,2	44,3
f1	43,1	45	45,2	44,4	44,3

prag sličnosti

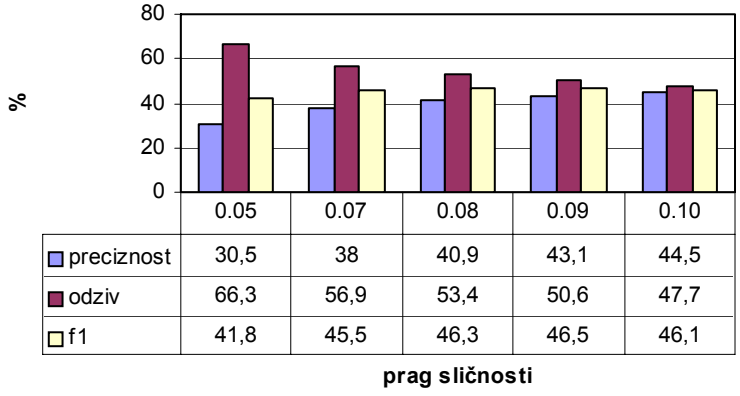
Kritična vrijednost = 10.83 (p = 0.001):



	0.05	0.06	0.07	0.08	0.10
preciznost	31,4	35,1	39,3	41,2	44,7
odziv	65	59,9	55,2	50,7	45,9
f1	42,3	44,3	45,9	45,5	45,3

prag sličnosti

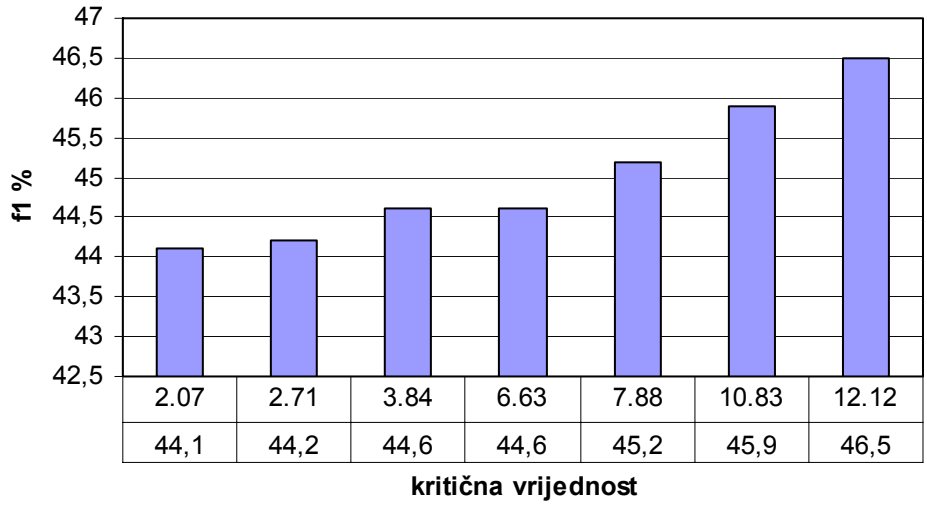
Kritična vrijednost = 12.12 (p = 0.0005):



	0.05	0.07	0.08	0.09	0.10
preciznost	30,5	38	40,9	43,1	44,5
odziv	66,3	56,9	53,4	50,6	47,7
f1	41,8	45,5	46,3	46,5	46,1

prag sličnosti

Ako se za svaku kritičnu vrijednost za omjer log-vjerodostojnosti uzme prag sličnosti koji daje najbolju f1 mjeru, rezultati su sljedeći:



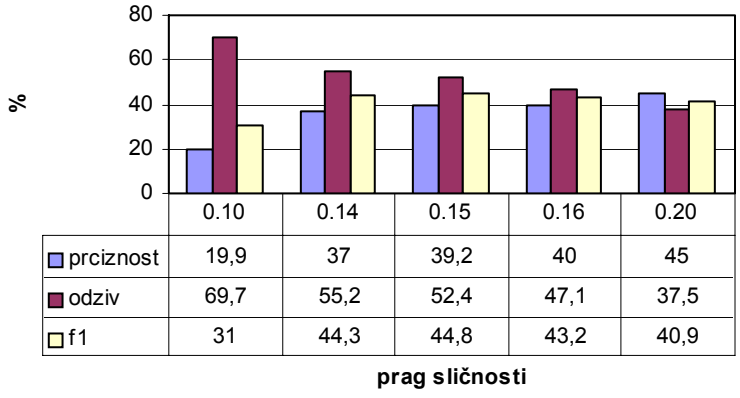
3. eksperiment

Korišteni parametri:

- Minimalni broj dokumenata po deskriptoru: 5
- Minimalni broj dokumenata u kojima se asocijat mora pojaviti: 1
- **Način odabira asocijata**: Frekvencija pojma u skupu dokumenata indeksiranih određenim deskriptorom
- Metode evaluacije: Mikro usrednjavanje
- Mjera sličnosti dokumenta i profila deskriptora: Kosinus

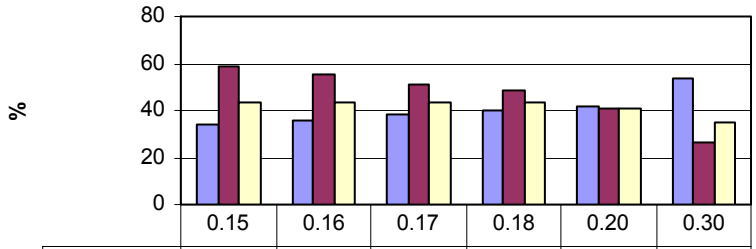
I ovaj je eksperiment rađen radi usporedbe načina odabira atributa. Ovdje se radi o frekvenciji pojma u skupu dokumenata indeksiranih određenim deskriptorom koja se mijenja.

Frekvencija pojma = 2:



	0.10	0.14	0.15	0.16	0.20
prciznost	19,9	37	39,2	40	45
odziv	69,7	55,2	52,4	47,1	37,5
f1	31	44,3	44,8	43,2	40,9

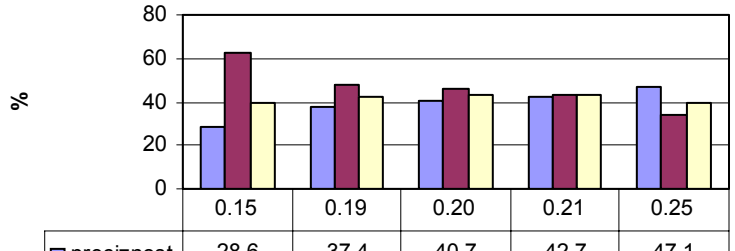
Frekvencija = 3:



