Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Matematički odsjek

Indeks sličnosti vrhova grafa i primjena u svrhu ekstrakcije sinonima

Tim 4M
Mateo Martinjak
Mateja Pejić
Mate Poljak
Mihaela Zima

Profesor:
Zlatko Drmač

Zagreb, prosinac 2021.

Prosinac 2021

Sadržaj

1 Uvod

Cilj ovog rada je prikazati mjeru sličnosti između vrhova grafa na primjeru rječnika. Ideja je napraviti program koji će iz rječnika izvući sve sinonime neke riječi. Za danu riječ želimo izdvojiti listu sinonima. Treba imati na umu da u ovom slučaju sinonimi imaju puno zajedničkih riječi u svojoj definiciji i korištene su u definiciji drugih zajedničkih riječi. Korišteni algoritam je generalizacija Kleinberg's web pretrage.

2 Osnovni pojmovi

2.1 Osnovne definicije teorije grafova

Uvodimo par definicija ključnih pojmova iz teorije grafova.

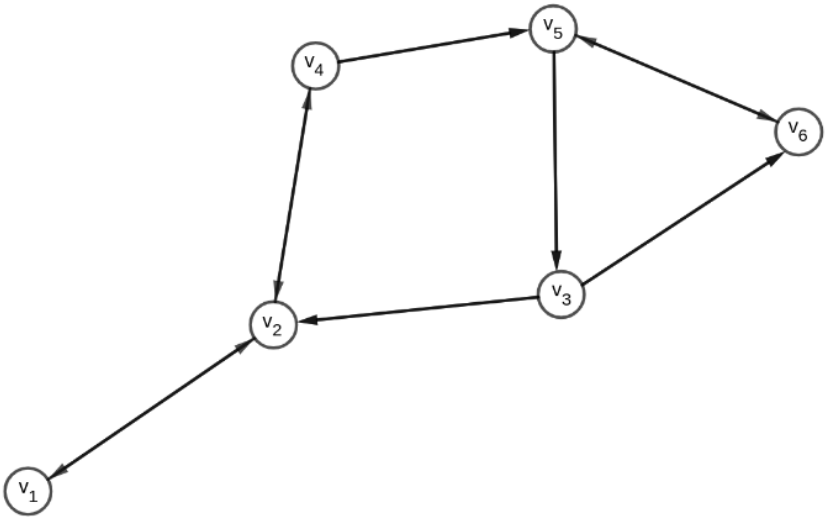
Definicija 1. Graf $G = (V, R)$ je skup V s relacijom $R \subset V \times V$.

Definicija 2. Vrhovi ili čvorovi grafa su elementi skupa V u oznaci v_i.

Definicija 3. Bridovi grafa su elementi relacije R, tj. uređeni parovi $(v_i, v_j) \in R$, u oznaci e_{ij}.

Graf koji ima usmjerene bridove zvat ćemo **usmjereni graf**.

Definicija 4. Za **vrhove** u i v kažemo da su **susjedni**, ako postoji brid (u, v) u tome grafu koji ih spaja. Za **bridove** e i f kažemo da su **susjedni**, ako postoji vrh u tome grafu koji im je zajednički.



Slika 1: Primjer usmjerenog grafa.

2.2 Matrice u teoriji grafova

Matrice su pogodne za obrađivanje različitih problema s grafovima, posebno za pohranjivanje grafova na računalima.

Matrica incidencije grafa $G = (V, R)$ je matrica $A \in M^{n\times n}$, gdje je n broj vrhova grafa. Elementi grafa su definirani sa

$$a_{ij} = \begin{cases} 0 & \text{za } (i,j) \notin R \\ 1 & \text{za } (i,j) \in R. \end{cases} \tag{1}$$

3 Hubovi i autoriteti

Internetske tražilice identificiraju iz skupa stranica relevantnih s obzirom na traženi pojam one stranice koje su dobri hub-ovi i one koji su dobri autoriteti. Dobri hub-ovi su stranice koje pokazuju na dobre autoritete i obratno.

Ukoliko identificiramo svaku web-stranicu kao vrh jednog usmjerenog grafa tada iterativnom metodom s obzirom na gornju definiciju dobrih hub-ova i autoriteta možemo svakom vrhu (web-stranici) pridružiti vrijednosti hub-ov i autoriteta.

