Mjera sličnosti vrhova grafa: Primjene u traženju sinonima i web tražilicama

Ivan Lazarić, Ivan Miošić, Damjan Perković

2. veljače 2020.

1 Uvod - generalizacija *hubova* i *autoriteta*

Efikasne web tražilice poput Googlea često se baziraju na otkrivanju važnih vrhova u grafu koji reprezentira veze između stranica na internetu. Jedna takva metoda, koju je predložio Kleinberg, identificira skup stranica relevantnih za traženi pojam kao podskup stranica koji su dobri *hubovi*, ili dobri *autoriteti* za taj određeni pojam. Na primjer, za pojam *"university"*, službene stranice Oxforda, Harvarda i drugih sveučilišta su dobri autoriteti, dok su stranice koje pokazuju na njih dobri hubovi - dakle, dobri hubovi su one stranice koje pokazuju na dobre autoritete, dok su dobri autoriteti stranice na koje pokazuju dobri hubovi. Iz ovih implicitnih relacija, Kleinberg je predložio iterativnu metodu koja svakoj stranici (tj. vrhu grafa) pridodaje njenu vrijednost autoriteta i vrijednost huba. Ti brojevi mogu se dobiti kao limesi konvergentnih iterativnih procesa koje ćemo sada opisati.

Neka je $G = (V, E)$ graf koji se sastoji od skupa vrhova V i skupa bridova E, te neka su a_j i h_j vrijednost autoriteta i vrijednost huba vrha j. Neka su te početne vrijednosti neki pozitivni brojevi. Sada istovremeno mijenjamo (tj. ažuriramo) sve vrijednosti autoriteta i hubova na sljedeći način: vrijednost huba vrha j postaje suma vrijednosti autoriteta svih vrhova na koje pokazuje vrh j, dok vrijednost autoriteta vrha j postaje suma vrijednosti huba svih vrhova koji pokazuju na vrh j, tj.

$$\begin{cases} h_j \leftarrow \sum_{i:(j,i)\in E} a_i \\ a_j \leftarrow \sum_{i:(i,j)\in E} h_i \end{cases}$$

Neka je B sada matrica koja na koordinatama (i, j) ima broj bridova između vrhova i i j u G, tj. matrica susjedstva grafa G. Prethodne jednakosti možemo sada zapisati kao:

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

Uvođenjem oznaka

$$x_k = \begin{bmatrix} h \\ a \end{bmatrix}_k, M = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}, \quad k = 0, 1, ...,$$

prethodnu jednakost možemo opet zapisati kao

$$x_{k+1} = Mx_k, \quad k = 0, 1, ...$$

Kako nas zanimaju samo relativne vrijednosti autoriteta i hubova (jer nas zanima koliko je koja stranica dobar autoritet, tj. hub u odnosu na druge), promatrat ćemo normaliziran niz vektora z_j, $j = 0, 1, ...$, gdje je

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

pri čemu je $\|\cdot\|_2$ Euklidska norma. Idejno, htjeli bi limes niza z_k uzeti kao vrijednosti autoriteta i hubova. No, pri tome imamo određene poteškoće - činjenicu da ovako definiran niz z_k ne mora nužno konvergirati, već često oscilira između vektora

$$z_{parni} = \lim_{k\to\infty} z_{2k} \quad \text{i} \quad z_{neparni} = \lim_{k\to\infty} z_{2k+1}$$

(pokaže se da ti limesi postoje jer je M simetrična s nenegativnim vrijednostima), te činjenicu da limesi z_{parni} i $z_{neparni}$ ovise o početnom vektoru z_0. Tada je skup svih mogućih limesa niza jednak

$$Z = \{z_{parni}(z_0),\ z_{neparni}(z_0) : z_0 > 0\},$$

te bi mi htjeli odabrati jedan vektor iz tog skupa, kao vektor vrijednosti autoriteta i hubova. Pokazuje se da za početni vektor $z_0 = \mathbf{1}$ (vektor sa svim jedinicama), limes $z_{parni}(\mathbf{1})$ ima neka zanimljiva svojstva, lagano se računa, te ima najveću 1-normu od svih vektora sadržanih u Z, pa je dobar odabir za inicijalni vektor. Zbog ovih svojstava, podvektore vektora $z_{parni}(\mathbf{1})$ uzimamo kao definicije vrijednosti autoriteta i hubova.