	0.15	0.16	0.17	0.18	0.20	0.30
preciznost	34	36,1	38	39,6	41,5	53,7
odziv	58,8	55,2	51,3	48,2	40,9	26,3
f1	43,1	43,6	43,6	43,5	41,2	35,3

prag sličnosti

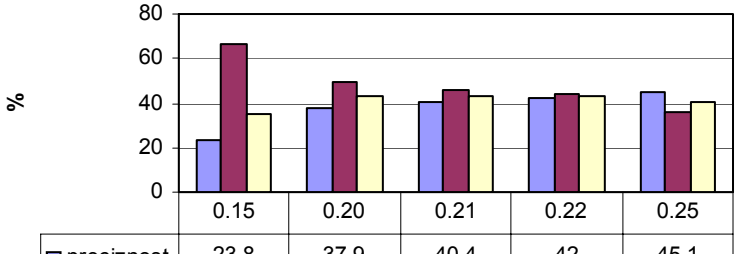
Frekvencija = 4:



	0.15	0.19	0.20	0.21	0.25
preciznost	28,6	37,4	40,7	42,7	47,1
odziv	62,7	47,6	45,9	43,4	33,9
f1	39,3	41,9	43,2	43,1	39,4

prag sličnosti

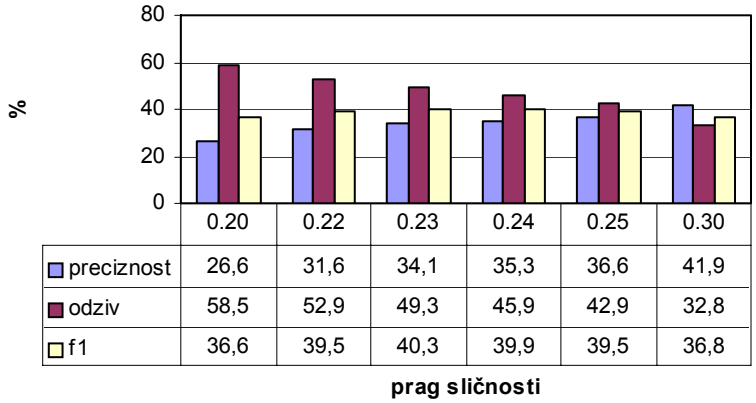
Frekvencija = 5:



	0.15	0.20	0.21	0.22	0.25
preciznost	23,8	37,9	40,4	42	45,1
odziv	66,7	49	46,2	44	36,1
f1	35	42,7	43,1	43	40,1

prag sličnosti

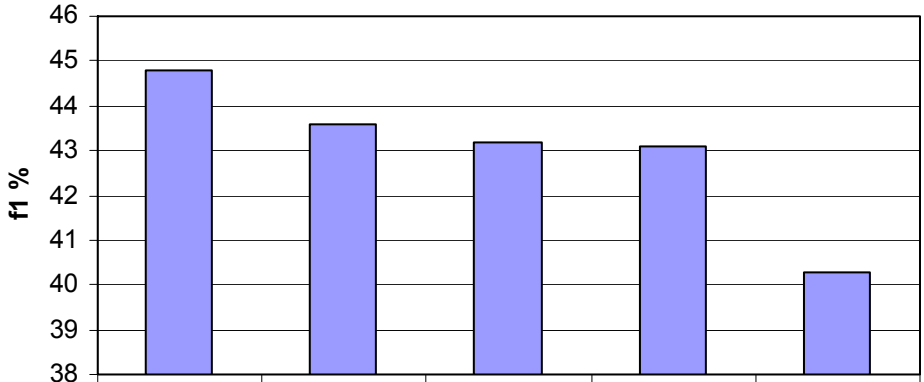
Frekvencija = 10:



	0.20	0.22	0.23	0.24	0.25	0.30
preciznost	26,6	31,6	34,1	35,3	36,6	41,9
odziv	58,5	52,9	49,3	45,9	42,9	32,8
f1	36,6	39,5	40,3	39,9	39,5	36,8

prag sličnosti

Ako se za svaku frekvenciju, uzme prag sličnosti koji daje najbolju f1 mjeru, rezultati su sljedeći:



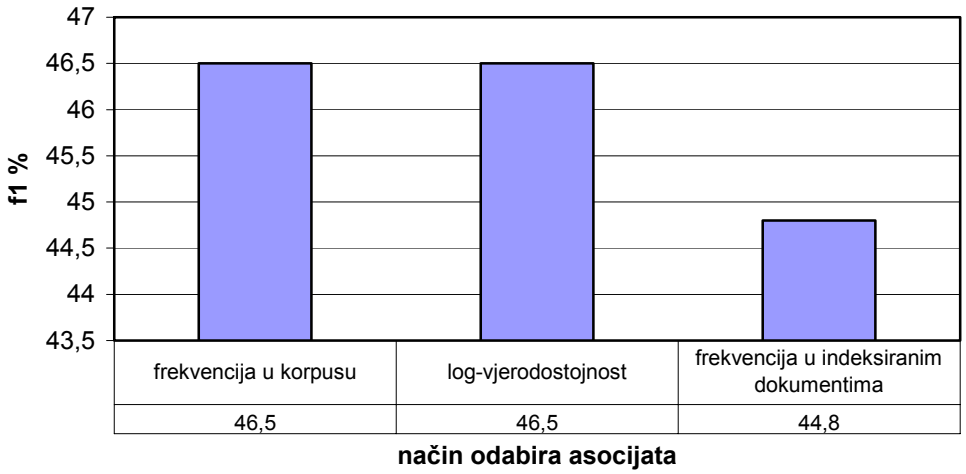
2	3	4	5	10
44,8	43,6	43,2	43,1	40,3

frekvencija

Vidi se da je za svaku frekvenciju pojmova ili kritičnu vrijednost za omjer log-vjerodostojnosti potrebno odrediti drugačije pragove sličnosti dokumenata i profila deskriptora. Pri manjim pragovima znatno su bolji rezultati za odziv, dok s porastom praga uspješnost odziva opada, a preciznosti raste.

Pri odabiru atributa pomoću bilo koje vrste frekvencija rezultati su dosta bolji pri nižim frekvencijama, odnosno kada su liste asocijata duže. Kada listu asocijata određuje omjer log-vjerodostojnosti, rezultati su upravo suprotni. Razlika je gotovo zanemariva, ali algoritam je nešto efikasniji sa višim kritičnim vrijednostima.

Usporedba tri metode odabira atributa prikazana je na sljedećem grafu. Kao što se može vidjeti, razlike nisu toliko velike, ali ipak najbolje rezultate daju frekvencija pojma u cijelom korpusu i omjer log-vjerodostojnosti, a nešto lošije frekvencija pojma u dokumentima indeksiranim pojedinim deskriptorom. Osnovna razlika u prve dvije metode je u tome što su u najboljim rezultatima kod omjera log-vjerodostojnosti liste asocijata znatno kraće nego u ostalim slučajevima, što dosta ubrzava cijeli postupak.



Rezultati dobiveni pomoću kritične vrijednosti za omjer log-vjerodostojnosti bi se možda mogli dodatno poboljšati odabirom nekih drugih vrijednosti. U ostalim slučajevima pokušaji daljnjih poboljšanja bi rezultirali odabirom svih ili gotovo svih pojmova za asocijate, što je ovdje upravo suprotan cilj.

