Laboratorij za tehnologije znanja (KTLab)
Zavod za elektroniku, mikrolektroniku, računalne i inteligentne sustave
Fakultet elektrotehnike i računarstva
Sveučilište u Zagrebu

Interni dokument

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI ZADATAK br. 1614

PRIMJENA TEORIJE GRAFOVA U DUBINSKOJ ANALIZI TEKSTA

Frane Šarić

Zagreb, rujan 2006.

Sadržaj

1. Sažetak

Posljednjih desetak godina web-tražilice počele su primjenjivati metode analize grafova radi rangiranja rezultata pretrage. Budući da su se te metode pokazale uspješnima, njihova primjena na razna područja sve više se istražuje.

U Blondelovom članku [3] opisana je mjera sličnosti među vrhovima dva grafa, te primjena te metode na nalaženje sinonima u rječniku. Nalaženje sinonima jedan je od alata koji se može koristiti prilikom dubinske analize teksta. U ovom radu proučena je primjena mjere sličnosti na automatsko indeksiranje dokumenata.

Uspješnost metode nalaženja sinonima ocijenjena je korištenjem engleskog i hrvatskog rječnika. Automatsko indeksiranje korištenjem mjere sličnosti evaluirano je na skupu dokumenata indeksiranim pojmovima iz hijerarhijskog pojmovnika EUROVOC [5].

U radu se koriste neki od rezultata projekta „Text mining system – Sustav za automatsko indeksiranje, kategorizaciju i semantičko pretraživanje teksta“ koji je nastao kao suradnja tima Fakulteta elektrotehnike i računarstva i Filozofskog fakulteta Sveučilišta u Zagrebu, Hrvatske informacijsko-dokumentacijske referalne agencije i Zajedničkog istraživačkog centra Europske komisije u Ispri, Italija.

2. Uvod

Porastom popularnosti Interneta, a posebno web-stranica, broj dostupnih dokumenata povećao se gotovo geometrijskom progresijom. Uz porast količine dokumenata pojavio se i problem nalaženja informacija, tj. dokumenata koji sadrže korisne informacije. Iako su prve web tražilice (npr. Excite[1], Hotbot[2], Altavista[3]) omogućile jednostavnu i brzu pretragu tog skupa, relevantni dokumenti često se nisu nalazili među prvim rezultatima pretrage. Nalaženje poretka pronađenih dokumenata prema tome koliko su oni relevantni za korisnički upit ostao je veliki problem. Najveći korak u rješavanju tog problema načinila je web tražilica Google[4] kada je prije desetak godina počela primjenjivati metode teorije grafova na analizu veza među web stranicama, tj. algoritam PageRank [6]. Od tada se pokušavaju slične metode primijeniti na mnoga druga područje, ne nužno vezana uz pretragu web stranica.

Područje u kojem se tek počinju primjenjivati slične metode jest dubinska analiza teksta. Dubinska analiza teksta podrazumijeva pronalaženje i sintezu informacija iz isključivo tekstnih podataka. Jedan od alata koji se koristi prilikom analize teksta jest nalaženje sinonima neke riječi. Ponekad je korisno svesti razne sinonime neke riječi u jednu kategoriju. Sinonimi se mogu koristiti i za transformaciju upita koji korisnik zadaje na način da riječi iz upita budu sličnije riječima korištenim u dokumentu [14, 13]. Blondel [4] opisuje metodu nalaženja sinonima u engleskom jeziku korištenjem engleskog rječnika. U ovom radu ta metoda primijenjena je na nalaženje sinonima u hrvatskom jeziku.

[1] http://www.excite.com

[2] http://www.hotbot.com

[3] http://www.altavista.com

[4] http://www.google.com

Prilikom pretrage web stranica korisnik zadaje upit slobodnim odabirom riječi. Ponekad je teško odabrati prave riječi čijim zadavanjem bi tražilica vratila rezultate koje korisnik traži. U slučaju da se pretražuju znatno manje kolekcije dokumenata (od svih web stranica na Internetu) može se svaki dokument unaprijed označiti sa skupom pojmova relevantnim za taj dokument. Sami pojmovi kojima se označavaju dokumenti odabire se iz kontroliranog rječnika ili pojmovnika. Proces dodjeljivanja relevantnih pojmova svakom dokumentu zovemo indeksiranje. Indeksirane dokumente korisnik može pretraživati tako da koristi samo pojmove iz kontroliranog rječnika. Ukoliko su pojmovi hijerarhijski organizirani, lako se mogu naći specifičniji ili općenitiji rezultati pretrage promjenom pojmova upita na specifičnije ili općenitije pojmove.

Jedan od problema kojima se bavi dubinska analiza teksta jest automatska kategorizacija dokumenata, tj. svrstavanje dokumenata u jednu ili više zadanih kategorija. Indeksiranje dokumenata možemo shvatiti kao kategorizaciju dokumenata u niz kategorija koje čine pojmovi kojima se indeksiraju dokumenti. U Blondelovom članku [3] opisana je mjera sličnosti među vrhovima dva grafa. U ovom radu pokušava se odgovoriti na pitanje kako se može koristiti slična metoda u svrhu automatskog indeksiranja dokumenata korištenjem pojmovnika EUROVOC [5].

U literaturi su opisane razne automatske metode nalaženja sinonima. Metode koje su u svojim radovima opisali Wu [24] i Curran [7] koriste samo velike količine tekstova u jednom jeziku. Poput Blondelove metode [4] koja je implementirana u ovom radu i metoda ArcRank [11] koristi rječnik za nalaženje sinonima.

Već su razvijeni sustavi za automatsko indeksiranje dokumenata korištenjem kontroliranog pojmovnika koji ne zahtijevaju eksperta, tj. čovjeka koji bi ručno zadao pravila dodjele pojmova dokumentima. Sustav sveučilišta u Berkeleyu [18] koristi sličnosti naziva pojmova i raznih dijelova dokumenta za indeksiranje dokumenta. Sustav CONDORCET [23] analizom sažetaka i

naslova dokumenata određuje kojim pojmovima se indeksira dokument. Pouliquen [19] razvio je sustav za indeksiranje dokumenata korištenjem pojmovnika EUROVOC. Ovaj sustav, koristeći skup indeksiranih dokumenata, svakom pojmu iz kontroliranog rječnika dodjeljuje niz ključnih riječi i težina. Prilikom indeksiranja novog dokumenta traže se pojmovi čije riječi najbolje odgovaraju riječima u dokumentu. U literaturi se metode analize grafova nisu koristile za potrebe indeksiranja dokumenata.

3. Analiza veza u grafovima

Značajni skok u kvaliteti rangiranja web stranica dogodio se kada je tražilica Google uvela metodu *PageRank* [6]. Google je počeo rangirati stranice ne samo prema analizi pojavljivanja upita na traženoj stranici, već i korištenjem mjere PageRank koja svakoj stranici dodjeljuje važnost neovisno o upitu korisnika. PageRank u potpunosti ovisi o vezama među raznim web stranicama.

Metoda čvorišta i autoriteta također koristi strukturu povezanosti web stranica. Tom metodom određuju se mjerodavnosti (koliko je neka stranica dobar autoritet) stranica s obzirom na dani korisnički upit. Generalizacijom ove metode dolazimo do pojma sličnosti među vrhovima u grafu, tj. do matrice sličnosti.

Ove ketode kao relevantnu karakteristiku web stranica uzimaju u obzir **veze** (hiperlinkove) koje određena stranica sadrži. Pri tome se gledaju samo veze koje ta stranica ima prema drugim stranicama, a ignoriraju se veze koje pokazuju na drugi dio iste stranice. Nadalje, ako stranica *a* ima vezu koja pokazuje na stranicu *b*, onda ćemo tu vezu zvati **izlaznom vezom** stranice *a* i **ulaznom vezom** stranice *b*.

3.1. PageRank

Uspjeh ove metode može se pripisati iskorištavanju strukture veza među web stranicama radi ocjenjivanja važnosti neke stranice. PageRank vrijednost neke stranice zapravo predstavlja vjerojatnost da će se korisnik nakon odabira slučajne veze u svom web-pregledniku zadržati na toj stranici. Pretpostavka je da će korisnik ako naiđe na stranicu koja nema izlaznih veza nastaviti sa pregledavanjem neke druge, slučajne, stranice na Internetu. Još jedna pretpostavka je da korisnik nastavlja postupak uz određenu vjerojatnost

– svakom korisniku nakon nekog vremena dosadi pregledavanje Interneta. Kada prestane odabirati veze, korisnik se zadržava na zadnjoj stranici koju je posjetio.

Vrijednost PR (PageRank) neke stranice može se izraziti sljedećom formulom:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}. \quad (3.1)$$

$PR(p)$ - vrijednost PageRank stranice p.

N - ukupni broj stranica.

$L(p)$ - broj veza na stranici p.

$M(p)$ - skup stranica čije veze pokazuju na p.

d - vjerojatnost da će korisnik nastaviti pregledavati Internet, tj. da će odabrati jednu od veza na stranici p.

Početni PageRank svake stranice je jednak. Iterativnim računom brzo se dolazi do konvergencije i PageRank se stabilizira. Izraz 3.1 može se napisati ovako:

$$R = \begin{bmatrix} (1-d)/N \\ (1-d)/N \\ \vdots \\ (1-d)/N \end{bmatrix} + d \cdot \frac{A}{\|A\|_1} \cdot R. \quad (3.2)$$

A je matrica susjedstva s elementima $a_{ij} = \begin{cases} 1, & \text{postoji veza od } p_j \text{ prema } p_i \\ 0, & \text{inače} \end{cases}$

$\|A\|_1$ je 1-norma matrice A.

