SVEUČILIŠTE U ZAGREBU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET
MATEMATIČI ODSJEK

Matrične i tenzorske metode u analizi podataka

Sličnost među vrhovima grafova

Ante Ćubela, Mirna Lovrić i Ermin Mustafić
10. prosinca 2022.

Sadržaj

Uvod

Grafovi se sve češće javljaju u primjenama kao koristan alat za modeliranje u raznim područjima, npr. u prometu, mreži komunikacija, znanosti o podacima itd. Postoje razne vrste grafova, no u ovom radu su uzeti u obzir samo usmjereni. Želimo odrediti nekakvu bliskost među grafovima. Cilj ovog rada je pronaći određene sličnosti među vrhovima grafova proučavajući usmjerene bridove sa susjednim vrhovima, te primijeniti dobivene rezultate na ekstrakciju sinonima iz jednojezičnog rječnika.

1 Osnovni pojmovi

Graf G je uređeni par (V, E), gdje elemente iz V nazivamo **vrhovima**, a elemente iz $E \subseteq V \times V$ **bridovima**. Vrhovi v_i i v_j su **susjedni** ako je $(v_i, v_j) \in E$, tada je i $(v_j, v_i) \in E$. Zadnja relacija motivira uvođenje pojma **usmjerenog grafa**. U usmjerenom grafu je dozvoljeno $(v_i, v_j) \in E$ i $(v_j, v_i) \notin E$ ili obratno. Tada (v_i, v_j) nazivamo **usmjerenim bridom**. Ukoliko je E multiskup, onda iz nekog vrha može postojati i više od jednog usmjerenog brida ka nekom istom vrhu.

Matricu A koja na mjestu (i, j) ima broj usmjerenih bridova iz vrha i u vrh j nazivamo **matricom susjedstva**.
Matrica A je **simetrična** ako je $A = A^T$. $A \in \mathbb{R}^{m \times n}$ nazivamo **nenegativnom**, u oznaci $A \geq 0$, ako je $a_{ij} \geq 0$ za sve i, j.

2 Matrica sličnosti

Pogledajmo prvo jednostavniji primjer, odnosno graf G_A

$$1 \to 2 \to 3.$$

Pripadna matrica susjedstva je

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Taj graf će nam zapravo biti i najzanimljiviji jer se kod traženja sinonima računa sličnost s vrhom 2. Neka je G_B graf s matricom susjedstva B i skupom bridova E. Svakom vrhu i u G_B pridružujemo tri koeficijenta sličnosti x_{i1}, x_{i2} te x_{i3} koji predstavljaju nekakvu mjeru sličnosti vrha i s vrhovima 1, 2 i 3 iz grafa G_A, respektivno. Nakon postavljanja nekih (pozitivnih) inicijalnih vrijednosti, nove ocjene računamo sljedećim formulama:

$$x_{i1}^{novi} = \sum_{(i,j) \in E} x_{j2}^{stari}, \tag{2.1}$$

$$x_{i2}^{novi} = \sum_{(j,i) \in E} x_{j1}^{stari} + \sum_{(i,j) \in E} x_{j3}^{stari}, \tag{2.2}$$

$$x_{i3}^{novi} = \sum_{(j,i) \in E} x_{j2}^{stari}. \tag{2.3}$$

Na primjer u (2.1) vidimo da je mjera sličnosti vrha i s vrhom 1 jednaka zbroju svih ocjena x_{j2} jer 1 ima svojstvo da pokazuje na neki vrh, a na njega ni jedan drugi vrh ne pokazuje. Slično

se interpretiraju ostale ocjene. Gornje jednakosti možemo jednostavno matrično zapisati kao

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B & 0 \\ B^T & 0 & B \\ 0 & B^T & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}_k, \quad k = 0, 1, ... \tag{2.4}$$

Općenito, uzmimo dva proizvoljna grafa G_A i G_B s n_A i n_B vrhova respektivno. Tada se gornje matrične iteracije mogu zapisati u jednostavnijem obliku:

$$X_{k+1} = BX_kA^T + B^TX_kA, \quad k = 0, 1, ... \tag{2.5}$$

Promatrat ćemo normalizirane iteracije zbog smislenijeg reda veličine samog koeficijenta sličnosti kao i zbog potencijalnog *overflowa*. Također, promatrat ćemo samo parne iteracije (v. Poglavlje 3). Konačno, uz $Z_0 = X_0$, imamo

$$Z_{k+1} = \frac{BZ_kA^T + B^TZ_kA}{||BZ_kA^T + B^TZ_kA||_F}, \quad k = 0, 1, ..., \tag{2.6}$$

gdje je $||\cdot||_F$ Frobeniusova norma. Sada definiramo **matricu sličnosti** kao

$$S = \lim_{k\to\infty} Z_{2k}. \tag{2.7}$$

3 Konvergencija i algoritam

Broj $\max\{|\lambda| : \lambda \in \sigma(A)\}$ zovemo **spektralni radijus** matrice A u oznaci $\rho(A)$. Prema Perron-Frobeniusovom teoremu, spektralni radijus nenegativne matrice je svojstvena vrijednost koju nazivamo **Perronov korijen**. Nadalje, postoji pripadni svojstveni vektor koji je nenegativan.