Pretpostavimo da imamo usmjereni graf $G = (V, E)$, gdje je V skup vrhova grafa, te E skup bridova, te neka su h_j i a_j vrijednosti hubova i autoriteta vrha j redom. Vrijednosti h_j i a_j za svaki pojedini vrh inicijaliziramo nekom pozitivnom vrijednošću. Te vrijednosti se kontinuirano mijenjaju pomoću sljedećih formula:

$$h_j = \sum_{i:(j,i)\in E} a_i \tag{2}$$

$$a_j = \sum_{i:(i,j)\in E} h_i \tag{3}$$

Neka je B matrica čije su vrijednosti elementa (i, j) brojevi stranica između vrhova i i j (matrica susjedstva). Neka su h i a vektori za vrijednosti hubova i autoriteta vrhova grafa B. Tada gornje jednadžbe (1) i (2) matrično možemo zapisati kao iterativnu metodu:

$$\begin{bmatrix} h \\ a \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} h \\ a \end{bmatrix}_k, \quad k = 0, 1, \ldots \tag{4}$$

Jednostavnije gornju iterativnu metodu možemo zapisati kao

$$x_{k+1} = Mx_k, \quad k = 0, 1, \ldots \tag{5}$$

gdje su

$$x_k = \begin{bmatrix} h \\ a \end{bmatrix}_k, \quad M = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}. \tag{6}$$

S obzirom da nas zanimaju relativne vrijednosti hubova i autoriteta gornju iterativnu metodu provodimo na normaliziranom nizu vektora:

$$z_0 = x_0 > 0, \quad z_{k+1} = \frac{Mz_k}{\|Mz_k\|_2}, \quad k = 0, 1, ... \tag{7}$$

Idelano bi htjeli kao definiciju vrijednosti huba i autoriteta vrhova grafa uzeti limes niza $(z_k)_k$.

Prvi problem koji se javlja kod ovakve definicije je taj da limes gornjeg niza uopće ne mora postojati. Međutim može se pokazati da limes oscilira između sljedeća dva limesa:

$$z_{even} = \lim_{k\to\infty} z_{2k} \quad and \quad z_{odd} = \lim_{k\to\infty} z_{2k+1}$$

Drugi problem je što limesi z_{even} i z_{odd} ovise o početnoj vrijednosti z_0, te nema nekog prirodnog odabira vektora z_0. Pokaže se da je uz početni vektor $z_0 = 1$, limes z_{even} dobar odabir za rješenje našeg problema (vrijednosti huba i autoriteta vrhova u grafu). Ustvari se vektor $z_{even}(1)$ sastoji od dva podvektora od kojih jedan određuje hub vrijednosti vrhova, a drugi vrijednosti autoriteta.

4 Matrica sličnosti dva grafa

Pretpostavimo da imamo dva usmjerena grafa G_A i G_B s vrhovima n_A i n_B, te bridovima E_A i E_B. Tada matrica sličnosti X ima nenegativne elemente x_{ij} za $i = 1, ..., n_B$, te $j = 1, ..., n_A$ koji se mijenjaju pomoću sljedeće relacije:

$$x_{ij} = \sum_{r:(r,i)\in E_B, s:(s,j)\in E_A} x_{rs} + \sum_{r:(i,r)\in E_B, s:(j,s)\in E_A} x_{rs} \tag{8}$$

Jednadžba se jednostavnije može zapisati u matričnoj formi. Neka je X_k n_B x n_A matrica čije su vrijednosti k-te iteracije gornje jednadžbe x_{ij}.Tada rekurzivna jednadžba ima oblik

$$X_{k+1} = BX_kA^T + B^TX_kA, \quad k = 0, 1, ... \tag{9}$$

gdje su A i B matrice susjedstva grafova G_A i G_B. Iako limes niza $(X_k)_k$ sam ne mora konvergirati vrijedi sljedeći teorem.

Teorem 5. *Neka su G_A i G_B dva grafa s matricama susjedstva Ai B, te $Z_0 > 0$ neka inicijalna pozitivna matrica. Definirajmo,*

$$Z_{k+1} = \frac{BZ_kA^T + B^TZ_kA}{\left\|BZ_kA^T + B^TZ_kA\right\|_F}, \quad k = 0, 1, ... \tag{10}$$

Tada podniz matrica Z_{2k} i Z_{2k+1} konvergiraju k matricama Z_{even} i Z_{odd}. Štoviše, među svim matricama skupa

$$\{Z_{even}(Z_0), Z_{odd}(Z_0) : Z_0 > 0\}$$

matrica $Z_{even}(1)$ je jedinstvena matrica najveće 1-norme.