Vektor $R = \begin{bmatrix} PR(p_1) \\ PR(p_2) \\ \vdots \\ PR(p_N) \end{bmatrix}$ sadrži vrijednosti PageRank svake stranice.

Vektor R može se računati iterativno tako da se u jednadžbi 3.2 na desnoj strani koristi stara vrijednost vektora R, a na lijevoj računa se nova vrijednost. Do konvergencije dolazi nakon malog broja iteracija, što je jedna od prednosti PageRank metode.

3.2. Metoda čvorišta i autoriteta

Uvjerivši se u učinkovitost PageRank metode, znanstvenici su sve više proučavali nove načine analize veza među web stranicama. Još jedan od razloga zašto je bilo potrebno naći neku novu metodu jest taj što je PageRank metoda patentirana. Kleinberg [13] predlaže novu metodu koja za dani korisnikov upit (pretragu) ocjenjuje koliko su stranice u rezultatima dobra čvorišta (engl. *hubs*) i koliko su stranice dobri autoriteti (engl. *authorities*). Na primjer, za pretragu „ZEMRIS“ dobar autoritet bila bi stranica www.fer.hr, a dobro čvorište stranica koja ima vezu na www.fer.hr, npr. www.carnet.hr/clanice/punopravne. Za danu pretragu korisniku tražilice prvo bi se predstavile stranice koje su dobri autoriteti.

Za razliku od PageRank metode, gdje je vrijednost PageRank unaprijed izračunata za svaku stranicu neovisno o upitu, metoda čvorišta i autoriteta dinamički, tj. nakon što korisnik pošalje upit, računa za neki reducirani skup stranica S koliko je svaka stranica dobro čvorište ili autoritet.

Skup S trebao bi biti relativno malen, ali u isto vrijeme i sadržavati većinu stranica koje su dobri autoriteti za zadani upit. Skup S gradi se na sljedeći način:

1. Nađe se skup stranica S_0 koji se sastoji od prvih *t* rezultata pretrage koristeći pretraživač koji uzima u obzir samo tekst stranice (npr. AltaVista[5]).

[5] http;//www.altavista.com

2. $S \leftarrow S_0$.
3. Skup S proširi se skupom svih stranica na koje pokazuju veze stranica iz skupa S_0.
4. Skup S proširi se skupom svih stranica koje pokazuju na stranice iz skupa S_0. Uzima se u obzir najviše *d* ulaznih veza svake stranice iz skupa S_0.

Prema metodi koju opisuje Kleinberg [13], odabiru se vrijednosti $t = 200$ i $d = 50$. Tvrdi se da tako izgrađen skup S sadrži većinu autoriteta za zadani upit, te je i dalje relativno malen, a stranice skupa S međusobno su dobro povezane. Te karakteristike važne su za postupak opisan u sljedećem poglavlju.

3.2.1. Nalaženje stranica autoriteta

Glavna ideja Kleinbergovog članka jest ta da su stranice dobri autoriteti ne samo ako imaju puno ulaznih veza, već i ako se stranice koje pokazuju na njih preklapaju. Postavljaju se sljedeće uzajamno podupiruće pretpostavke:

stranica je dobar autoritet ako su stranice koje pokazuju na nju dobra čvorišta

stranica je dobro čvorište ako pokazuje na dobre autoritete

Svakom vrhu i, tj. stranici p_i, pridružit ćemo nenegativne vrijednosti a_i (a dolazi od engl. *authority*) i h_i (h dolazi od engl. *hub*) koje označavaju koliko je stranica p_i dobar autoritet ili dobro čvorište. Veće vrijednosti znače da je stranica bolji autoritet, odnosno bolje čvorište.

Koristeći skup S možemo konstruirati usmjereni graf $G = (V, E)$ čiji su vrhovi stranice iz skupa S, a bridovi veze među tim stranicama. Formalno, $V = \{i : p_i \in S\}$, te $E = \{(i,j) : p_i, p_j \in V \wedge p_i \in M(p_j)\}$. Sada možemo numerički zapisati dvije pretpostavke:

$$a_j \leftarrow \sum_{i:\,(i,j)\in E} h_i,$$
$$h_j \leftarrow \sum_{i:\,(j,i)\in E} a_i. \qquad (3.3)$$

Ako s B označimo matricu susjedstva grafa G, tj. matricu u kojoj je element u retku i i stupcu j jednak broju bridova između vrhova i i j, onda relacije 2.3 možemo zapisati na sljedeći način:

$$\begin{bmatrix} h \\ a \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} h \\ a \end{bmatrix}_k, \qquad k = 0, 1, \ldots$$

Skraćeno zapisano,

$$x_{k+1} = M\, x_k, \qquad k = 0, 1, \ldots \qquad (3.4)$$
$$x_k = \begin{bmatrix} h \\ a \end{bmatrix}_k, \; M = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}.$$

Zanimaju nas jedino relativni rezultati, pa stoga promatramo normirani niz gdje vektore x_k normiramo Euklidskom normom $\| \quad \|_2$:

$$z_{k+1} = \frac{M\, z_k}{\|M\, z_k\|_2}, \qquad k = 0, 1, \ldots$$

Idealno bi bilo kada bi mogli definirati vrijednosti h i a kao limes vektora z_k kada $k \to \infty$. S takvom definicijom postoje dva problema. Prvi problem je taj što limes u velikom broju slučajeva i ne postoji, već alternira između vrijednosti $z_{parni} = \lim_{k\to\infty} z_{2k}$ i $z_{neparni} = \lim_{k\to\infty} z_{2k+1}$. Drugi problem jest odabir početne vrijednosti z_0. Naime, vrijednosti z_{parni} i $z_{neparni}$ ovise o z_0, pa ćemo ih i označavati sa $z_{parni}(z)$ i $z_{neparni}(z)$.

U Kleinbergovom [13] članku odabire se $z_0 = \mathbf{1}$, gdje $\mathbf{1}$ predstavlja vektor ili matricu sa svim elementima jednakim 1. Blondel [3] obrazlaže odabir $z_0 = \mathbf{1}$ time što se tada mogu jednostavno izračunati $z_{parni}(\mathbf{1})$ i $z_{neparni}(\mathbf{1})$, a $z_{parni}(\mathbf{1})$ je u tom slučaju najveći vektor od svih $z_{parni}(z)$ i $z_{neparni}(z)$ gledano po 1-normi.

Formalni dokaz za postojanje $z_{parni}(\mathbf{1})$ nećemo iznositi jer je detaljno opisan u Kleinbergovom i Blondelovom članku, ali ćemo iznijeti skicu. Vektori u nizu z_k imaju smjerove (ne nužno i magnitude)

$$x_k, Mx_k, M^2x_k, \ldots$$

ili

$$\begin{bmatrix} h \\ a \end{bmatrix}_k, \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}\begin{bmatrix} h \\ a \end{bmatrix}_k, \begin{bmatrix} BB^T & 0 \\ 0 & B^TB \end{bmatrix}\begin{bmatrix} h \\ a \end{bmatrix}_k, \ldots$$

iz čega vidimo da je:

$$\begin{bmatrix} h \\ a \end{bmatrix}_{k+2} = t\begin{bmatrix} BB^T & 0 \\ 0 & B^TB \end{bmatrix}\begin{bmatrix} h \\ a \end{bmatrix}_k.$$

Sa t označavamo faktor koji uzima u obzir magnitudu, tj. normalizaciju vektora. Raspišemo li jednadžbe, dobijemo:

$$h_{k+2} = tBB^Th_k, \quad a_{k+2} = t\,B^TBa_k,$$

te,

$$h_{2k} = (BB^T)^k t\mathbf{1}, \quad a_{2k} = (B^TB)^k t\mathbf{1}.$$

Ako je neka matrica M simetrična, a svojstveni vektor w koji odgovara dominantnoj (po apsolutnoj vrijednosti najvećoj) svojstvenoj vrijednosti matrice M nije okomit na neki vektor v, onda $\lim_{k\to\infty} M^k v = w$ [10]. Također, ako M nema negativnih vrijednosti, nema ih ni w.

Ako znamo da je $v = t\mathbf{1}$ i da su matrice BB^T i B^TB simetrične i nemaju negativnih elemenata (jer ni matrica susjedstva B nema negativnih elemenata) onda vrijedi $w \cdot t\mathbf{1} > 0$ (jer je suma komponenti vektora w pozitivna). Zato vrijedi i da su $\lim_{k\to\infty}(BB^T)^k t\mathbf{1}$ i $\lim_{k\to\infty}(B^TB)^k t\mathbf{1}$ dominantni svojstveni vektori matrica BB^T i B^TB.

Za računanje egzaktnih vrijednosti a_i i h_i svake stranice mogu se naći dominantni svojstveni vektori matrica BB^T i B^TB. Budući da za potrebe rangiranja web-stranica nisu potrebne egzaktne vrijednosti, već samo aproksimacije, a vrijednosti a_i i h_i konvergiraju brzo, iterativni postupak češće se koristi.

3.3. Sličnost među vrhovima dva grafa

U daljnjem tekstu opisujemo kako se u Blondelovom članku generalizira metoda čvorišta i autoriteta.