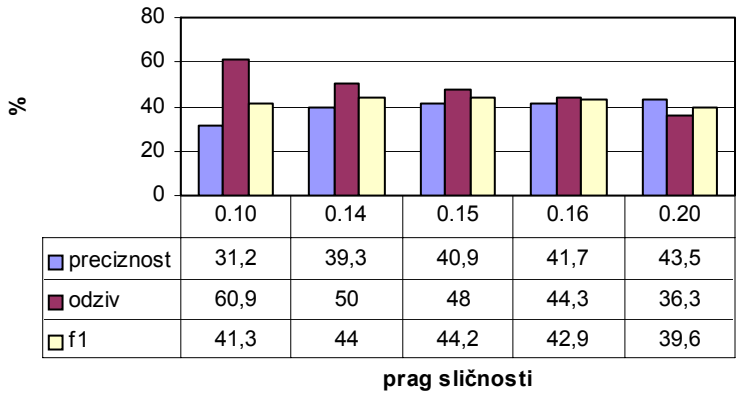
4. eksperiment

Korišteni parametri:

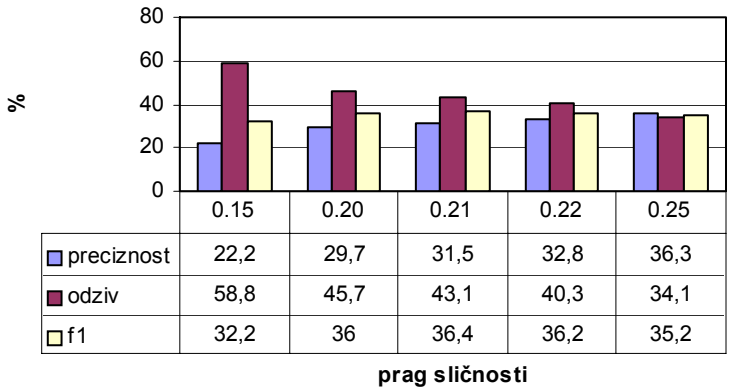
- Minimalni broj dokumenata po deskriptoru: 5
- **Minimalni broj dokumenata u kojima se asocijat mora pojaviti**: 1-3
- Način odabira asocijata: omjer log-vjerodostojnosti
- Metode evaluacije: Mikro usrednjavanje
- Mjera sličnosti dokumenta i profila deskriptora: Kosinus

Kao primjer usporedbe rezultata koji se dobivaju uz određivanje minimalnog broja dokumenata indeksiranih određenim deskriptorom u kojima se asocijat mora pojaviti i rezultata bez tog ograničenja prikazan je omjer log-vjerodostojnosti sa kritičnom vrijednosti od 12.12.

Uz minimalni broj dokumenata postavljen na samo 2, rezultati su sljedeći:



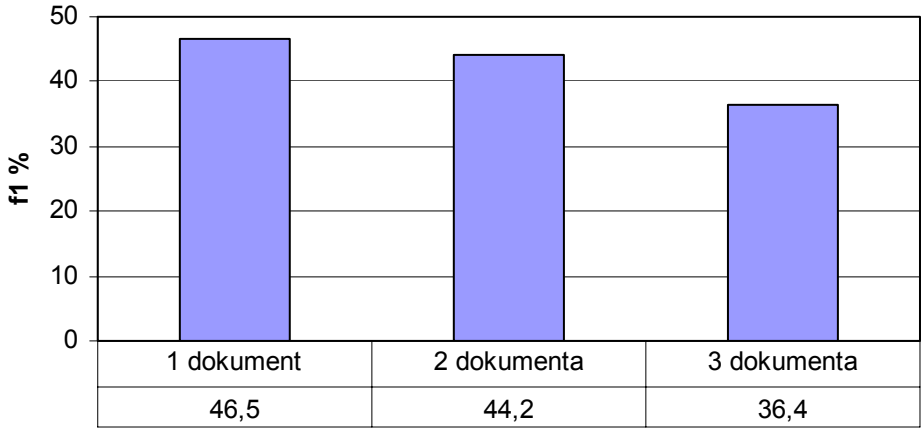
Ako se minimalni broj dokumenata postavi na 3, graf će izgledati ovako:



	0.15	0.20	0.21	0.22	0.25
preciznost	22,2	29,7	31,5	32,8	36,3
odziv	58,8	45,7	43,1	40,3	34,1
f1	32,2	36	36,4	36,2	35,2

prag sličnosti

Najbolji rezultati iz oba slučaja su na sljedećem grafu uspoređeni sa eksperimentom u kojem nije bio zadan minimalni broj dokumenata, odnosno kada je prag iznosio 1.



1 dokument	2 dokumenta	3 dokumenta
46,5	44,2	36,4

minimalni broj dokumenata po deskriptoru u kojima se javlja pojam

Iz grafova se vidi da se dodavanjem ovog dodatnog ograničenja pri izboru asocijata, rezultati pogoršavaju. Isti slučaj se događa i pri drugim

metodama odabira atributa, kao i pri drugim parametrima. U nekim slučajevima je razlika i znatno veća, pa rezultati postaju potpuno neupotrebljivi. Ova pojava može biti i rezultat učenja s relativno malom količinom tekstova. Cijeli skup asocijata nije velik, pa se uvođenjem dodatnih uvjeta on pretjerano smanji. Zbog toga bi se moglo zaključiti da bi uz nešto veći skup za učenje i rezultati mogli biti znatno drugačiji.

5. eksperiment

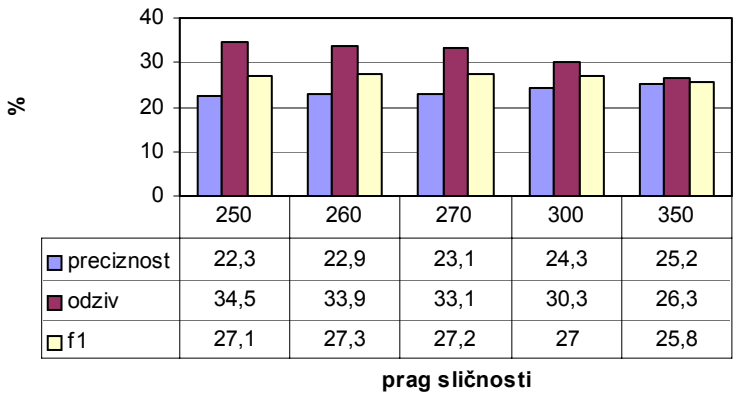
Korišteni parametri:

- Minimalni broj dokumenata po deskriptoru: 5
- Minimalni broj dokumenata u kojima se asocijat mora pojaviti: 1
- Način odabira asocijata: omjer log-vjerodostojnosti
- Metode evaluacije: Mikro usrednjavanje
- **Mjera sličnosti dokumenta i profila deskriptora**: kosinus, skalarni produkt, okapi, kombinacije

U prethodnim eksperimentima je kao mjera sličnosti dokumenta i profila deskriptora korišten isključivo kosinus kuta među vektorima. Ovdje će se pokazati i utjecaj ostalih mjera sličnosti na rezultate, kao i njihovih kombinacija. Za parametre su uzeti oni koji su se u prethodnim testovima pokazali kao uspješni. Kao metoda odabira asocijata odabran je omjer log-vjerodostojnosti s kritičnom vrijednosti 12.12.