1.1 Generalizacija konstrukcije

Sada ćemo generalizirati ovu konstrukciju. Problem možemo alternativno promatrati na ovaj način: vrijednost autoriteta vrha j u grafu G možemo promatrati kao mjeru sličnosti između vrha j u G i vrha *autoritet* u grafu

$$hub \to autoritet,$$

te na sličan način vrijednost huba vrha j u G možemo promatrati kao mjeru sličnosti između vrha j u G i vrha *hub* ovako definiranog grafa. No, prethodnu

ideju ažuriranja vrijednosti možemo primijeniti i na grafove koji se razlikuju od *hub - autoritet* tipa grafa. Na primjer, neka je dan početni graf $G = (V, E)$, matrica susjedstva B, te neka je *graf strukture* zadan s

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

Za ovako definiran *graf strukture*, za svaki vrh j u G su asocirane tri vrijednosti: x_{j1}, x_{j2} i x_{j3} (analog tome u prethodnom *grafu strukture* su bile vrijednosti autoriteta i hubova). Na isti način kao i prije, kao početne vrijednosti uzimamo neke pozitivne brojeve. U ovom primjeru, ažuriranje vrijednosti vrši se analogno:

$$\begin{cases} x_{i1} \leftarrow \sum_{j:(i,j)\in E} x_{j2} \\ x_{i2} \leftarrow \sum_{j:(j,i)\in E} x_{j1} + \sum_{j:(i,j)\in E} x_{j3} \\ x_{i3} \leftarrow \sum_{j:(j,i)\in E} x_{j2} \end{cases}$$

Kao i prije, ove jednakosti možemo zapisati i matrično:

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

što opet označavamo kao $x_{k+1} = Mx_k$. U ovom slučaju opet vrijede ista svojstva kao i prije - M je simetrična s nenegativnim vrijednostima, nizovi normaliziranih vektora parnih i neparnih indeksa opet konvergiraju, te je $z_{parni}(\mathbf{1})$ opet vektor s najvećom 1-normom od svih mogućih limesa. Dakle, kao u prethodnom slučaju, komponente limesa $z_{parni}(\mathbf{1})$ uzimamo kao definicije vrijednosti sličnosti s_1, s_2 i s_3, te definiramo matricu sličnosti kao $\mathbf{S} = \begin{bmatrix} s_1 & s_2 & s_3 \end{bmatrix}$.

Sada možemo opisati i općeniti slučaj. Neka su dani grafovi $G_A = (V_A, E_A)$ i $G_B = (V_B, E_B)$, s n_A i n_B vrhova, respektivno. G_A nam u ovom slučaju označava *graf strukture*, što su nam u prethodnim slučajevima bili grafovi *hub* $\to$ *autoritet* i $1 \to 2 \to 3$. Ovaj put je, na analogan način, za svaki vrh i u G_B asocirano j vrijednosti x_{ij}, $j = 1, 2, ..., n_A$, $i = 1, 2, ..., n_B$. Kao i prije, ažuriramo vrijednosti na analogan način:

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

Ovu nejednakost možemo zapisati i matrično, kao i u prethodnim slučajevima. Neka je X_k matrica u kojoj su zapisane vrijednosti x_{ij} u k-toj iteraciji (dimenzija $n_B \times n_A$). Tada ažurirajuće jednakosti možemo zapisati kao

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

gdje su A i B matrice susjedstva grafova G_A i G_B, respektivno. Pokaže se da i u ovom slučaju nizovi normaliziranih vektora parnih i neparnih indeksa konvergiraju, te da je opet $Z_{parni}(\mathbf{1})$ jedinstven limes s najvećom 1-normom, te opet

definiramo matricu sličnosti kao taj limes.

Ova generalizacija ima i zanimljivu interpretaciju: ako promatramo produkt grafova G_B i G_A, tj. graf s $n_A \cdot n_B$ vrhova, koji ima brid između vrhova (i_1, j_1) i (i_2, j_2) ako postoji brid između vrhova i_1 i i_2 u G_A i brid između vrhova j_1 i j_2 u G_B, tada je gore definirana jednakost jednaka ažuriranju vrijednosti vrha produkta grafova sa sumom vrijednosti svih vrhova povezanih s tim vrhom (u bilo kojem smjeru), za svaki vrh produkta grafova.