Neka je $G = (V, E)$ graf kojemu se računale vrijednosti a_i i h_i, tj. mjere koliko je neka stranica dobar autoritet ili dobro čvorište za određeni upit. Vrijednost a_i vrha i možemo poistovjetiti s mjerom sličnosti vrha i i vrha *autoritet* sljedećeg grafa:

$$čvorište \longrightarrow autoritet.$$

Slično, vrijednost h_i možemo izjednačiti s mjerom sličnosti vrha i i vrha $čvorište$. Za računanje sličnosti grafa G s grafom $čvorište \longrightarrow autoritet$ može se koristiti metoda čvorišta i autoriteta.

Ako vrijednosti h_i i a_i označimo redom s x_{i1} i x_{i2} onda izraz 3.3 možemo zapisati ovako:

$$\begin{aligned} x_{i1} &\leftarrow \textstyle\sum_{j:\,(i,j)\in E} x_{j2}, \\ x_{i2} &\leftarrow \sum_{j:\,(j,i)\in E} x_{j1}. \end{aligned} \qquad (3.5)$$

Postavlja se pitanje mogu li se relacije 3.5 poopćiti na složenije grafove od grafa oblika $1 \longrightarrow 2$. Prvo ćemo pokušati pokazati kako bi se računala matrica sličnosti s grafom oblika $1 \longrightarrow 2 \longrightarrow 3$. Za računanje sličnosti vrhova grafa G i vrhova 1 i 3 grafa $1 \longrightarrow 2 \longrightarrow 3$ relacije će biti slične relaciji 3.5. Nešto drukčija će biti relacija za računanje sličnosti s vrhom 2:

$$\begin{aligned} x_{i1} &\leftarrow \sum_{j:\,(i,j)\in E} x_{j2}, \\ x_{i2} &\leftarrow \sum_{j:\,(j,i)\in E} x_{j1} + \sum_{j:\,(i,j)\in E} x_{j3}, \end{aligned} \qquad (3.6)$$

$$x_{i3} \leftarrow \sum_{j:\,(j,i)\in E} x_{j2}.$$

Po uzoru na izraz 3.4 pišemo:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B & 0 \\ B^T & 0 & B \\ 0 & B^T & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_k, \qquad k = 0, 1, \ldots \tag{3.7}$$

Sva promatranja vezana uz granične vrijednosti koja su vrijedila za graf $1 \longrightarrow 2$ vrijede i za graf $1 \longrightarrow 2 \longrightarrow 3$. Za početnu vrijednost odabire se $[1 \quad 1 \quad 1]^T$. Sličnost grafa G s grafom $1 \longrightarrow 2 \longrightarrow 3$ definiramo kao graničnu vrijednost za parne elemente niza iz relacije 3.7.

U općem slučaju promatramo grafove $G_A = (V_A, E_A)$ i $G_B = (V_B, E_B)$, uz $n_A = |V_A|$ i $n_B = |V_B|$, te računamo matricu sličnosti $X = (x_{ij})$ dimenzija $n_B \times n_A$ na sljedeći način:

$$x_{ij} \leftarrow \sum_{\substack{r:\,(r,i)\in E_B \\ s:\,(s,j)\in E_A}} x_{rs} + \sum_{\substack{r:\,(i,r)\in E_B \\ s:\,(j,s)\in E_A}} x_{rs}, \qquad \begin{matrix} i \in \{1, \ldots, n_B\} \\ j \in \{1, \ldots, n_A\} \end{matrix} \tag{3.8}$$

Operacija nalaženja matrice sličnosti može se promatrati kao produkt grafova G_A i G_B čiji je rezultat graf $G_{A\times B} = (V_{A\times B}, E_{A\times B})$, gdje je

$$V_{A\times B} = \{(i,j) : i \in V_A \wedge j \in V_B\},$$
$$E_{A\times B} = \{((i_1, j_1), (i_2, j_2)) : (i_1, i_2) \in E_A \wedge (j_1, j_2) \in E_B\}.$$

U takvom grafu vrijednost svakog vrha jednaka je sumi vrijednosti svih susjednih vrhova. Vrijednost vrha (i,j) novonastalog grafa jednaka je elementu matrice sličnosti u retku i i stupcu j.

Promatrajući izraz 3.7 vidimo da se izraz 3.8 može napisati u kraćem obliku:

$$X_{k+1} = BX_kA^T + B^TX_kA, \tag{3.9}$$

gdje su A i B matrice susjedstva grafova G_A i G_B.

Izraz u 3.9 konvergira za parne k uz $X_0 = \mathbf{1}$ (prisjetimo se, $\mathbf{1}$ nije jedinična matrica već matrica ili vektor čiji su svi elementi jednaki 1). Dokaz ove tvrdnje može se naći u Blondelovom članku [3],.

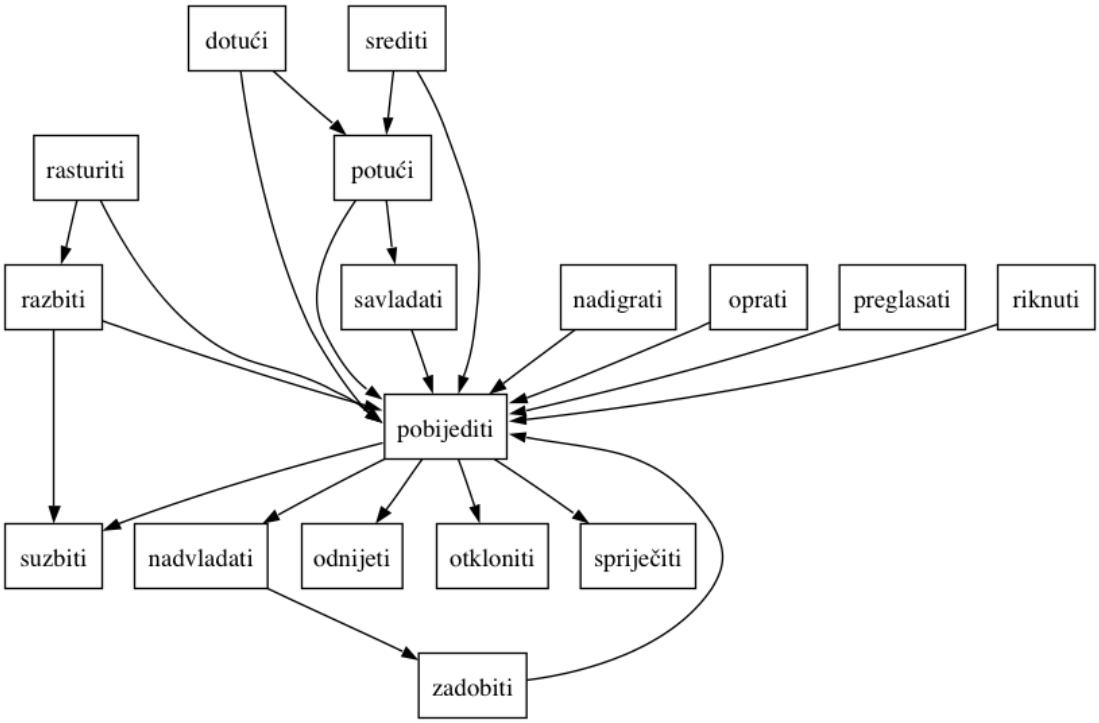
4. Nalaženje sinonima u rječniku

4.1. Primjena matrice sličnosti

Možemo se prisjetiti da se u metodi čvorišta i autoriteta izdvajao onaj podgraf grafa povezanosti svih web stranica koji je sadržavao prvih 200 rezultata korisničke pretrage. Zatim se taj podgraf proširio koristeći ulazne i izlazne veze od početnih 200 stranica. Na kraju se tražila sličnost vrhova dobivenog podgrafa s vrhom 2 grafa $1 \longrightarrow 2$. Ta sličnost predstavljala je mjeru autoritativnosti web stranice.

Metoda nalaženja sinonima u rječniku koju opisuje Blondel [4] koristi englesko-engleski rječnik. Glavna pretpostavka u njegovom radu jest da sinonimi često imaju zajedničke riječi u svom opisu, te da se sinonimi često koriste u opisu istih riječi.

Prvi korak algoritma nalaženja sinonima jest konstruiranje **grafa rječnika** G_R čiji su vrhovi sve riječi u rječniku. Postoji usmjereni brid od riječi *a* do riječi *b* u tom grafu samo ako se riječ *b* pojavljuje u opisu riječi *a*. Po uzoru na postupak koji opisuje Kleinberg [13] drugi korak algoritma jest konstrukcija grafa G, podgrafa grafa G_R za zadanu riječ w (upit). Graf G sadrži samo one vrhove iz grafa G_R koji su prvi susjedi vrha w i sam vrh w. Graf G nazivamo **graf susjedstva** riječi w.



Slika 1. Graf susjedstva riječi *pobijediti*.