Prvo će se pokazati usporedba pojedinih mjera sličnosti, a za to su potrebni sljedeći grafovi:

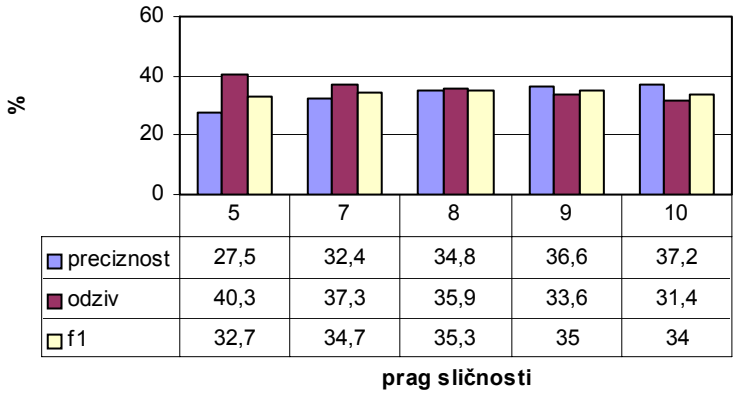
- **Skalarni produkt**



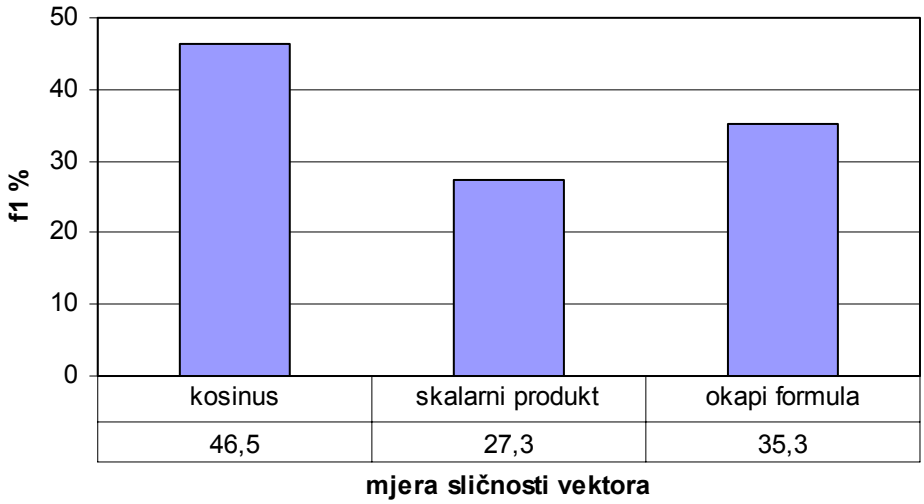
	250	260	270	300	350
preciznost	22,3	22,9	23,1	24,3	25,2
odziv	34,5	33,9	33,1	30,3	26,3
f1	27,1	27,3	27,2	27	25,8

prag sličnosti

- **Okapi formula**



Usporedba samostalnih metoda za određivanje sličnosti može se vidjeti na sljedećem grafu:



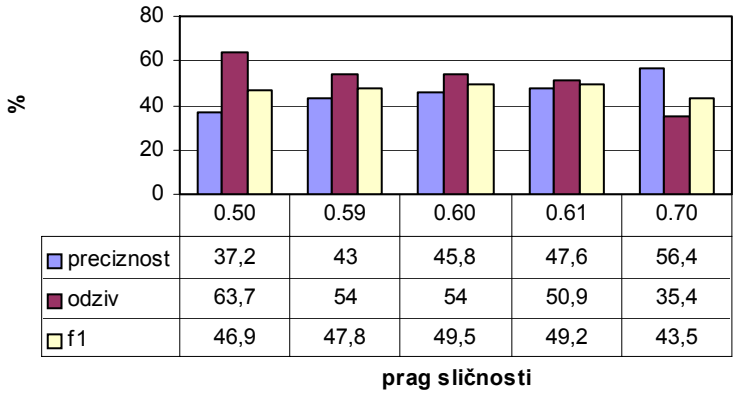
Kao najbolja pojedinačna mjera pokazao se kosinus kuta među vektorima. Radi dodatnog poboljšanja rezultata slijede testovi s kombinacijama kosinusa i ostalih mjera. Iako se skalarni produkt i okapi formula nisu pojedinačno pokazali kao naročito uspješni, njihov učinak može svejedno povećati uspješnost postupka, što se i vidi na sjedećim grafovima.

- **Kombinacija kosinusa i skalarnog produkta**

Uzeti su različiti omjeri utjecaja kosinusa i skalarnog produkta. U prvom slučaju utjecaji su jednaki i formula za računanje sličnosti izgleda ovako:

$$0.5 \cdot \frac{ko\sin us}{Max(ko\sin us)} + 0.5 \cdot \frac{Skal\,\Pr od}{Max(Skal\,\Pr od)}.$$

Omjer kosinus-skalarni produkt: 50-50%:

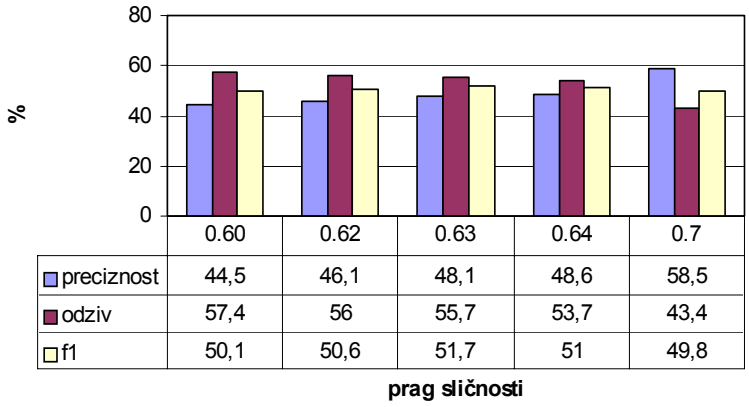


	0.50	0.59	0.60	0.61	0.70
preciznost	37,2	43	45,8	47,6	56,4
odziv	63,7	54	54	50,9	35,4
f1	46,9	47,8	49,5	49,2	43,5

prag sličnosti

Pošto se kosinus pokazao kao bolja samostalna mjera od skalarnog produkta omjer utjecaja je promijenjen u korist kosinusa i uzeti su omjeri 60-40% i 70-30%.