2 Preko slicnosti do autoriteta i hubova. Sredisnji rang

2.1 Autoriteti i hubovi

Autori u radu pokazuju da matrica slicnosti izmedu zadanog grafa G_B i strukturnog grafa:

$$hub \longrightarrow authority$$

sadrzi vrijednosti rangova autoriteta i hub-ova cvorova u grafu G_B. Preciznije, neka je B matrica susjedstva grafa G_B. Tada je matrica slicnosti S s navedenim strukturnim grafom limes iteracija zadanih s

$$X_{k+1} = B^T X_k \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + B X_k \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

i $X_0 = \mathbf{1}$. Koristeci operator vec za razvoj matrice u vektor-stupac, te njegovu vezu s Kroneckerovim produktom matrica, dobijemo da vrijedi

$$\mathrm{vec}(X_{k+1}) = \left(\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \otimes B + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \otimes B^T \right) \mathrm{vec}(X_k).$$

Izraz u zagradi na desnoj strani jednak je matrici

$$M = \begin{bmatrix} 0 & B \\ B^T & 0 \end{bmatrix}$$

ciji kvadrat je $M^2 = \begin{bmatrix} BB^T & 0 \\ 0 & B^T B \end{bmatrix}$. Sada vidimo da smo dosli do matrica kojima u orignalnoj Kleinbergovoj definiciji odredujemo rangove autoriteta i hub-ova.

2.2 Sredisnji rang

Generalizacija do koje dolazimo motivirani prethodnim rezultatom je definicija sredisnjeg ranga cvorova grafa, kojim mjerimo stupanj centralnosti cvora. Promatramo matricu slicnosti zadanog grafa G_B sa sljedecim strukturnim grafom:

$$1 \longrightarrow 2 \longrightarrow 3$$

te proglasimo sredisnji stupac dobivene matrice (dakle onaj koji predstavlja slicnost cvora s cvorom 2 iz strukturnog grafa) kao *sredisnji rang* cvorova u G_B.

Analognim racunom kao u prethodnoj tocki dolazimo i do neposrednijeg rezultata: vektor sredisnjih rangova jednak je dominantnom svojstvenom vektoru matrice $B^T B + BB^T$, gdje je B matrica susjedstva grafa G_B.

U radu je na jednostavnom primjeru pokazana prednost promatranja sredisnjeg ranga nad rangovima autoriteta i hub-ova. Takoder, sredisnji rang pokazao se uspjesnim u automatskoj identifikaciji sinonima.

3 Matrice slicnosti ranga 1

Kod matrica ranga 1, sve informacije zapisane su u jednom stupcu, zbog cega su iznimno lagane za manipulaciju. Autori identificiraju dva vazna slucaja kad matrica slicnosti ima rang 1:

1. jedan od grafova je regularan;
2. jedan od grafova ima normalnu matricu susjedstva.

U drugi slucaj spadaju i neusmjereni grafovi, jer je njihova matrica susjedstva simetricna pa time i normalna.

Na osnovu zajednickih svojstava ova dva primjera, autori postavljaju sljedecu generalnu hipotezu:

HIPOTEZA. *Matrica slicnosti grafova ima rang* 1 *ako i samo ako matrica susjedstva jednog od grafova zadovoljava uvjet* $\rho(D + D^T) = 2\rho(D)$.

4 Primjena na automatiziranu ekstrakciju sinonima

Ideja sredisnjeg ranga primjenjiva je u trazenju sinonima iz zadanog rjecnika. Slijedi opis metode.

Najprije iz zadanog rjecnika konstruiramo usmjereni *graf rjecnika* G na sljedeci nacin: svaka rijec iz rjecnika postaje cvor grafa G, te postoji veza od cvora u prema cvoru v ako se rijec v pojavljuje u definiciji rijeci u. Dalje, za promatranu rijec w konstruiramo *graf susjedstva* G_w kao podgraf grafa G generiran svim cvorovima koji ili pokazuju na w ili w pokazuje na njih. Konacno, izracunamo sredisnji rang cvorova grafa G te poredamo rijeci po dobivenom rangu. Najboljih desetak rijeci proglasavamo dobrim kandidatima za sinonime polazne rijeci.

Rjecnik koji koristimo je e-verzija Websterova rjecnika. Graf kojeg dobijemo sadrzi $112,169$ cvorova i $1,398,424$ veza. Implementacija metode bit ce demonstrirana na prezentaciji seminara.