Riječi u grafu susjedstva trebamo poredati po važnosti, tj. za svaku riječ trebamo procijeniti koliko je ta riječ dobar sinonim riječi $\boldsymbol{w}$. Razumna ideja jest da rangiramo vrhove prema tome koliko su oni dobra čvorišta ili autoriteti. Drugim riječima, mogli bismo izračunati sličnost grafa $\boldsymbol{G}$ sa grafom $\mathbf{1} \longrightarrow \mathbf{2}$, te koristiti sličnost sa vrhom $\mathbf{1}$ ili sličnost sa vrhom $\mathbf{2}$ grafa $\mathbf{1} \longrightarrow \mathbf{2}$. Time bi postupak bio gotovo identičan postupku rangiranja web stranica metodom čvorišta i autoriteta. Ipak, rezultati dobiveni takvom metodom nisu zadovoljavajući. Razlog tome je što pravi sinonimi u ovako konstruiranom grafu susjedstva riječi $\boldsymbol{w}$ ne odgovaraju ni pojmu čvorišta ni pojmu autoriteta. Na primjeru grafa susjedstva riječi *pobijediti* (slika 1), riječ *suzbiti* će imati relativno visok rezultat kao autoritet, a riječi *rasturiti*, *dotući*, *srediti* relativno visok rezultat kao čvorišta. Ipak, očito je da te riječi nisu sinonimi jer nemaju ni jednu izlaznu ili pak nemaju ni jednu ulaznu vezu.

U Blondelovom članku [4] predlaže se umjesto mjerenja sličnosti grafa G sa vrhovima grafa $1 \longrightarrow 2$ mjerenje sličnosti grafa G sa vrhom 2 grafa $1 \longrightarrow 2 \longrightarrow 3$. Naime, početna pretpostavka bila je da se više sinonima nalazi

u opisu zajedničke riječi i da sinonimi u svom opisu imaju zajedničke riječi. Mjereći sličnost vrhova grafa G s vrhom 2 zapravo mjerimo koliko strukture tih riječi odgovaraju strukturi tipičnog sinonima. Usporedimo li međusobno te vrijednosti, redovito će riječ w biti najbolji sinonim (tj. riječ će sama sebi biti vrlo dobar sinonim).

4.2. Homografi i polisemi

Homografi su riječi koje se isto pišu, ali imaju različito značenje. U rječnicima se obično oni homografi koji su različite vrste riječi smatraju zasebnim pojmovima, tj. odvojeno su definirani. Npr. riječ „liječnički“ može biti pridjev (koji se odnosi na liječnike) ili prilog (kao liječnik, na način liječnika). Polisemi su riječi koje se isto pišu, imaju različita značenja, ali dijele istu osnovu. Npr. riječ „brod“ je imenica muškog roda, ali može predstavljati plovilo ili uzdužni dio crkve. Svi polisemi obično su navedeni u definiciji jednog pojma. U rječnicima je uz svaku definiciju pojma navedena i vrsta riječi (imenica, zamjenica, itd.).

Za pojavnicu *B* koja se nalazi u opisu riječi *A*, ne možemo uvijek odrediti o kojoj vrsti riječi se radi. To se događa u slučajevima ako postoje riječi druge vrste koje se pišu kao i riječ *B* (tj. homografi su). Prilikom konstrukcije grafa rječnika u idealnom slučaju povezali bi smo *A* s *B* samo ako su *A* i *B* iste vrste. Nema smisla za sinonim imenice kao rezultat dati glagol.

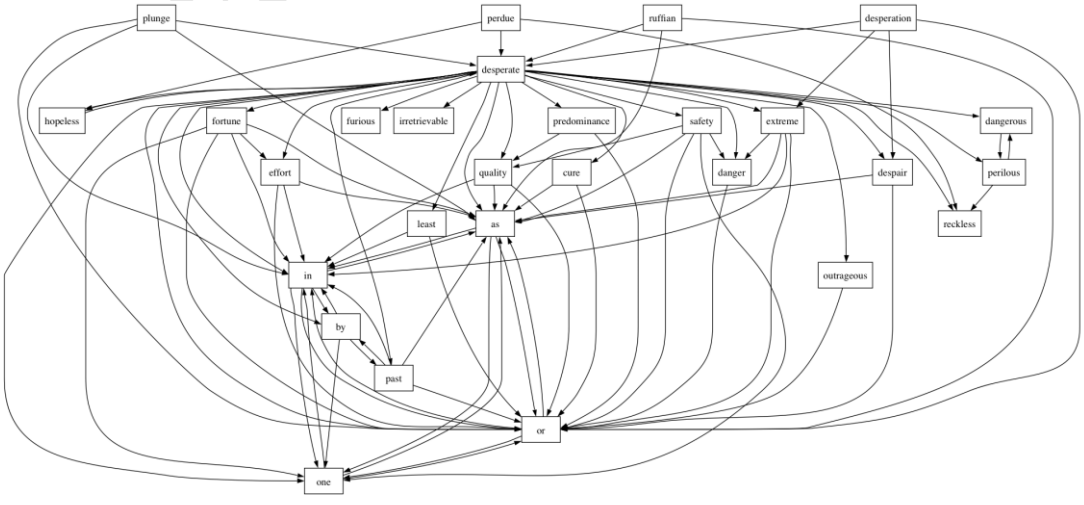
Problem možemo riješiti tako da provedemo pojednostavljenja navedena u Blondelovom članku [4]:

> opisi svih homografa i polisema spojeni su u jedan opis i promatraju se kao jedna riječ;
>
> reći ćemo da je vrsta novonastale riječi skup vrsta svih riječi od kojih je nastala;
>
> kako nema smisla tražiti sinonime među riječima različite vrste, prilikom konstrukcije grafa povezanosti riječi spojiti ćemo dvije riječi samo ako

njihove vrste imaju neprazan presjek (tj. riječi od kojih se sastoje imaju barem jednu zajedničku vrstu).

4.3. Problem zaustavnih riječi

Prilikom konstrukcije grafa susjedstva riječi $\boldsymbol{w}$ često se događa da $\boldsymbol{w}$ ima vezu prema nekoj zaustavnoj riječi (npr. engleske riječi *of, and, by,* itd.). To samo po sebi ne bi trebao biti problem, jer je pretpostavka da će takve zaustavne riječi imati samo ulazne veze, a neće imati izlazne veze, budući da je mala vjerojatnost da određena zaustavna riječ ima vezu sa sinonimom kojeg tražimo. Zbog provjere da sinonimi moraju biti riječi iste vrste ili da im se barem vrste moraju preklapati, nameće se zaključak da neće postojati veze od zaustavnih riječi ili prema njima. Međutim, takve se veze ipak javljaju. Primjerice, riječ *by* može biti prijedlog, pridjev ili prilog, riječ *in* može biti prijedlog, prilog ili imenica, itd. Riječi *in* i *by* međusobno će se spominjati u opisima i zato će biti vrlo slične vrhu $\mathbf{2}$ grafa $\mathbf{1} \longrightarrow \mathbf{2} \longrightarrow \mathbf{3}$. Stoga će se te zaustavne riječi činiti kao idealni sinonimi bilo koje riječi (vidi sliku 2).



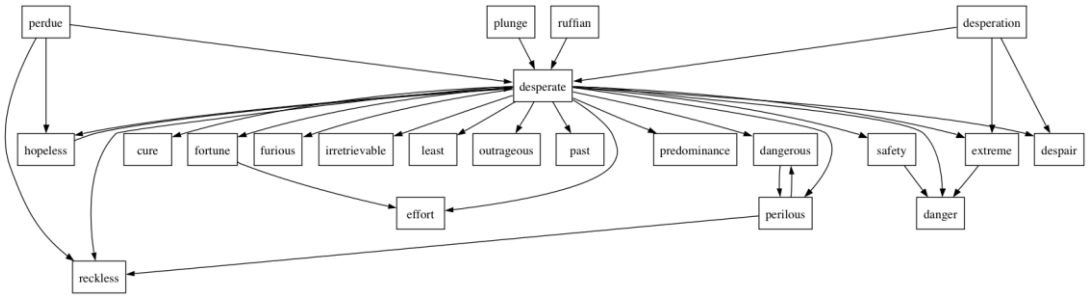
Slika 2. Graf susjedstva riječi *desperate* bez filtriranja zaustavnih riječi. *On, of, by, and* i *one* imaju jako puno ulaznih i izlaznih veza, te znatno pogoršavaju rad metode.

Ovaj problem može se jednostavno riješiti na način da se iz grafa izbace sve riječi koje se pojavljuju u više od k opisa. U slučaju engleskog rječnika odabran je $k = 1000$.

Zaustavna riječ	*Broj pojavljivanja*
of	67337
a	47188
the	43402
or	40835
to	31441
in	23600
as	22245
and	16491
an	13705
by	12061
one	11185
with	10753
which	10326
see	8422
is	8350

Tablica 1. Popis najčešćih zaustavnih riječi u engleskom rječniku.

Nakon što su izbačene zaustavne riječi, grafovi susjedstva postaju znatno jednostavniji i upotrebljiviji (slika 3).



Slika 3. Graf susjedstva riječi *desperate* znatno je jednostavniji nakon što su izbačene zaustavne riječi.

Negativan utjecaj zaustavne riječi ovisi o načinu sastavljanja rječnika, a ponajviše o samom jeziku, pa će o tome i ovisiti izbor k.

5. Primjena na indeksiranje dokumenata

Indeksiranje dokumenata jest postupak dodjeljivanja jedne ili više oznaka ili pojmova svakom dokumentu. Skup pojmova kojima se dokumenti označuju, tj. indeksiraju, unaprijed je određen. Ako su dokumenti kvalitetno indeksirani, pretraživanje dokumenata može biti znatno lakše. Umjesto da se pretraga vrši nad cijelim skupom dokumenata, može se ograničiti samo na manji dio koji je označen pojmom vezanim uz pretragu. Indeksiranje se može vršiti ručno, poluautomatski ili automatski. U prvom slučaju čovjek sam odabire pojmove za koje misli da najbolje opisuju dokument. U slučaju automatskog indeksiranja računalni program sam odabire sve pojmove koji odgovaraju nekom dokumentu, a kod poluautomatskog indeksiranja računalo samo predlaže skraćenu listu pojmova koja koristi čovjeku da brže indeksira dokument. Samo indeksiranje može se svesti na problem klasifikacije, tj. svrstavanje dokumenta u jednu ili više kategorija.