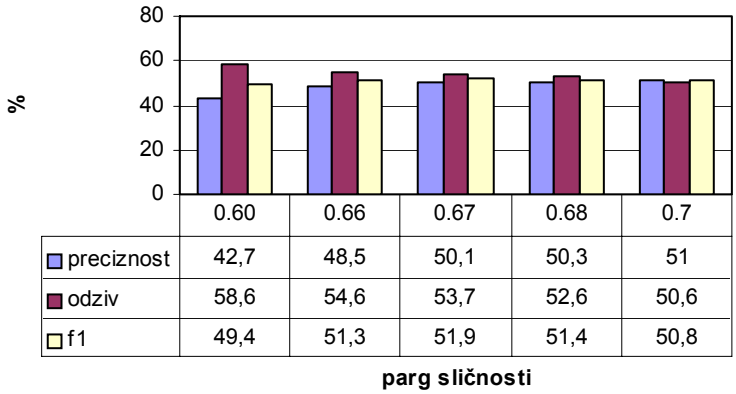
Omjer kosinus-skalarni produkt: 60-40%:



	0.60	0.62	0.63	0.64	0.7
preciznost	44,5	46,1	48,1	48,6	58,5
odziv	57,4	56	55,7	53,7	43,4
f1	50,1	50,6	51,7	51	49,8

prag sličnosti

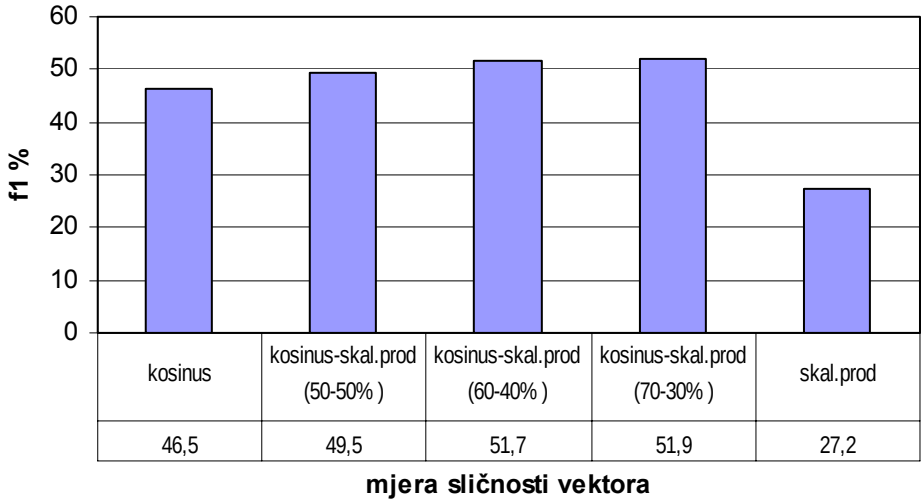
Omjer kosinus-skalarni produkt: 70-30%:



	0.60	0.66	0.67	0.68	0.7
preciznost	42,7	48,5	50,1	50,3	51
odziv	58,6	54,6	53,7	52,6	50,6
f1	49,4	51,3	51,9	51,4	50,8

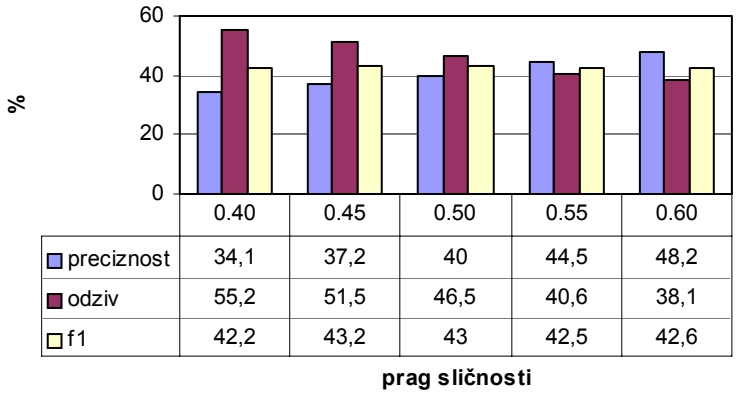
parg sličnosti

Usporedba različitih omjera utjecaja kosinusa i skalarnog produkta dana je na sljedećem grafu. Može se vidjeti da kombinacija ovih metoda daje najbolje rezultate, a posebno u slučaju kada je veća težina dana kosinusu kuta.

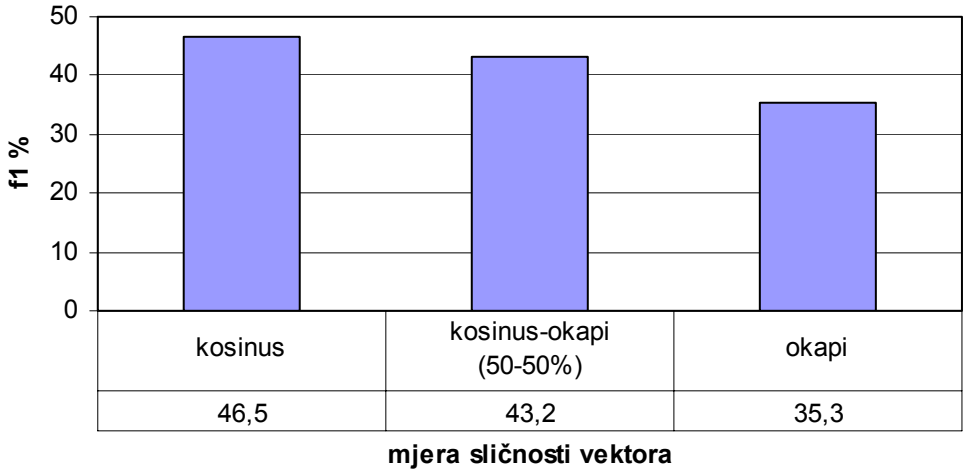


- **Kombinacija kosinusa i okapi formule**

<u>Omjer kosinus-okapi formula: 50-50%:</u>



Usporedba učinkovitosti kosinusa, okapi formule i njihove kombinacije vidi se na sljedećem grafu. U ovom se slučaju kao najuspješnija pokazala samostalna kosinus mjera.

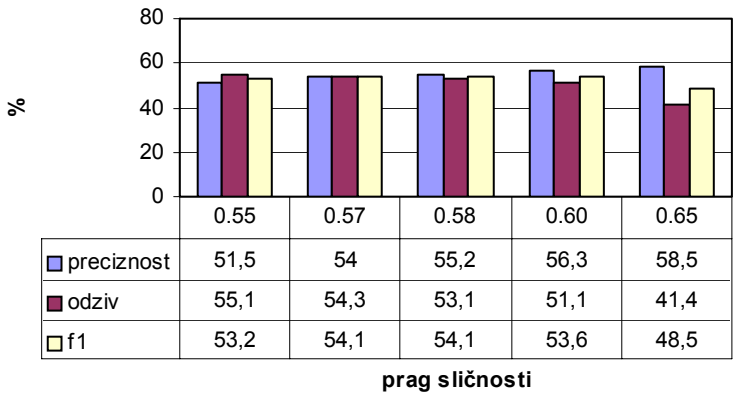


- **Kombinacija sve tri mjere**

Pošto se kosinus mjera pokazala kao najbolja, ona će imati najveći utjecaj u ukupnoj formuli koja glasi:

$$0.6 \cdot \frac{ko\sin us}{Max(ko\sin us)} + 0.2 \cdot \frac{Skal\,\Pr od}{Max(Skal\,\Pr od)} + 0.2 \cdot \frac{Okapi}{Max(Okapi)}.$$