Teorem 3.1. *Neka je M nenegativna simetrična matrica s spektralnim radijusom ρ. Tada su algebarske i geometrijske kratnosti Perronovog korijena ρ jednake. Štoviše, postoji nenegativna matrica $V \geq 0$ čiji stupci razapinju invarijantni potprostor Perronovog korijena.*

Sljedeći teorem opravdava definiciju (2.7):

Teorem 3.2. *Neka je M nenegativna simetrična matrica spektralnog radijusa ρ te $z_0 > 0$. Tada, ukoliko $-\rho$ nije svojstvena vrijednost od M,*

$$z_{k+1} = \frac{Mz_k}{||Mz_k||_2}, \quad k = 0, 1, ... \tag{3.1}$$

konvergira k $\frac{\tilde{\Pi}z_0}{||\tilde{\Pi}z_0||_2}$, gdje je $\tilde{\Pi}$ ortogonalni projektor na invarijantni potprostor Perronovog korijena ρ. Ako $-\rho$ je svojstvena vrijednost matrice M, onda imamo:

$$z_{par}(z_0) = \lim_{k\to\infty} z_{2k} = \frac{\Pi z_0}{||\Pi z_0||_2} \quad i \quad z_{nepar}(z_0) = \lim_{k\to\infty} z_{2k+1} = \frac{\Pi M z_0}{||\Pi M z_0||_2}, \tag{3.2}$$

gdje je Π *ortogonalni projektor na sumu invarijantnih potprostora od* $-\rho$ *i* ρ*. U oba slučaja je skup svih mogućih limesa dan s*

$$Z = \{z_{par}(z_0), z_{nepar}(z_0) : z_0 > 0\} = \{\frac{\Pi z}{||\Pi z||_2} : z > 0\} \tag{3.3}$$

i vektor $z_{par}(\mathbf{1})$ *je jedinstveni vektor najveće 1-norme u tom skupu.* $\mathbf{1} = (1, ..., 1)$.

Algoritam za računanje matrice sličnosti S:

1. `Stavi` $Z_0 = \mathbf{1}$.
2. `Iteriraj parno puta`
$$Z_{k+1} = \frac{BZ_kA^T + B^TZ_kA}{||BZ_kA^T + B^TZ_kA||_F}$$
`do konvergencije`.
3. S `je jednak zadnjem` Z_k.

Neka su n_A i n_B broj vrhova te e_A i e_B broj bridova grafova G_A i G_B, respektivno. Uvedimo koeficijente $\delta_A := \frac{e_A}{n_A}$ i $\delta_B := \frac{e_B}{n_B}$. Neka je ρ spektralni radijus matrice $A \otimes B + A^T \otimes B^T$ i μ druga po redu najveća svojstvena vrijednost te matrice. Tada je za točnost ε potrebno napraviti

$$8n_An_B(\delta_A + \delta_B)\frac{\log \varepsilon}{\log \mu - \log \rho} \tag{3.4}$$

iteracija.

4 Hubovi, autoriteti i centralne ocjene

Kod pretraživanja nekog upita na internetu, tražilice moraju nekako sortirati pretragu kako bi dobili što relevantnije informacije. Često za neki upit dobijemo više stotina tisuća rezultata, stoga je potrebno napraviti neki algoritam kako bi dobili kvalitetne informacije. Jedna od metoda je ocjena upita prema tome je li dobar hub ili autoritet. Internet se modelira kao usmjereni graf, gdje vrh i ima usmjeren brid na vrh j ako stranica i ima link na stranicu j.

U usmjerenom grafu G, vrh i smatramo dobrim

- **hubom** ukoliko sadrži linkove na vrhove koji u sebi sadrže neke kvalitetne informacije;
- **autoritetom** ako na njega pokazuju dobri hubovi.