Uspješnost metoda koje poboljšavaju kvalitetu pretrage kod Internet tražilica uvelike je ovisila o strukturi grafa koji opisuje veze među web stranicama. U mnogim područjima koja nisu vezana uz pretragu web-a pokušavaju se iskoristiti slične metode tako da se pokuša izgraditi struktura slična web-u. Na primjer, za automatsko generiranje sažetaka radova razvijen je algoritam LexRank [9] po uzoru na algoritam PageRank. Dokumenti su razdijeljeni u manje segmente, a slični segmenti su međusobno povezani. Metodom LexRank mjeri se korisnost svake rečenice za uključivanje u sažetak. Druga primjena je rješavanje višeznačnosti riječi uz korištenje PageRank algoritma, koju je opisao Agirre [1]. Sve riječi koje su se pojavljivale uz višeznačnicu u korpusu činile su graf. Dvije riječi bile su povezane ako su se pojavile zajedno u grafu. Na kraju, prikazana je i primjena u slučaju traženja sinonima zadane riječi. Ono što je svojstveno svakoj od navedenih primjena jest inovativan način na koji su se konstruirali grafovi.

Treba primijetiti da se ni u jednoj primjeni nije koristilo znanje naučeno na primjerima. Metoda za rangiranje stranica nije iskoristila podatke nekih

korisnika da im ta stranica za određeni upit odgovara ili ne odgovara. Slično tome, metoda nalaženja sinonima ne može unaprijed iskoristiti primjere za učenje te pokušati iz njih generalizirati naučeno. To nas navodi na sljedeći zaključak: ako bi za svrhu klasifikacije konstruirali graf koji na neki način ima ugrađene informacije dobivene od rezultata klasifikacije dobivenih na primjerima za učenje, te informacije ne bismo mogli iskoristiti tijekom rada na skupu za testiranje. Zato ostaje otvoreno pitanje kako iskoristiti podatke za učenje ako primjenjujemo metode analize grafova.

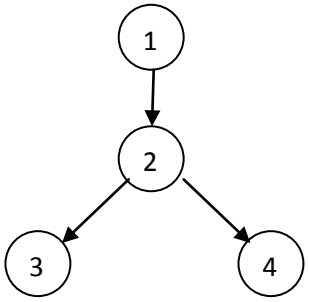
5.1. Nalaženje sličnih dokumenata

Jedan od načina za iskorištavanje rezultata indeksiranja dobivenih na primjerima za učenje jest da novi dokument usporedimo s već indeksiranim dokumentima i iskoristimo pojmove kojima su indeksirani najsličniji dokumenti. Time svodimo problem na problem traženja sličnih dokumenata. Odabir strukture grafa kojim bi to ostvarili otvoren je problem.

Kako sada na raspolaganju imamo operaciju nalaženja sličnosti među vrhovima dva grafa, logično rješenje bilo bi da konstruiramo grafove na način da usporedbom sličnosti vrhova dva grafa dobijemo mjeru sličnosti među dokumentima koje predstavljaju ti vrhovi. Barem jedan od tih grafova trebao bi imati ugrađene informacije o nekim svojstvima tog dokumenta. Svojstvo dokumenta mogu biti, primjerice, riječi koje se pojavljuju u dokumentu. Kako tekstni dokumenti mogu biti vrlo veliki, a njihov broj može biti također velik, bilo bi teško analizirati dobiveni graf radi njegove veličine i vremena potrebnog za analizu. Jedno od svojstava dokumenta jesu pojavljivanja pojmova kojima pokušavamo indeksirati te iste dokumente, tako da njih možemo koristiti pri izgradnji grafa.

Operacija nalaženja sličnosti među grafovima funkcionira i na primjeru jednog grafa, tj. kad tražimo samosličnost nekog grafa [3]. Na slici 4 vidimo primjer matrice samosličnosti. Izgleda da je dovoljno da izgradimo graf kojemu će

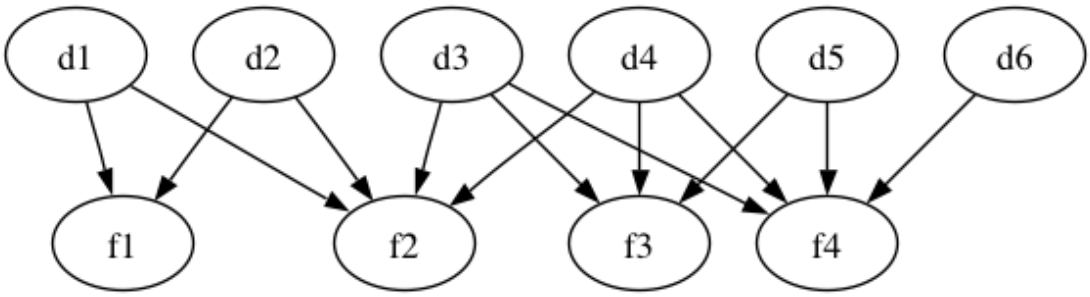
neki čvorovi predstavljati dokumente, a drugi čvorovi svojstva tih dokumenata. Kao moguća svojstva dokumenata možemo uzeti sve pojmove kojima indeksiramo dokumente. Ako se u tekstu dokumenta pojavljuju pojmovi kojima inače indeksiramo dokumente, onda možemo u grafu povezati dokument s vrhovima koji predstavljaju te pojmove. Sličnost među dokumentima nalazit ćemo računajući sličnost među vrhovima koji predstavljaju te dokumente.



$$\begin{bmatrix} 0.41 & 0 & 0 & 0 \\ 0 & 0.41 & 0 & 0 \\ 0 & 0 & 0.41 & 0.41 \\ 0 & 0 & 0.41 & 0.41 \end{bmatrix}$$

Slika 4. Matrica sličnosti grafa sa samim sobom.

Jedan graf koji zadovoljava ta svojstva je graf prikazan na slici 5.



Slika 5. Graf u kojemu je svaki od šest dokumenata spojen sa vrhovima koji predstavljaju svojstva tih dokumenata. Vidimo da su dokumenti *d1* i *d2*, te *d3* i *d4* slični.

Matricu sličnosti računamo iterativnim postupkom, a njena veličina može biti npr. 5000 (broj dokumenata) s 10000 (broj svojstava). Matrica susjedstva je tada veličine 15000 s 15000. Zato trebamo pojednostaviti izraze za računanje matrice sličnosti za ovaj slučaj. Ako s A označimo matricu susjedstva i uvrstimo u izraz 3.9 za računanje matrice sličnosti, dobijemo sljedeći izraz:

$$X_{k+1} = AX_kA^T + A^TX_kA. \qquad (5.1)$$

Matrica A ima sljedeći oblik:

$$A = \begin{bmatrix} \mathbf{0} & B \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \qquad (5.2)$$

gdje je B matrica dimenzija $M \times N$, M broj dokumenata, a N broj svojstvenih vrhova. $\mathbf{0}$ označava nul-matricu ili vektor. Uvrstimo 5.2 u 5.1, te uvrstimo vrijednost za $X_{k=0} = \mathbf{1}_{M\times N}$. $\mathbf{1}_{M\times N}$ je matrica dimenzija $M \times N$ čiji su svi elementi jednaki 1. Dobijemo sljedeći izraz:

$$X_1 = \begin{bmatrix} \mathbf{0} & B \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{1}_{M\times M} & \mathbf{1}_{M\times N} \\ \mathbf{1}_{N\times M} & \mathbf{1}_{N\times N} \end{bmatrix}\begin{bmatrix} \mathbf{0} & \mathbf{0} \\ B^T & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ B^T & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{1}_{M\times M} & \mathbf{1}_{M\times N} \\ \mathbf{1}_{N\times M} & \mathbf{1}_{N\times N} \end{bmatrix}\begin{bmatrix} \mathbf{0} & B \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

Nakon sređivanja:

$$X_1 = \begin{bmatrix} B\mathbf{1}_{N\times N}B^T & \mathbf{0} \\ \mathbf{0} & B^T\mathbf{1}_{M\times M}B \end{bmatrix}. \qquad (5.3)$$

Računamo X_2 uvrštavanjem 5.3 u 5.1:

$$X_2 = \begin{bmatrix} BB^T\mathbf{1}_{M\times M}BB^T & \mathbf{0} \\ \mathbf{0} & B^TB\mathbf{1}_{N\times N}B^TB \end{bmatrix}.$$

Sada je očito da je:

$$X_{2k} = \begin{bmatrix} (BB^T)^k\mathbf{1}_{M\times M}(BB^T)^k & \mathbf{0} \\ \mathbf{0} & (B^TB)^k\mathbf{1}_{N\times N}(B^TB)^k \end{bmatrix}.$$

Nas ne zanima sličnost među dokumentima i svojstvenim vrhovima, već samo među dokumentima. Zato nam je potreban samo gornji lijevi dio matrice. Dobivamo S, matricu sličnosti među dokumentima:

$$S_k = (BB^T)^k\mathbf{1}_{M\times M}(BB^T)^k.$$

Bez dokaza navodimo da je $\lim_{k}(S_k)^k = ww^T$, gdje je w dominantni svojstveni vektor matrice BB^T. Takvu matricu koristit ćemo za usporedbu dva dokumenta. Znači:

$$\lim_{k\to\infty} X_{2k} = \begin{bmatrix} ww^T & \mathbf{0} \\ \mathbf{0} & qq^T \end{bmatrix}, \qquad (5.4)$$

gdje je w dominantni svojstveni vektor matrice BB^T, a q dominantni svojstveni vektor matrice B^TB. Budući da ćemo koristiti samo gornji lijevi dio matrice, tj. relaciju sličnosti među dokumentima, trebamo izračunati samo vektor w.