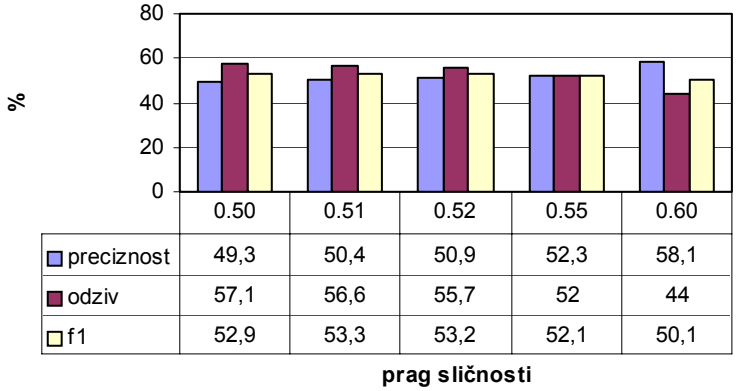
Omjer kosinus-skalarni produkt-okapi formula: 60-20-20%:



U dosadašnjim testovima se i kombinacija kosinusa i skalarnog produkta pokazala prilično uspješnom, pa su u sljedećem eksperimentu te

dvije mjere uzete u nešto većem omjeru, dok je okapi formula prilično zanemarena.

Omjer kosinus-skalarni produkt-okapi formula: 45-45-10%:

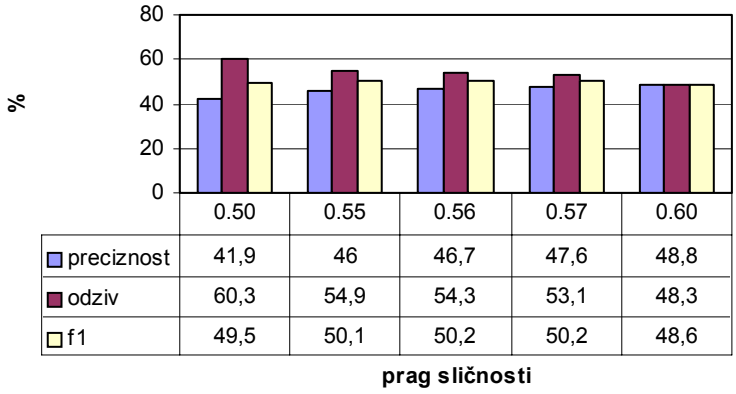


	0.50	0.51	0.52	0.55	0.60
preciznost	49,3	50,4	50,9	52,3	58,1
odziv	57,1	56,6	55,7	52	44
f1	52,9	53,3	53,2	52,1	50,1

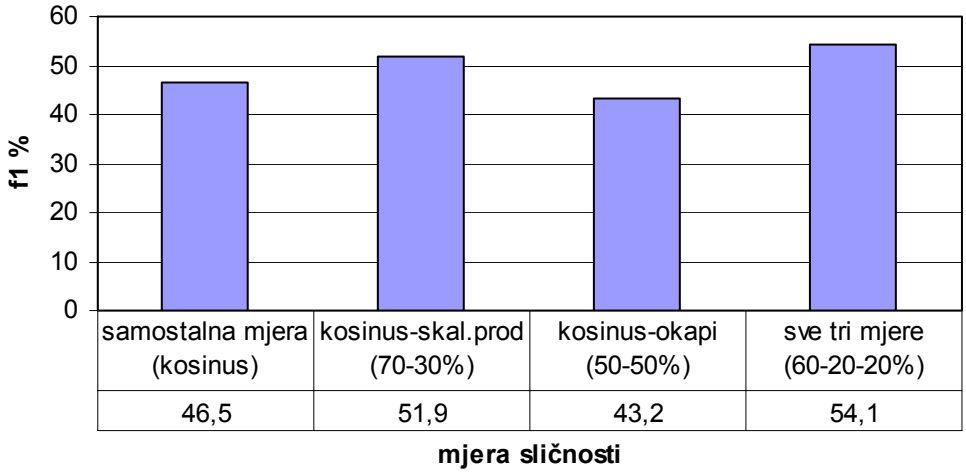
prag sličnosti

Iz grafova se može vidjeti da je nešto veća uspješnost postignuta u prvom slučaju, kada je najveća težina dana mjeri kosinusa kuta među vektorima.

U dosadašnjim se primjerima kao težina riječi u novim dokumentima koje je potrebno indeksirati koristila frekvencija pojma. U sljedećem je eksperimentu ponovljena kombinacija sve tri mjere sličnosti u omjerima 60-20-20 %, ali s tom razlikom da se prilikom računanja kosinusa i skalarnog produkta uzimalo u obzir jedino pojavljivanje riječi. Kao što se i može vidjeti na sljedećem grafu, ova se metoda nije pokazala naročito uspješnom i rezultati su za nekoliko postotaka lošiji.



Usporedba različitih kombinacija mjera sličnosti može se vidjeti na sljedećem grafu. Kao najuspješnija mjera pokazala se kombinacija sve tri metode u omjeru 60-20-20% uz težinu pojma u novom dokumentu određenu na temelju njegove frekvencije.



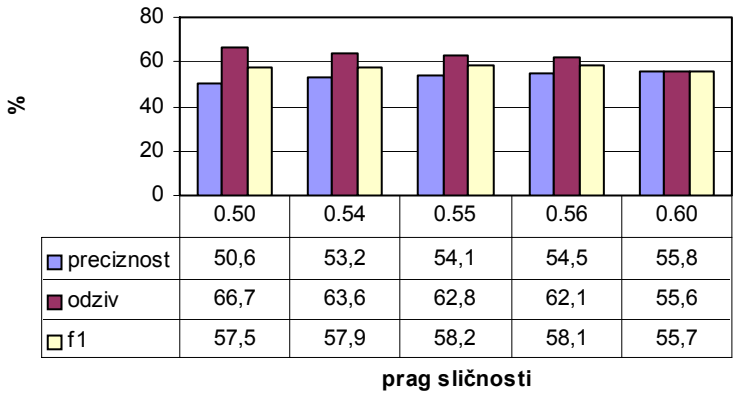
6. eksperiment

Korišteni parametri:

- **Minimalni broj dokumenata po deskriptoru**: 5-20
- Minimalni broj dokumenata u kojima se asocijat mora pojaviti: 1
- Način odabira asocijata: omjer log-vjerodostojnosti
- Metode evaluacije: Mikro usrednjavanje
- Mjera sličnosti dokumenta i profila deskriptora: Kombinacija sve tri mjere

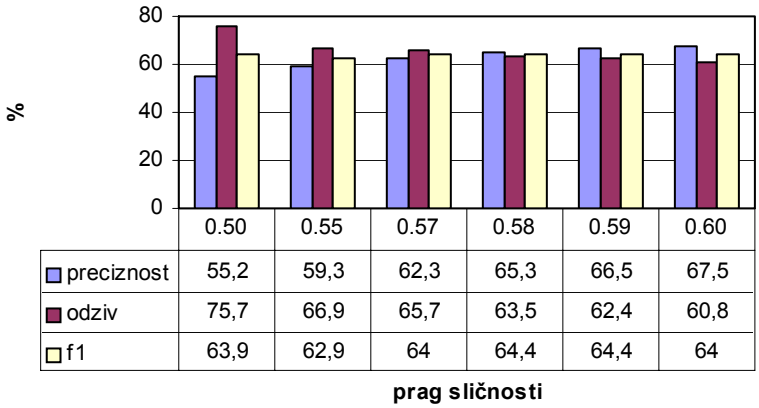
Kako bi se poboljšala kvaliteta profila, povećan je minimalni broj dokumenata po deskriptoru koji je potreban da bi se profil uopće stvarao. Za odabir asocijata korišten je omjer log-vjerodostojnosti s kritičnom vrijednosti 12.12, a kao mjera sličnosti kombinacija sve tri mjere, koja se do sada pokazala kao najuspješnija.