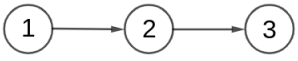
Posljedica ovog teorema je upravo ta da matrica $S = \lim_{k\to\infty} Z_{2k}$ služi kao definicija matrice sličnosti između vrhova G_A i G_B. Iz ove definicije odmah možemo uočiti da je matrica sličnosti grafova G_B i G_A upravo transponirana matrica sličnosti između grafova G_A i G_B. Pseudo-algoritam za implementaciju metode na računalu se sastoji od sljedećih koraka:

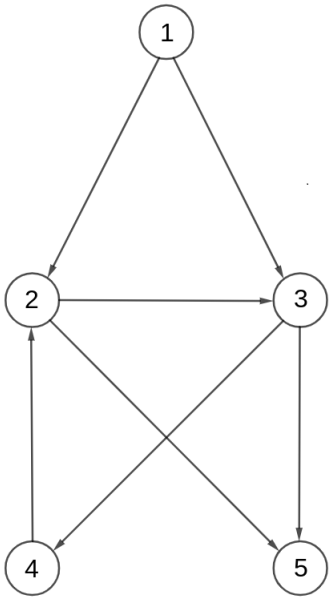
1. Inicijaliziraj $Z_0 = 1$.
2. Iteriraj paran broj puta

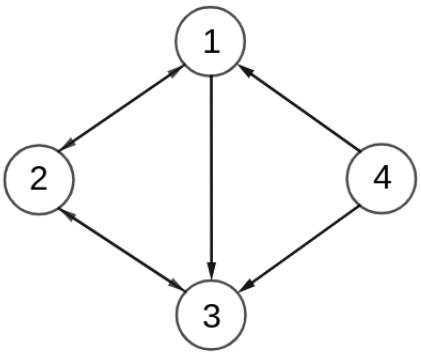
$$Z_{k+1} = \frac{BZ_kA^T + B^TZ_kA}{\left\|BZ_kA^T + B^TZ_kA\right\|_F} \tag{11}$$

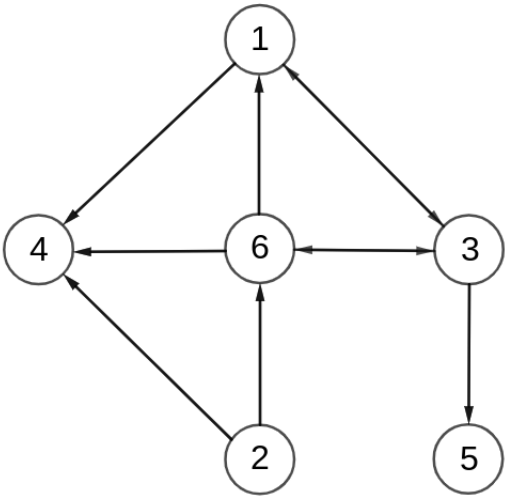
dok se ne postigne konvergencija.

3.Matrica sličnosti S je posljednja matrica Z_k iz iteracija.





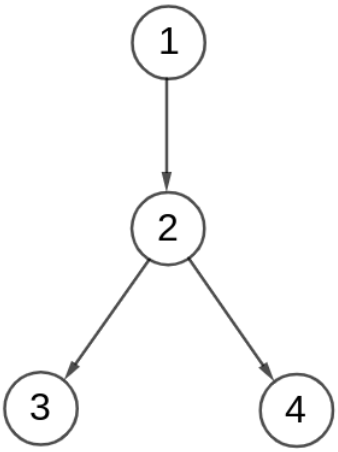




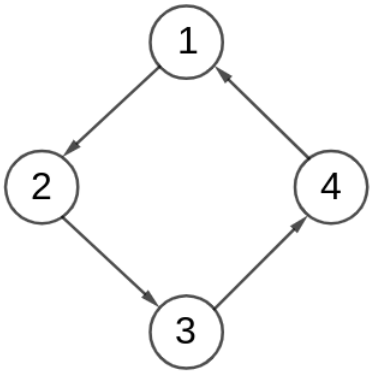
5 Matrica sličnosti grafa sa samim sobom

Poseban slučaj prilikom proučavanja sličnosti vrhova između dva grafa je taj kada je $G_A = G_B = G$. Tada je matrica sličnosti S kvadratna matrica čije su vrijednosti upravo sličnosti između pojedina dva vrha istog grafa G. Očekujemo da će vrhovi imati najveću sličnost upravo sami sa sobom, odnosno, očekujemo da će najveće relativne vrijednosti sličnosti biti upravo na dijagonali matrice S.