Sada možemo iskoristiti metodu poput metode traženja k najbližih susjeda [8], gdje se nekom dokumentu pronađe, koristeći određenu metriku, k najbližih susjeda. Dokument se svrsta u kategoriju u koju je svrstano najviše od tih k dokumenata. Da bi ocijenili rad metode koristit ćemo sljedeći vrlo jednostavan sustav.

5.2. Osnovni sustav

Neka je $P = \{p_1, p_2, \ldots, p_N\}$ skup pojmova kojim indeksiramo dokumente, a $D = \{d_1, d_2, \ldots, d_N\}$ skup dokumenata. Načinit ćemo matricu $A_{M\times N} = (a_{ij})$, u kojoj je element a_{ij} jednak broju pojavljivanja pojma p_j u dokumentu d_i. Neka su a_{it_1}, a_{it_2}, …, a_{tk} k najvećih elemenata i-tog retka matrice A, gdje je k neka unaprijed utvrđena konstanta. Tada ćemo indeksirati dokument d_i pojmovima p_{t_1}, p_{t_2}, …, p_{tk}. Očekujemo da ćemo postići poboljšanje ako umjesto broja ponavljanja pojma u dokumentu za element matrice a_{ij} upotrijebimo TFIDF mjeru [12] na elemente matrice A.

TFIDF računamo na sljedeći način:

$$TFIDF = TF \cdot IDF,$$

gdje je:

$$TF = \text{broj pojavljivanja pojma } p \text{ u dokumentu } d$$

$$IDF = \log\left(\frac{\text{broj dokumenata}}{\text{broj dokumenata koji sadrže pojam p}}\right)$$

6. Rezultati

6.1. Nalaženje sinonima

U ovom poglavlju prikazani su rezultati metode nalaženja sinonima, te rezultati eksperimenata s automatskim indeksiranjem deskriptora. Radi provjere valjanosti algoritma i boljeg shvaćanja rada, prvi cilj pri izradi programa je bio pokušaj repliciranja rezultata iz Blondelovog rada [4]. Nakon što je provjerena ispravnost implementacije algoritma, krenulo se u obavljanje eksperimenta s hrvatskim jezikom.

6.1.1. Nalaženje sinonima u engleskom jeziku

Na temelju projekta *„Project Gutenberg Etext of Webster's Unabridged Dictionary“* koji je temeljen na rječniku *„1913 US Webster's Unabridged Dictionary“* nastao je *„Online Plain Text English Dictionary“* [16], rječnik koji koristimo pri nalaženju sinonima.. Taj rječnik sastoji se od 27 HTML datoteka formatiranih na vrlo jednostavan način: svaka definicija u rječniku pojavljuje se u zasebnom redu datoteke. Definicija, vrsta i opis riječi omeĐeni su uvijek istim HTML tagovima. Homografi (riječi koje se jednako pišu, a imaju različito značenje) i polisemije (slučaj kada više oblika imaju istu osnovu, ali različita i nesrodna značenja) imaju zaseban unos, tj. navedeni su u odvojenim redovima. Rječnik ima ukupno oko 176000 unosa, odnosno riječi, ako brojimo zasebno homografije i polisemije.

Jedan od problema u ovom rječniku je taj što definicija svake riječi počinje velikim početnim slovom. Radi lakšeg stvaranja grafa svih riječi u rječniku, sve su riječi i opisi riječi prebačeni u mala slova. OdreĐene kategorije riječi izbačene su iz rječnika:

definicije prefiksa ili sufiksa. Npr. *„Aero-“ ili „-escent“,*

fraze i drugi višerječni unosi jer je teško odrediti kada određeni niz riječi u opisu predstavlja više zasebnih riječi ili jedan unos, te
sve definicije koje sadrže znamenke.

Nakon spajanja opisa kako je opisano u poglavlju 4.2, broj unosa smanjio se na 111617 riječi (za usporedbu, Blondel je dobio 112169 riječi). Svaki spojeni unos označen je jednom ili više od ukupno 309 vrsta riječi. Ukupan broj pojavnica korišten u opisima svih riječi je otprilike 1957000.

6.1.2. Evaluacija

Za evaluaciju automatskog nalaženja sinonima obično se koriste dva pristupa. Jedan pristup evaluaciji jest automatska usporedba s ručno izrađenim rječnikom sinonima [24], npr. s rječnikom WordNet [1]. Problem kod tog pristupa jest što ručno izrađenim rječnicima često nedostaju mnogi sinonimi. Zato će preciznosti rječnika koji se ocjenjuje često biti preniska. Drugi pristup je ljudska evaluacija više različitih metoda (npr. [4]). Jasno je da su mane ovog pristupa vremenska zahtijevnost, te nekonzistentnost. Ponekad se koristi i hibridni pristup, tj. uz automatsku evaluaciju s ručno izrađenim rječnikom vrši se ljudska evaluacija nad podskupom riječi koji nije pokriven u ručno izrađenom rječniku [17].

U Blondelovom članku [4] uspoređuje se pet metoda nalaženja sinonima: mjera udaljenosti, ArcRank [11], rječnik WordNet [15] i rječnik iz programa Microsoft Word 97. Svakom metodom nađeni su sinonimi za engleske riječi *disappear*, *sugar*, *parallelogram*, *science*. Deset osoba ocijenilo je svaku listu sinonima ocjenom 0-10. Lako se uočava propust ove metode: ispitane su samo četiri riječi iz rječnika od preko 100000 riječi.

6.1.3. Usporedba s implementacijom u Blondelovom članku

Ponovno ćemo naći sinonime za četiri riječi koje su navedene u Blondelovom radu [3]. To su riječi ***disappear***, ***parallelogram***, ***sugar*** i ***science***. Kao razlog zašto su baš ove četiri riječi odabrane navodi se prva riječ ima razne sinonime, druga ih nema, treća ima mnogo značenja, a četvrta je vrlo česta i teško je uopće reći što je njen sinonim.

Blondelov članak	*Ovaj rad*
vanish	vanish
pass	pass
die	die
wear	faint
faint	cease
fade	dissipate
sail	evanesce
light	wear
dissipate	light
ease	ship

Tablica 2. Sinonimi riječi *dissapear*.

Blondelov članak	*Ovaj rad*
square	square
rhomb	rhomb
parallel	parallel
gure	equal
prism	opposite
equal	figure
opposite	parallelopiped
angles	altitude
quadrilateral	quadrilateral
rectangle	rhomboid

Tablica 3. Sinonimi riječi *parallelogram*

Blondelov članak	*Ovaj rad*
cane	cane
starch	sucrose
sucrose	starch
milk	molasses
sweet	juice
dextrose	milk
molasses	sweet
juice	dextrose
glucose	glucose
lactose	lactose

Tablica 4. Sinonimi riječi *sugar*.

Blondelov članak	*Ovaj rad*
art	art
branch	branch
law	study
study	practice
practice	learning
natural	natural
knowledge	application
learning	skill
theory	theory
principle	history

Tablica 5. Sinonimi riječi *science*

Na primjeru tablice 2 do tablice 5 vidimo usporedbu implementacije algoritma u ovom radu i u Blondelovom radu. Za svaku riječ nađeno je deset najbolje rangiranih sinonima, s time da se sama riječ za koju su se tražili sinonimi nije uzimala u obzir. Postoje male razlike u dobivenim listama sinonima, ali one se mogu pripisati malo drukčijoj konstrukciji grafa riječi, te činjenicom da se u Blondelovom radu [4] transformiralo 309 kategorija riječi u kombinaciju 5 kategorija (detalji nisu navedeni). Unatoč manjim razlikama, možemo reći da su implementacije vrlo slične.

Učinkovitost algoritma isprobana je i na nekim drugim riječima. Primijećen je jedan sustavni nedostatak ove metode: u listi sinonima često se pojavljuju riječi suprotnog značenja. Neobično je što u svom radu Blondel [4] uopće ne spominje taj problem. U članku je čak i prikazan graf susjedstva riječi *likely*, ali ispod slike piše da graf nije potpun (ručno su odstranjeni neki vrhovi). Primjerice, među prvih 10 sinonima riječi *quick* i *likely* pojavljuju se riječi *slow* i *unlikely* koje imaju suprotno značenje.

6.1.4. Nalaženje sinonima u hrvatskom jeziku

Prilikom primjene opisane metode na riječi iz hrvatskog jezika javlja se veliki problem u izboru hrvatskog rječnika. Naime, metoda zahtijeva korištenje hrvatskog rječnika koji za svaku definiciju riječi ima opis te riječi na istom jeziku. Za razliku od engleskog jezika, besplatnih i javno dostupnih materijala za hrvatski jezik ima vrlo malo. Javno je dostupan hrvatski rječnik[6] koji ne sadrži ni opise ni vrste riječi. Dostupan je i englesko-hrvatsko-engleski rječnik[7], pa bi mogli probati koristiti taj rječnik za prijevod hrvatske riječ na englesku, a zatim za prijevod liste sinonima s engleskog jezika na hrvatski. Ta metoda bila bi vrlo neprecizna jer se u prijevodu svake hrvatske riječi nalazi nekoliko engleskih i obratno. Prijevodom samo u jednom smjeru (hrvatski u engleski) već znatno opada preciznost.