Promatramo sličnost s grafom

$$hub \to authority.$$

Neka je G graf s matricom susjedstva B i skupom bridova E. Slično kao u (2.1) i (2.3), *hub-score* vrha i h_i i *authority-score* vrha i a_i računamo sljedećim formulama:

$$h_i = \sum_{(i,j)\in E} a_j, \tag{4.1}$$

$$a_i = \sum_{(j,i)\in E} h_j. \tag{4.2}$$

Kao i ranije, pripadne matrične iteracije su

$$\begin{bmatrix} h \\ a \end{bmatrix}_{k+1} = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} h \\ a \end{bmatrix}_k, \quad k = 0, 1, ... \tag{4.3}$$

Hubovi i autoriteti se računaju metodom potencija, preciznije imamo sljedeći teorem:

Teorem 4.1. *Neka je G_B graf s matricom susjedstva B. Normalizirani hub i authority scoreovi vrhova u G_B su dani kao normalizirani dominantni svojstveni vektori matrica BB^T i B^TB ako su pripadni Perronovi korijeni kratnosti 1. Inače su normalizirane projekcije vektora* **1** *na respektivne dominantne invarijantne potprostore.*

Vidimo da je gornja shema specijalan slučaj traženja sličnosti s grafom kao na početku Poglavlja 2. Za hubove trazimo sličnost s vrhom 1, a za autoritete, sličnost s vrhom 3.

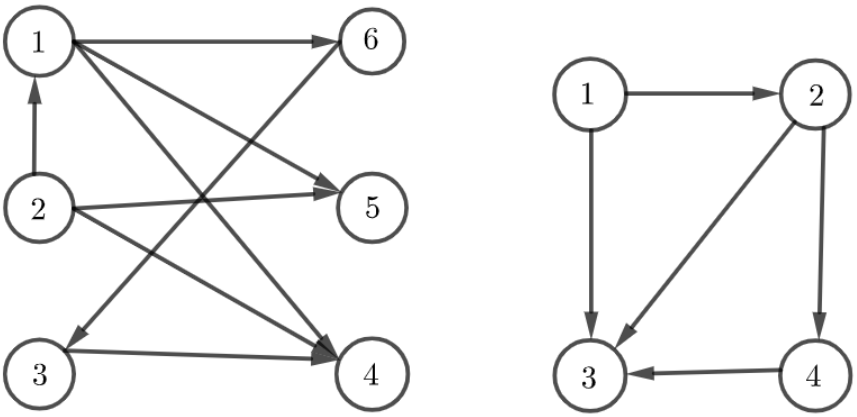
Kod traženja sinonima ćemo promatrati centralne *scoreove*, odnosno sličnost s vrhom 2. Za neku riječ w uzimamo podgraf koji se sastoji od svih riječi koji u definiciji koriste riječ w te sve riječi koje se pojavljuju u definiciji riječi w. Sada je vrh w takav da na njega pokazuju drugi vrhovi te on pokazuje na neke druge vrhove pa ima smisla tražiti sličnost s vrhom 2 jer očekujemo da sinonimi koriste slične riječi u svojim definicijama te da se pojavljuju u definicijama sličnih riječi.

Kao za hubove i autoritete, imamo analogan teorem:

Teorem 4.2. *Neka je G_B graf s matricom susjedstva B. Normalizirani centralni scoreovi vrhova u G_B su dani kao normaliziran dominantan svojstveni vektor matrice $B^TB + BB^T$ ako je pripadni Perronov korijen kratnosti 1. Inače je normalizirana projekcija vektora* **1** *na dominantan invarijantan potprostor.*

5 Primjena

Primjer 5.1. Za donja dva grafa

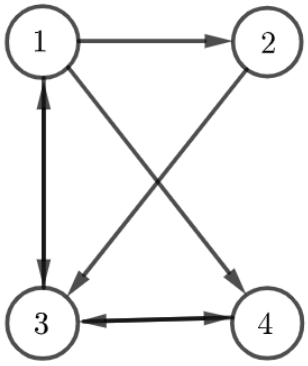


Slika 5.1 Slika prikazuje grafove G_A (lijevo) i G_B (desno).

matrica sličnosti je jednaka

$$\begin{bmatrix} 0.2708 & 0.3215 & 0.1179 & 0 & 0 & 0.0350 \\ 0.3413 & 0.3106 & 0.1281 & 0.1467 & 0.1224 & 0.0863 \\ 0.1710 & 0 & 0.0414 & 0.4240 & 0.3514 & 0.1805 \\ 0.2616 & 0.1955 & 0.1054 & 0.1611 & 0.1347 & 0.0790 \end{bmatrix}.$$

Primjer 5.2 Promotrimo sličnost grafa G sa samim sobom:



Slika 5.2. Slika prikazuje graf G.