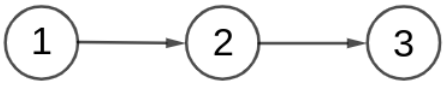
Teorem 6. *Matrica sličnosti grafa G sa samim sobom je pozitivno definitna. Najveći element matrice se pojavljuje na dijagonali.*



U primjeru cikličkog grafa očekujemo da će matrica sličnosti imati iste vrijednosti na svakom svom elementu.



Teorem 7. *Matrica sličnosti grafa puta sa samim sobom je uvijek dijagonalna.*



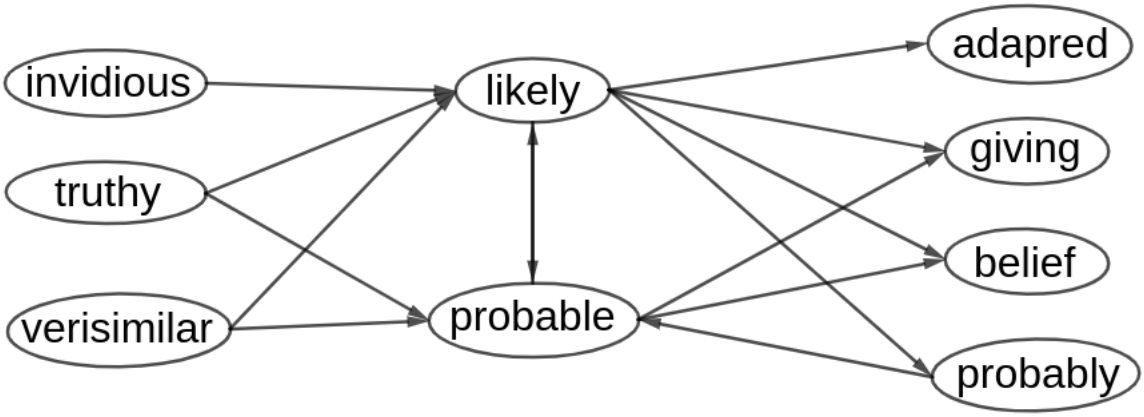
6 Primjena u svrhu ekstarkcije sinonima

Ilustrirat ćemo primjenu indeksa centralnosti za automatsku ekstrakciju sinonima iz rječnika. Detalji se mogu pronaći na [1]. Metoda se temelji na uporabi grafa koji je konstruiran iz rječnika. Pretpostavka je da sinonimi imaju puno zajedničkih riječi u svojim definicijama, te se pojavljuju zajedno u definicijama mnogih riječi.

Pomoću rječnika konstruiramo usmjereni graf G tako da svaka riječ rječnika predstavlja jedan vrh grafa. Između svaka dva vrha u i v postoji brid iz u u v ako se v pojavljuje u definiciji od u. Za izabranu riječ konstruiramo graf susjedstva G_w koji je podgraf grafa G čiji vrhovi su vrhovi grafa G koji pokazuju na w ili w pokazuje na njih. Zatim računamo indeks sličnosti vrhova grafa G_w s centralnim vrhom grafa dane strukture

$$1 \longrightarrow 2 \longrightarrow 3$$

Riječi rangiramo u padajućem poretku. Očekujemo da će riječi s najvećim centralnim indeksom biti dobri sinonimi.



Slika 2: Primjer grafa dobiven gore opisanom metodom.

Samu izabranu riječ ne uvrštavamo u skup mogućih sinonima s obzirom na to da nema previše smisla jer je za očekivati da će najbolji sinonim za svaku riječ biti upravo ona sama.

	Naša metoda	WordNet
1	make	survey
2	form	work
3	work	report
4	object	discipline
5	knowledge	subject
6	art	field
7	subject	sketch
8	mind	
9	purpose	
10	see	

Tablica 1: Predloženi sinonimi za riječ study.

	Naša metoda	WordNet	Distance
1	art	skill	art
2	branch		branch
3	law		nature
4	principle		law
5	knowledge		knowledge
6	cause		principle
7	life		life
8	practice		natural
9	nature		electricity
10	natural		biology

Tablica 2: Predloženi sinonimi za riječ science.

	Naša metoda	WordNet
1	water	sea
2	sea	
3	land	
4	large	
5	surface	
6	body	
7	more	
8	sound	
9	great	
10	river	

Tablica 3: Predloženi sinonimi za riječ ocean.

Literatura

[1] P. P. Senellart V. D. Blondel. *Automatic extraction of synonyms in a dictionary*. 2002.