Ako se granica za minimalni broj dokumenata po deskriptoru postavi na 10 rezultati izgledaju ovako:



	0.50	0.54	0.55	0.56	0.60
preciznost	50,6	53,2	54,1	54,5	55,8
odziv	66,7	63,6	62,8	62,1	55,6
f1	57,5	57,9	58,2	58,1	55,7

Uz minimalni broj dokumenata po deskriptoru koji iznosi 20, rezultati su sljedeći:

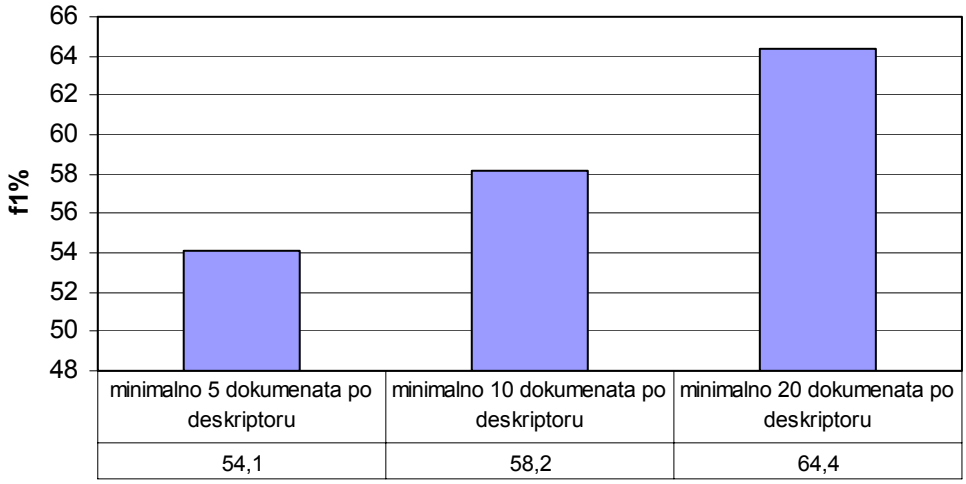


	0.50	0.55	0.57	0.58	0.59	0.60
preciznost	55,2	59,3	62,3	65,3	66,5	67,5
odziv	75,7	66,9	65,7	63,5	62,4	60,8
f1	63,9	62,9	64	64,4	64,4	64

prag sličnosti

U ovom je slučaju dobivena najveća uspješnost rezultata do sada od 64.4%. Ako se promatraju preciznost i odziv, koji se u nekim slučajevima mogu pokazati kao značajnije mjere, rezultati su i znatno bolji. Na grafovima su prikazane samo vrijednosti oko najbolje f1 mjere, ali pomicanjem praga sličnosti preciznost i odziv mogu postići i vrijednosti znatno bolje od navedenih.

Usporedba rezultata uz različite minimalne brojeve dokumenata po deskriptoru vidi se na sljedećem grafu. Očito je da se uz povećanje broja dokumenata rezultati znatno poboljšavaju. Jedini nedostatak visoke donje granice broja dokumenata je u tome što se na taj način mnogi deskriptori eliminiraju. Za rad s većim brojem deskriptora bilo bi potrebno raditi sa znatno većim skupom za učenje, koji u ovom slučaju nije dostupan.


66
64
62
60
58
56
54
52
50
48
f1%
minimalno 5 dokumenata po deskriptoru
minimalno 10 dokumenata po deskriptoru
minimalno 20 dokumenata po deskriptoru
54,1
58,2
64,4

11. Zaključak

Zadatak diplomskog rada je bio primijeniti model za indeksiranje dokumenata na temelju sličnosti vektora (Pouliquen i Steinberger [4]) i prilagoditi postupak za dokumente na hrvatskom jeziku. Unatoč nedostatku veće količine indeksiranih dokumenata potrebnih za proces kvalitetnog učenja i morfološkog bogatstva hrvatskog jezika, koje otežava računalnu obradu, model se može smatrati primjenjivim. Optimalna kombinacija parametara implementiranog modela se razlikuje od onih koji su korišteni za druge jezike, ali se i neke od tih razlika mogu pripisati nedovoljno kvalitetnim ulaznim podacima, pa su temeljem toga i rezultati različiti. Zbog toga se ovaj model primijenjen na hrvatski jezik može smatrati uvodom u područje indeksiranja dokumenata na hrvatskom, a rezultati usporedivi s ostalim svjetskim sustavima mogu se dobiti upotrebom istih algoritama uz kvalitetniju obradu ulaznih podataka.

Uspoređujući rezultate dobivene pomoću implementiranog modela korištenjem F1 mjere, najveća postignuta uspješnost indeksiranja iznosila je 64.4%. Budući da se sustav može koristiti za automatsko ili poluautomatsko indeksiranje, F1 mjera ne mora uvijek određivati optimalne rezultate. U slučaju automatskog indeksiranja bitno je ostvariti što veću preciznost, jer se očekuje ispravnost svih dodijeljenih deskriptora. U slučaju poluautomatskog indeksiranja, veća se težina treba dati odzivu, jer na temelju većeg broja dodijeljenih deskriptora korisnik sam odabire odgovarajuće. Procjenjujući rad modela korištenjem mjere preciznosti ili odziva dobiva se uspješnost indeksiranja od 75% i više.

12. Literatura

[1] C. D. Manning i H. Schutze, Fundations of Statistical Natural Language Processing, The MIT Press, Cambridge, Massachusetts, 2003.

[2] D. Sulivan, "The Need for Text Mining in Business Intelligence", DM Direct Special Report, 2004. http://www.dmreview.com/article_sub.cfm?articleId=8100 [3.7.2005.].

[3] K. Kells, What is Indexing?, Index West, Olympia, http://www.indexw.com/whatido.htm [25/6/2005].

[4] R. Steinberger i B. Pouliquen, Cross-lingual Indexing, European Commission – Joint Research Centre (JRC), Institute for the protection and Security of the Citizen (IPCS), Ispra, Italija, 2003.

[5] Automatic vs. manual indexing, European Library Automation Group, 2001, http://www.stk.cz/elag2001/Workshop/ws4.doc [25/6/2005].

[6] M. Kolar, I. Vukmirović, B. Dalbelo Bašić, J. Šnajder, Computer Aided Document Indexing System, 27th International Conference Information Technology Interfaces ITI 2005, pp. 343-348, Cavtat, 2005.