U tom slučaju je matrica sličnosti pozitivno semidefinitna i najveći element je na dijagonali. Dodatno, ako je dijagonalni element 0, onda je cijeli pripadni redak jednak 0:

$$\begin{bmatrix} 0.3421 & 0.1693 & 0.3113 & 0.2252 \\ 0.1693 & 0.1041 & 0.1903 & 0.1479 \\ 0.3113 & 0.1903 & 0.3969 & 0.3037 \\ 0.2252 & 0.1479 & 0.3037 & 0.2476 \end{bmatrix}.$$

Primjer 5.3. Promotrimo graf G_A kao na početku poglavlja 2 te G_B kao na Slici 5.1. U

tom je slučaju matrica sličnosti jednaka

$$\begin{bmatrix} 0.3770 & 0.2309 & 0 \\ 0.3270 & 0.4122 & 0.1159 \\ 0 & 0.3387 & 0.4797 \\ 0.1700 & 0.3127 & 0.2069 \end{bmatrix}.$$

Primjetimo da je sličnost vrha 3_B s 1_A jednaka 0 jer vrh 3_B ne pokazuje ni na jedan drugi vrh, dok vrh 1_A samo pokazuje na drugi vrh.

Primjer 5.4. (Traženje sinonima)
Uz standardnu metodu, napravili smo i modifikaciju filtriranja riječi koje se pojavljuju u jako velikom broju definicija, npr. *a, the, to*, itd. Za svaku riječ w u rječniku smo izračunali broj pojavljivanja te riječi u definicijama drugih riječi, u oznaci N_w, te pripadni koeficijent sličnosti $s_{w,2}$ zamijenili s:

$$\hat{s}_{w,2} = \frac{s_{w,2}}{N_w + k}.$$

gdje smo za k uglavnom uzeli $k = \pi$ (U tablicama 5.10 - 5.13 smo varirali k). Zatim smo rangirali sinonime prema koeficijentima $\hat{s}$.

Bez modifikacije	**S modifikacijom**
pass	evanid
vanish	blinkard
light	eliminate
die	elimination
view	vanish
wear	dissipate
sail	fade
ship	fugitive
appear	fordwine
melt	efface

Tablica 5.1 Lista riječi s najvećim koeficijentom sličnosti za riječ *disappear*.

Bez modifikacije	S modifikacijom
side	parallelopiped
square	rhomb
figure	parallelogrammatic
parallel	gnomon
opposite	rhomboid
equal	longilateral
right	quadrilateral
prism	right-lined
rhomb	parallelogrammical
gnomon	prism

Tablica 5.2 Lista riječi s najvećim koeficijentom sličnosti za riječ *parallelogram*.

Bez modifikacije	S modifikacijom
more	hotpress
state	syzygy
act	octant
under	joinder
degree	anticoherer
union	continuative
word	cobelligerent
sentence	synodical
moon	corradiation
being	polysyndeton

Tablica 5.3 Lista riječi s najvećim koeficijentom sličnosti za riječ *conjuction*.

Bez modifikacije	S modifikacijom
person	zwinglian
life	machiavelian
christian	seljukian
character	antenicene
right	confessional
moral	traditionlism
make	laving
subject	parole
trust	glassite
duty	neophyte

Tablica 5.4 Lista riječi s najvećim koeficijentom sličnosti za riječ *faith*.

Bez modifikacije	S modifikacijom
part	versor
being	vector
another	factor
regard	stretch
direction	ratio
render	tense
ratio	stretching
stretch	regard
length	muscle
factor	length

Tablica 5.5 Lista riječi s najvećim koeficijentom sličnosti za riječ *tensor*.

Bez modifikacije	S modifikacijom
fleshy	berseem
pulp	cherimoyer
pulpy	stapelia
berry	carob
juice	chard
full	succulency
dry	carnous
houseleek	sappy
sedum	aloe
stonecrop	sedum

Tablica 5.6 Lista riječi s najvećim koeficijentom sličnosti za riječ *succulent*.

Bez modifikacije	S modifikacijom
large	miserere
point	fan-tan
under	stockwork
short	layshaft
little	liner
being	cresset
having	guidon
head	pedrail
light	regie
very	opeidoscope

Tablica 5.7 Lista riječi s najvećim koeficijentom sličnosti za riječ *small*.

Bez modifikacije	S modifikacijom
fear	adrad
apprehension	lisp
sneak	apprehensive
fearful	sneak
apprehensive	fearful
lisp	dare
dare	apprehension
adrad	fear
afeard	afeard
impressed	impressed

Tablica 5.8 Lista riječi s najvećim koeficijentom sličnosti za riječ *afraid*.

