PageRank

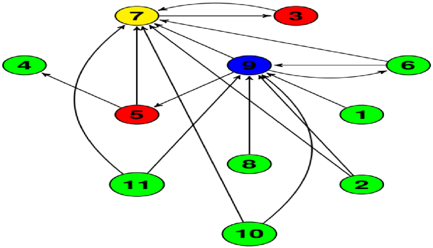
Katarina Fabek, Goran Lalić, Ivan Leverić

17. prosinca 2020.

Sadržaj

1. Uvod
2. Google matrica G i PageRank
3. Lumping
4. Algoritam za PageRank
    - Nekoliko vektora visećih vrhova
5. PageRank visećih i nevisećih vrhova
6. Web sastavljen od visećih vrhova
7. Problemi s pohranom

Uvod



$$G = \alpha S + (1 - \alpha)E \text{ for } \alpha \in [0, 1\rangle$$

Google matrica G i PageRank

• n=ukupni broj web stranica, k= broj nevisećih vrhova, $1 \leq k < n$
• $n \times n$ matrica P predstavlja strukturu weba $\Rightarrow$ dobijemo matricu H prebacivanjem nul-redaka na kraj

$$H = \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix}$$

• $n - k$ nul-retci= viseći vrhovi
• H_{11}, H_{12}

$$H_{11} \geq 0 \ , \ H_{12} \geq 0$$

$$H_{11}e + H_{12}e = e \tag{1}$$

gdje je e vektor jedinica

Google matrica G i PageRank

- Želimo: H stohastička matrica; problem: $n-k$ nul-redaka
- Rješenje: dangling node vector w

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad w \geq 0,$$

$$\|w\| = w^T e = 1 \tag{2}$$

- w_1 je $k \times 1$, w_2 $(n-k) \times 1$
- Označimo $d = \begin{bmatrix} 0 \\ e \end{bmatrix}$.

Google matrica G i PageRank

$$
\begin{aligned}
S = H + dw^T &= \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ e \end{bmatrix} \begin{bmatrix} w_1^T & w_2^T \end{bmatrix} \\
&= \begin{bmatrix} H_{11} & H_{12} \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ ew_1^T & ew_2^T \end{bmatrix} \\
&= \begin{bmatrix} H_{11} & H_{12} \\ ew_1^T & ew_2^T \end{bmatrix}
\end{aligned}
$$

- $S \geq 0$ i ((1) & (2))$\Rightarrow Se = e$

$\Rightarrow$ S stohastička!

Google matrica G i PageRank

$$G = \alpha S + (1-\alpha)E, \quad S, E \text{ stohastičke}$$

- v=personalization vector, $E = ev^T$

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \quad v \geq 0,$$
$$||v|| = 1$$

$$\Rightarrow G = \alpha S + (1-\alpha)ev^T, \quad \text{za } \alpha \in [0,1\rangle \tag{3}$$

Google matrica G i PageRank

• koristeći $\|\|_\infty$, za proizvoljnu svojstvenu vrijednost matrice G i njen pripadni svojstveni vektor z

$$|\lambda_i| \, ||z||_\infty = ||\lambda_i z||_\infty = ||Gz||_\infty \leq ||G||_\infty \, ||z||_\infty$$

$\Rightarrow |\lambda_i| \leq ||G||_\infty = 1$, dakle $\lambda_1 = \lambda = 1$ je dominantna svojstvena vrijednost

$$|\lambda_i| \leq |\lambda_1| \text{ za svaki } i = 2, ..., n$$

Iz (3), za $\lambda_1 = 1$, $\lambda_2(S)$,..., $\lambda_n(S)$ svojstvene vrijednosti od S, svojstvene vrijednosti od G su

$$1, \alpha\lambda_2(S), ..., \alpha\lambda_n(S)$$

Google matrica G i PageRank

• $\lambda = 1$ ima geometrijsku kratnost 1, uz to su sve ostale svojstvene vrijednosti ograničene s α

$$\Rightarrow G \text{ ima jedinstvenu stacionarnu distribuciju } \pi$$

$$\pi^T G = \pi^T \ , \quad \pi \geq 0,$$

$$||\pi|| = 1$$

• π = PageRank

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}$$

• π_1= PageRank nevisećih vrhova
• π_2= PageRank visećih vrhova

Lumping

- Ideja
- Primjer: Neka je P stohastička

$$\sum_{z \in t_j} p(n, z) = \sum_{z \in t_j} p(m, z)$$

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \\ \frac{5}{8} & \frac{1}{8} & \frac{1}{4} & 0 \\ 0 & \frac{1}{8} & \frac{7}{8} & 0 \\ \frac{1}{8} & 0 & \frac{5}{8} & \frac{1}{4} \end{bmatrix}$$

Lumping

- P je "skupljiva" (lumpable) u odnosu na particiju $t = \{(1,2),(3,4)\}$

$$P_t = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{8} & \frac{7}{8} \end{bmatrix}$$

Lumping

- "Lumpability" u matričnim terminima s matricom permutacija P i

$$PMP^T = \begin{bmatrix} M_{11} & \dots & M_{1,k+1} \\ \vdots & & \vdots \\ M_{k+1,1} & \dots & M_{k+1,k+1} \end{bmatrix}$$

particijom stohastičke matrice M.