[7] M. F. Moens, Automatic indexing and Abstracting of Document Texts, Kluwer Academic Publishers, Massachusetts, 2000.

[8] G. Salton, Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Massachusetts, 1989.

[9] A. Nikolić-Hoyt, Konceptualna leksikografija, Prema tezaurusu hrvatskog jezika, Hrvatska sveučilišna naklada, Zagreb, 2004.

[10] M. M. K. Hlava, Automatic Indexing - Return on Investment (ROI), A Case Study Comparison of Rule Base and Statistical Approaches, http://www.dataharmony.com/papers/roiMaiComparison.html [24./5./2005.].

[11] T. M. Mitchell, Machine Learning, The McGraw-Hill Companies, 1997.

[12] T. Lahtinen, Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods, doktorska disertacija, University of Helsinki, Department of General Linguistics, Finland, 2000.

[13] Y. Yang, J. O. Pedersen, A Comparative Study on Feature Selection in Text Categorization, Fourteenth International Conference on Machine Learning, pp. 412-420, 1997.

[14] I. Levner, Feature selection and nearest centroid classification for protein mass spectrometry, BMC Bioinformatics, 2005.

[15] P. E. Rayson, Matrix: A statistical method and software tool for linguistic analysis through corpus comparison, Doktorska disertacija, Lancaster University, Lancaster, UK, 2002.

[16] G. Forman, An Extensive Empirical Study of Feature Selection Metrics for Text Classification, Journal of Machine Learning Research 3, pp. 1289-1305, 2003.

[17] P. Radivojac, N. V. Chawla, A. K. Dunker, Z. Obradovic, Classification and knowledge discovery in protein databases, Journal of Biomedical Informatics 37, pp. 224–239, 2004.

[18] S. Simon, STATS - Steve's Attempt to Teach Statistics, http://www.cmh.edu/stats/definitions/or.htm [3.7.2005.].

[19] D. Mladenič, How to Approach Data Analisys of Text, Journal of Information and Organizational Sciences, vol. 28, 2004.

[20] F. Tang, Automatic Identification in Document Indexing, Statistic Department, Berkeley, http://stat-www.berkeley.edu/users/vigre/undergrad/reports/tang.pdf [3./3./2005.].

[21] H. Fang, T. Tao, C. Zhai, A Formal Study of Information Retrieval Heuristics, ACM SIGIR 2004, pp. 49-56, 2004.

[22] S. E. Robertson, S. Walker, S. Jones, M. M. Hanoock-Beaulieu, M. Gatford, Okapi in TREC-3, Text Retrieval Conference TREC-3, pp. 109-128, 1994.

[23] M. M. Beaulieu, M. Gatford, X. Huang, S. E. Robertson, S. Walker, P.Williams, Okapi at TREC-5, The Fifth Text Retrieval Conference (TREC-5), pp. 143-165, 1997.

[24] S. E. Robertson, S. Walker, M. Beaulieu, Okapi at TREC-7: automatic ad hoc, filtering, VLC and interactive track, The Seventh Text Retrieval Conference (TREC-7), pp. 253-264, 1999.

[25] M. Malenica, Primjena jezgrenih metoda u kategorizaciji teksta, Diplomski rad, Fakultet elektrotehnike i računarstva, Zagreb, 2004.

[26] R. Ozcan, Y. A. Aslandogan, Concept Based Information Access Using Ontologies and Latent Semantic Analysis, Tehnički izvještaj CSE-2004-8, University of Texas at Arlington, Arlington, 2004.

[27] B. Ripplinger, P. Schmidt, Automatic Multilingual Indexing and Natural Language Processing, SIGIR 2000, New Orleans, 2000.

[28] R. Ferber, Information Retrieval, Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web, dpunkt.verlag, Heidelberg, 2003. http://information-retrieval.de/irb/ir.part_3.chapter_5.section_6.html [5.7.2005.].

[29] N. Fuhr, S. Hartmann, G. Lustig, M. Schwantner, K. Tzeras, G. Knorz, AIR/X – Rule-Based multistage Indexing System for Large Subject Fields, RIAO'91, pp. 606-623, 1991.

[30] N. Fuhr, S. Hartman, G. Lustig, K. Tzeras, G. Knorz, M. Schwantner, Automatic Indexing in Operation: The Rule Based System AIR/X for Large Subject Fields, Tehnički izvještaj, Techniche Hochschule Darmstadt, Njemačka, 1993.

[31] K. Tzeras, S. Hartmann, Automatic Indexing Based on Bayesian Interface Networks, 16th ACM SIGIR conference on Research and development in information retrieval, pp. 22-35, 1993.

[32] Y. Yang, C. G. Chute, A Linear Least Squares Fit Mapping Method for Information Retrieval from Natural Language Texts, COLING-92, pp. 23-28, 1992.

[33] B. Burnside, H. Strasberg i D. Rubin, Automated Indexing of Mammography Reports Using Linear Least Squares Fit, Stanford Medical Informatics, Stanford, CA.

[34] B. v. Bakel, R. T. Boon, N. J. I. Mars, E. Oltmans, Condorcet Final Report, Vossius Laboratory, University of Twente, Enschede, The Netherlands, Tehnički izvještaj CTIT TR-00-02, 2000.

[35] E. Oltmans, A Knowledge-Based Approach to Robust Parsing, Doktorska disertacija, Centre for Telematics and Information Technology (CTIT), The Netherlands, 1999.

[36] C. Plaunt, B. A. Norgard, (1998), An Association-Based Method for Automatic Indexing with a Controlled Vocabulary, Journal of the American Society for Information Science 49(10), pp. 888–902, 1998.

[37] K. P. Burnham, G. C. White, S. Converse, B. McClintock, The Binomial Likelihood Function, Colorado State University, 2004. http://www.cnr.colostate.edu/class_info/_fw663/BinomialLikelihood.PDF [10.5.2005.].

[38] J. P. Silvester, M. T. Genuardi, P. H. Klingbiel, Machine-Aided Indexing at NASA, NASA CASI, Information Processing & Management, Vol. 30, No. 5, pp. 631-645, 1994.

[39] Hidra, http://www.hidra.hr/ [30.6.2005.].

[40] B. Pouliquen, Automatic Eurovoc Indexing, European Commission – Joint Research Centre (JRC), Institute for the protection and Security of the Citizen (IPCS), Ispra, Italija, 2004.

[41] On-line Chi-sqared table, http://perdana.fsktm.um.edu.my/~tehyw/Principles%20of%20Biology%20-%20BIOL1-BIOL2%20%20Chi-Squared%20Table.htm [3.7.2005.]

13. Dodatak

13.1. χ^2 kritične vrijednosti

df \ p	0.25	0.20	0.15	0.10	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.59	6.74	7.78	9.49	11.14	11.67	13.23	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.33	15.09	16.75	18.39	20.51	22.11
6	7.84	8.56	9.45	10.64	12.53	14.45	15.03	16.81	13.55	20.25	22.46	24.10

.
.
.