U ovom trenutku jedini dostupan hrvatsko hrvatski rječnik u elektronskom izdanju je „Veliki rječnik hrvatskoga jezika“ Vladimira Anića [2], pa su svi eksperimenti rađeni korištenjem izdanja tog rječnika na CD-u. Rječnik se sastoji od približno 70000 natuknica. Homografi se nalaze u zasebnim natuknicama, a svi polisemi u jednoj. U rječniku je korištena 21 vrsta riječi[8].

Za razliku od engleskog rječnika koji je korišten, u hrvatskom rječniku možemo razlikovati između riječi pisanih velikim i malim slovom. Opisi riječi nikada nisu pune rečenice i ne počinju velikim slovom samo zbog početka rečenice. Iz svake riječi uklonjeni su naglasci – npr. riječ nàstānjenōst zamijenjena je s riječi nastanjenost. Opisi riječi sadrže gramatičke natuknice, etimologije, frazeologije, te sintagme. Te dodatne informacije su uklonjene.

---

[6] http://cvs.linux.hr/spell/ispell/

[7] http://web.math.hr/~igaly/EHrjecnik.htm

[8] Vrsta riječi koja se spominje u tekstu ne odgovara nužno jednoj od 10 vrsta riječi u hrvatskom jeziku. U rječniku se ponekad pod vrstom riječi spominje „srednji rod“, što je specifičnije od vrste riječi „imenica“ u gramatici.

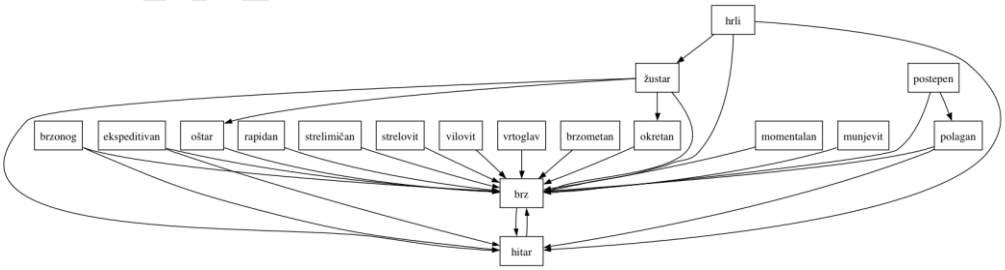
Nakon pretprocesiranja rječnika, tj. izbacivanja višerječnih ulaza, micanja natuknica koje su samo prefiksi i sufiksi, te spajanja svih homografa dobijemo oko 65000 riječi opisanih s oko 705000 riječi. Vidimo da engleski rječnik ima skoro dvostruko više riječi sa skoro tri puta dužim opisima. U slučaju hrvatskog jezika problem zaustavne riječi nije toliko izražen. Jedan od razloga je taj što većina zaustavnih riječi pripada vrsti riječi koja nije ista kao i vrsta riječi čije sinonime tražimo.

Automatsku evaluaciju u ovom trenutku gotovo je nemoguće napraviti jer ne postoji javno dostupan rječnik sinonima u elektronskom obliku. Zato ćemo samo navesti nekoliko primjera i proučiti njihove karakteristike.

brz	***jama***	***nestati***	***ukrasti***	***kotač***
hitar	rupa	umrijeti	zdipiti	kolotur
žustar	spilja	zaći	maznuti	velosiped
polagan	vapnenica	proći	odnijeti	zupčanik
hrli	budža	izvjetriti	drpnuti	točak
brzonog	rovača	zamaknuti	gepiti	kotur
ekspeditivan	golubinka	otići	šćapiti	čelnik
okretan	krečana	utonuti	otuĐiti	bucanj
oštar	fojba	utišati	dignuti	razmjer
postepen	graba	iščeznuti	mrknuti	senzor

Tablica 6. Deset predloženih sinonima za odabrane riječi

Za riječ *brz* predloženi su i sinonim *polagan* i *postepen*, što nikako nije dobro. Ako pogledamo graf susjedstva na slici 6, vidimo da riječi *polagan* i *postepen* spominju riječ *brz* u svom opisu. Kako riječ *postepen* spominje i riječ *polagan*, vrh riječi *polagan* postaje sličan vrhu 2 grafa $1 \longrightarrow 2 \longrightarrow 3$.



Slika 6. Graf susjedstva riječi *brz*.

Kod riječi *kotač* primjećuju se riječi koje nikako ne mogu biti sinonimi, jer je teško pronaći deset sinonima za riječ *kotač*. Svi ostali pronađeni sinonimi dobro su odabrani.

6.2. Automatsko indeksiranje dokumenata

Za ulazni skup korišten je skup dokumenata indeksiran pojmovnikom EUROVOC [5]. Pojmovnik EUROVOC koristi se prilikom indeksiranja službenih dokumenata institucija Republike Hrvatske koje provodi Hrvatska informacijsko-dokumentacijska referalna agencija (HIDRA). EUROVOC je hijerarhijski pojmovnik koji se sastoji od oko 6500 deskriptora (pojmova) svrstanih u 21 područje i 127 podpodručja. Preveden je na 20 službenih jezika Europske Unije, a preveden je i na hrvatski jezik. HIDRA je dodala oko 3700 pojmova specifičnih za Hrvatsku. Primjer dijela EUROVOCa može se vidjeti na slici 7.

Slika 7. Mali dio pojmovnika EUROVOC. Pojmovi kojima se indeksiraju dokumenti označeni su tamnom bojom.

Skup dokumenata na kojem je isprobavano indeksiranje sastoji se od 4556 službenih dokumenata Republike Hrvatske iz Narodnih novina. Dokumente je indeksirala HIDRA ručnim postupkom. U skupu je oko 3300

dokumenata indeksirano jednim pojmom, oko 1200 sa dva pojma, te svega 9 dokumenata sa tri pojma.

Za pronalaženje pojavljivanja pojmova u dokumentu korištena je metoda koju u svom radu opisuju Šarić i suradnici [21]. Pojmovi se mogu pojavljivati u bilo kojem padežu zahvaljujući korištenju lematizacijske metode „Automatska morfološka normalizacija“ koju je razvio Šnajder [22].

Korištene su tri mjere za evaluaciju indeksiranja: preciznost, odaziv i F1. Definiramo ih na sljedeći način:

$$\text{preciznost} = \frac{\text{broj točnih pojmova koje je odabrao program}}{\text{ukupni broj pojmova koje je odabrao program}},$$

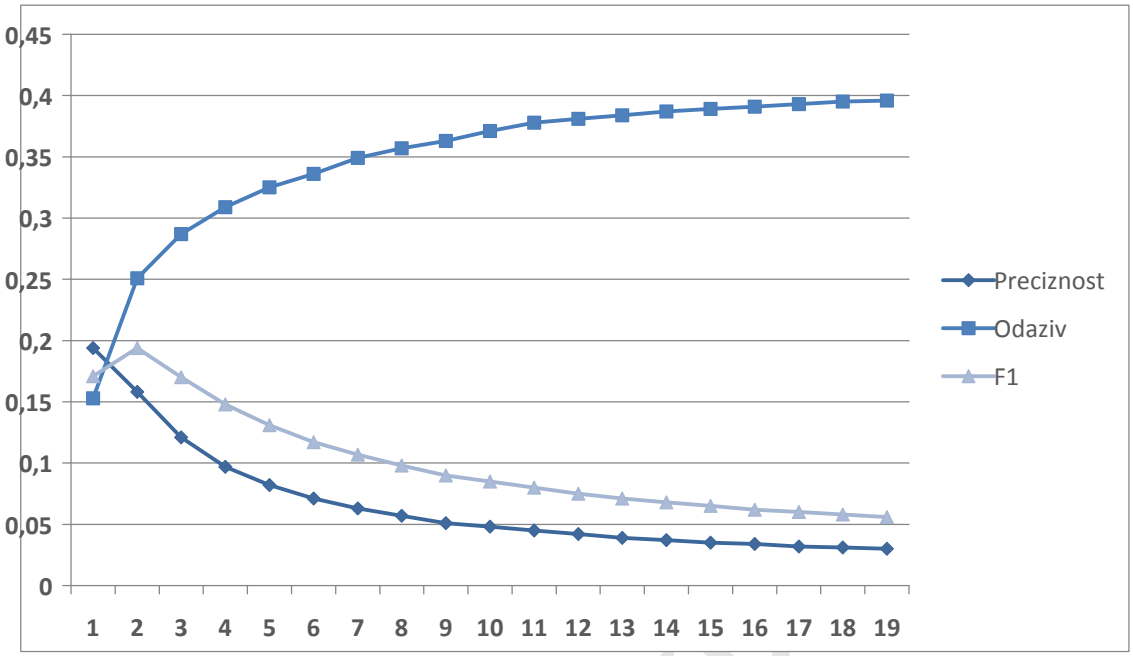
$$\text{odaziv} = \frac{\text{broj točnih pojmova koje je odabrao program}}{\text{ukupni broj pojmova koje su odabrali indeksatori}},$$

a F1 je mjera koja objedinjuje preciznost i odaziv računajući njihovu harmonijsku sredinu:

$$\text{F1} = \frac{2}{\frac{1}{\text{preciznost}} + \frac{1}{\text{odaziv}}}.$$

6.2.1. Osnovna metoda

Za osnovnu metodu skup nije bilo potrebno dijeliti u skup za učenje i skup za testiranje jer ta metoda ne koristi već klasificirane primjere. Ovisno o k, broju odabranih pojmova po dokumentu dobiveni su rezultati za osnovnu metodu, prikazani na slici 8.