Bez modifikacije	S modifikacijom
large	moneyed
rich	opulent
full	wealthiness
possession	affluent
money	plutocracy
having	ample
opulent	satisfactory
state	wealthful
ample	substantial
hence	pecunious

Tablica 5.9 Lista riječi s najvećim koeficijentom sličnosti za riječ *wealthy*.

bez mod.	k=1	k=pi	k=5	k=10	k=20	k=50	k=100	k=300
wheel	trailer	trailer	trailer	trailer	trailer	snakehead	controller	coach
see	motorcycle	motorcycle	motorcycle	snakehead	snakehead	trailer	snakehead	automobile
horse	handwheel	handwheel	snakehead	motorcycle	tram	controller	tram	truck
railroad	motorcar	snakehead	trolly	tram	controller	tram	trailer	pinch
rail	motoring	trolly	handwheel	trolly	motorcycle	indicator	pinch	drag
train	double-decker	motorcar	motorcar	controller	gondola	pinch	indicator	switch
vehicle	checkstring	motoring	stateroom	dummy	trolly	gondola	automobile	controller
coach	drawrod	double-decker	motoring	handwheel	dummy	hopper	hopper	rail
wagon	mahovo	checkstring	double-decker	tramway	tramway	motorcycle	switch	freight
bear	prizing	drawrod	tram	stateroom	indicator	dummy	truck	nakehead

Tablica 5.10 Lista riječi s najvećim koeficijentom sličnosti za riječ *car*.

bez mod.	**k=1**	**k=5**	**k=10**	**k=20**	**k=50**	**k=100**	**k=300**
pass	evanid	evanid	pathos	pathos	vanish	vanish	vanish
vanish	pathos	pathos	evanid	evanid	pathos	pathos	pathos
will	fleeting	fleeting	vanish	vanish	evanid	fugitive	fugitive
like	fugitive	fugitive	fugitive	fugitive	fugitive	evanid	evanid
vapor	vanish	vanish	fleeting	fleeting	fleeting	fleeting	vapor
liable	imperceptible	imperceptible	imperceptible	imperceptible	imperceptible	vapor	fleeting
pathos	vanishing	transitory	transitory	transitory	vapor	imperceptible	pass
fugitive	transitory	vanishing	vanishing	vapor	transitory	liable	liable
notice	vapor	vapor	vapor	vanishing	liable	notice	notice
evanid	joy	liable	liable	liable	notice	transitory	away

Tablica 5.11 Lista riječi s najvećim koeficijentom sličnosti za riječ *evanescent*.

bez mod.	**k=1**	**k=pi**	**k=5**	**k=10**	**k=20**	**k=50**
substance	galactin	photosynthesis	photosynthesis	photosynthesis	photosynthesis	photosynthesis
see	skilligalee	galactin	galactin	galactin	wassail	glucose
state	osmogene	skilligalee	wassail	wassail	galactin	inversion
used	poluria	osmogene	skilligalee	andropogon	glycogen	glycogen
kind	pinole	wassail	andropogon	sulphinide	andropogon	cellulose
white	ratafia	poluria	osmogene	purl	purl	wassail
starch	kama	andropogon	sulphinide	glycogen	sulphinide	dextrose
milk	potting	pinole	poluria	skilligalee	saccharinic	dextrin
plant	muscovado	ratafia	sorbin	saccharinic	inulin	candy
taste	invertase	kama	pinole	sorbin	racemic	purl

Tablica 5.12 Lista riječi s najvećim koeficijentom sličnosti za riječ *sugar*.

k=100	**k=300**
photosynthesis	starch
glucose	glucose
inversion	photosynthesis
cellulose	inversion
candy	cellulose
dextrin	cane
dextrose	candy
glycogen	dextrin
starch	preposition
trash	dextrose

Tablica 5.13 Lista riječi s najvećim koeficijentom sličnosti za riječ *sugar*.

Literatura

[1] Z. Drmač: *MTMAP predavanja 2022./23.*. PMF Zagreb, 2022.

[2] I. Lo: *Graph Theory, The Mathematics of Networks*. Berkeley Math Circle, 2019.

[3] D. Bakić: *Linearna algebra*. PMF Zagreb, 2008.

[4] V. D. Blondel, A. Gajardo, M. Heymans, P. Senellart, P. Van Dooren: *A Measure of Similarity between Graph Vertices: Applications to Synonym Extraction and Web Searching*. SIAM, Society for Industrial and Applied Mathematics, 2004.

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