Slika 8. Rezultati osnovne metode ovisno o broju izabranih pojmova po dokumentu.

Za odabrani $k = 2$ postiže se $F1 = 0.194$. Ako dodatno odbacimo sve pojmove s TFIDF vrijednošću manjom od 0.3, dobijemo $F1 = 0.218$. Bez TFIDF mjere najbolji rezultat je $F1 = 0.143$.

6.2.2. Korištenje matrice sličnosti

Skup od 4556 dokumenata podijeljen je u omjeru na način da je 70% slučajno odabranih dokumenata pripalo skupu za učenje, a 30% skupu za testiranje. Za svaki dokument nađeno je k njemu najsličnijih dokumenata. Taj dokument indeksiran je svim pojmovima kojima je indeksirano preko 50% od k dokumenata. Sličnost dva dokumenta izračunata je metodom opisanom u poglavlju 5.1.

Najbolji rezultat dobio se korištenjem $k = 2$ i iznosi $F1 = 0.03$. Tako loši rezultati bili su izvan svakog očekivanja, pa je valjalo pronaći uzrok za to. Numerički podatci su više puta provjereni i isprobani na znatno manjem

primjeru. Glavni razlog za loše performanse naĐen je u činjenici da operacija računanja matrice sličnosti ne daje intuitivne (očekivane) rezultate.

Ime metode impliciralo bi da se tom operacijom stvarno računa sličnost meĐu vrhovima dva grafa. Jedno od očekivanja takvog razmišljanja bilo je da će nakon računanja sličnosti grafa sa samim sobom svakom vrhu najsličniji vrh biti upravo on sam. Iako se u Blondelovom radu [3] dokazuje tvrdnja da je najveći element u matrici sličnosti kod računanja samosličnosti upravo element na dijagonali, to ne znači da su svi elementi na dijagonali najveći u svom retku (ili stupcu).

Ovakav rezultat mogao se predvidjeti malo boljim razmatranjem jednadžbe 5.4. Primijetimo da za ovako konstruiran graf cijela matrica sličnosti ovisi o dominantnim svojstvenim vektorima BB^T i B^TB. Drugim riječima, cijela matrica sličnosti ovisi o $M+N$ realnih vrijednosti (M i N su brojevi dokumenata, odnosno pojmova u pojmovniku), a očekivali smo red veličine $M^2/2$ korisnih informacija iz te matrice. Jasno je da ovako definirana operacija sličnosti meĐu vrhovima grafova ne daje očekivane rezultate. U praktičnoj primjeni takva operacija jedino je bila korištena kod metode čvorišta i autoriteta (iz koje je i generalizirana), te kod metode nalaženja sinonima.

7. Zaključak

U radu je proučena metoda nalaženja sličnosti među vrhovima dva grafa koja je nastala generalizacijom metode čvorišta i autoriteta, poznate i učinkovite metode za rangiranje web-stranica. Učinkovitost metoda koje koriste analizu grafova ovisi o dobro izgrađenoj strukturi grafa, u kojoj su veze među vrhovima najveći nositelj informacija.

Metoda nalaženja sinonima opisana u Blondelovom radu [3] iskorištava specifičnu strukturu rječnika, tj. gradi veze među riječima u rječniku na način koji podsjeća na Kleinbergovu [13] metodu rangiranja web stranica. Blondel je također uspješno modificirao Kleinbergovu metodu tako da daje bolje rezultate pri rangiranju sinonima. Ta metoda implementirana je u ovom radu i evaluirana na primjeru engleskog rječnika i na primjeru hrvatskog rječnika. Zanemarujući neke nedostatke, Blondelova modifikacija Kleinbergove metode pokazala se kao dobrom metodom nalaženja sinonima.

U radu je operacija traženja sličnosti među čvorovima istog grafa iskorištena za nalaženje sličnosti među dokumentima indeksiranim pojmovima pojmovnika EUROVOC. Izrađena je i vrlo jednostavna osnovna metoda kako bi se bolje ocijenila učinkovitost metode temeljene na nalaženju matrice samosličnosti jednog grafa. Primjena nove metode na navedenom skupu dokumenata nije pokazala zadovoljavajuće rezultate. Ova metoda nalaženja sličnosti među grafovima morala bi se primijeniti na drukčije izgrađene grafove. Način kako bi se to postiglo radi bolje klasifikacije tekstualnih dokumenata ostaje otvoren problem.

8. Literatura

[1] Agirre E; Martinez D; Lopez de Lacalle, O; Soroa A: *Two graph-based algorithms for state-of-the-art WSD*, Proc. of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 585–593, Sydney, 2006.

[2] Anić, V: *Veliki rječnik hrvatskoga jezika*, Novi Liber, Zagreb, 2004.

[3] Blondel, V. D; Gajardo, A; Heymans, M; Senellart, P. P; Van Dooren, P: *A measure of similarity between graph vertices: Applications to synonym extraction and web searching*, SIAM Review, Vol. 46, No. 4. pp. 647-666, 2004.

[4] Blondel, V. D; Senellart, P. P: *Automatic extraction of synonyms in a dictionary*, Technical Report 89, Université catholique de Louvain, Louvainla-neuve, Belgium, 2001.

[5] Bratanić, M., ur: *Pojmovnik EUROVOC, drugo izdanje*, HIDRA, Zagreb, 2000.

[6] Brin, S; Page, L: *The anatomy of a large-scale hypertextual Web search engine*, Proc. of the 7th International World Wide Web Conference, 1998.

[7] Curran, J: *Ensemble Methods for Automatic Thesaurus Extraction*, In Proc. of the First International Workshop on Multiple Classier Systems, pp. 1-15., 2000.

[8] Dasarathy, B. V., ur: *Nearest Neighbor (NN) Norms: NN Pattern Classification Technique*, IEEE Computer Society Press, Los Alamitos, 1991.

[9] Erkan, G; Radev, D. R: *Graph-based lexical centrality as salience in text summarization*, Journal of Artificial Intelligence Research, Vol 22, pp. 457-479, 2004.

[10] Ivanšić, I: *Numerička matematika*, Element, Zagreb, 2002.

[11] Jannink, J; Wiederhold G; *Thesaurus entry extraction from an online dictionary*, Proc. of Fusion 1999. Sunnyvale CA, 1999.

[12] Joachims, T: *A probabilistic analysis of the Rocchio algorithm TFIDF for text categorization*, Proceedings of the International Conference on Machine Learning, Nashville, 1997.

[13] Kleinberg, J. M: *Authoritative sources in a hyperlinked environment*, Journal of the ACM, 46:5, pp. 604-632, 1999.

[14] Mandala, R; Tokunaga, T; Tanaka, H: *Combining Multiple Evidence from Different Type of Thesaurus for Query Expansion*, In Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, Berkeley, 1999.

[15] Miller, G: *WordNet: A lexical database for English*, Communications of the ACM 38(11), pp. 39-41, 1995.

[16] *The Online Plain Text English Dictionary*, http://msowww.anu.edu.au/~ralph/OPTED/, 2000.
[1. 6. 2006.]

[17] van der Plas, L; Tiedemann, J: *Finding synonyms using automatic word alignment and measures of distributional similarity*, Proc. of the COLING/ACL, pp. 866-873, Sydney, 2006

[18] Plaunt, C; Norgard, B: *An association-based method for automatic indexing with a controlled vocabulary*, Journal of the American Society for Information Science, Vol. 49(10), 1998.

[19] Pouliquen, B; Steinberger, R; Ignat, C: *Automatic annotation of multilingual text collections with a conceptual thesaurus*, Proc. of In: Proceedings of the Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology - Its Potential and Practicalities (EUROLAN'2003), Bucharest, 2003.

[20] Radev, D; Qi, H; Zheng, Z; Goldensohn, S; Zhang, Z; Fang, W; Prager, J: *Mining the Web for Answers to Natural Language Questions*, In the 10th International ACM Conference on Information and Knowledge Management, Atlanta, 2001.

[21] Šarić, F; Šnajder, J; Dalbelo-Bašić, B; Eklić, H: *Enhanced thesaurus terms extraction for document indexing*, Proc. of the 27th International Conference ITI, Cavtat, 2005.

[22] Šnajder, J: *Rule-based automatic acquisition of large-coverage morphological lexicons for information retrieval*, Tech. Report, MZOŠ 2003-082, ZEMRIS, FER, University of Zagreb, 2005.

[23] van der Vet, P. E; Mars, N. J. I: *Condorcet Query Engine: A Query Engine for Coordinated Index Terms*, Journal of the American Society for Information Science Vol. 50(6), pp. 485-492, 1999.

[24] Wu, H; Zhou, M: *Optimizing synonym extraction using monolingual and bilingual resources*, Proc. of the Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003), Sapporo, Japan, 2003.

9. Dodatak

9.1. Popis tablica

9.2. Popis slika