- M je "lumpable" u odnosu na danu particiju ako je svaki vektor $M_{ij}e$ višekratnik vektora e.

Lumping

- Primjer:

$$M_{11} = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} \\ \frac{5}{8} & \frac{1}{8} \end{bmatrix}, \quad M_{12} = \begin{bmatrix} \frac{1}{8} & \frac{1}{8} \\ \frac{1}{4} & 0 \end{bmatrix}$$

$$M_{21} = \begin{bmatrix} 0 & \frac{1}{8} \\ \frac{1}{8} & 0 \end{bmatrix}, \quad M_{22} = \begin{bmatrix} \frac{7}{8} & 0 \\ \frac{5}{8} & \frac{1}{4} \end{bmatrix}$$

$\Rightarrow M_{11}e = \frac{3}{4}e$, $M_{12}e = \frac{1}{4}e$, $M_{21}e = \frac{1}{8}e$ i $M_{22}e = \frac{7}{8}e$

Lumping

- G je "lumpable" ako su svi viseći čvorovi spojeni u jedan čvor.

$$G = \begin{bmatrix} G_{11} & G_{12} \\ eu_1^T & eu_2^T \end{bmatrix}, \tag{4}$$

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \alpha w + (1-\alpha)v$$

- G_{11} je $k \times k$, a G_{12} $(n-k) \times k$

Lumping - Transformacija sličnosti

- Transformacija sličnosti:

$$\begin{bmatrix} G^{(1)} & * \\ 0 & 0 \end{bmatrix} \text{ gdje je } G^{(1)} = \begin{bmatrix} G_{11} & G_{12}e \\ u_1^T & u_2^T e \end{bmatrix}$$

- $G^{(1)}$ je stohastička reda $k+1$ i ima iste svojstvene vrijednosti$\neq 0$ kao G

Algoritam za PageRank

- σ=stacionarna distribucija od $G^{(1)}$

$$\sigma^T \begin{bmatrix} G_{11} & G_{12}e \\ u_1^T & u_2^T e \end{bmatrix} = \sigma^T, \quad \sigma \geq 0,$$
$$||\sigma|| = 1$$

i particija $\sigma^T = \begin{bmatrix} \sigma_{1:k}^T & \sigma_{k+1} \end{bmatrix}$, gdje je σ_{k+1} skalar

$$\Rightarrow \pi^T = \begin{bmatrix} \sigma_{1:k}^T & \sigma^T \begin{pmatrix} G_{12} \\ u_2^T \end{pmatrix} \end{bmatrix}$$
$$\Rightarrow \pi \text{ računamo preko } \sigma$$

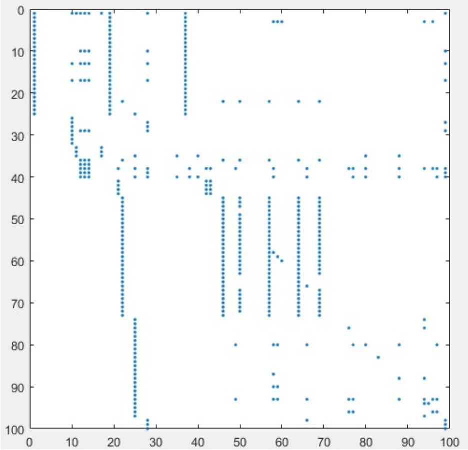
Algoritam za PageRank

Algorithm:
% Inputs: H, v, w, α; Output: $\hat{\pi}$
% We applied a Power method to $G^{(1)}$:
Choose a starting vector $\hat{\sigma}^T = \begin{bmatrix} \hat{\sigma}_{1:k}^T & \hat{\sigma}_{k+1} \end{bmatrix}$, where $\hat{\sigma} \geq 0, ||\hat{\sigma}|| = 1$

While not converged
$\hat{\sigma}_{1:k}^T = \alpha \hat{\sigma}_{1:k}^T H_{11} + (1-\alpha) v_1^T + \alpha \hat{\sigma}_{k+1} w_1^T$
$\hat{\sigma}_{k+1} = 1 - \hat{\sigma}_{1:k}^T e$
end while

% We need to recover PageRank:
$\hat{\pi}^T = \begin{bmatrix} \hat{\sigma}_{1:k}^T & \alpha \hat{\sigma}_{1:k}^T H_{12} + (1-\alpha) v_2^T + \alpha \hat{\sigma}_{k+1} w_2^T \end{bmatrix}$

Algoritam za PageRank - Primjer



Algoritam za PageRank- O konvergenciji

- stopa konvergencije metode potencija matrice $G = \alpha$

- $G^{(1)} \rightarrow$ isto, no brža je konvergencija

$\rightarrow$ manja matrica, manje operacija (ovisi o k),...

- $\Rightarrow$ metoda potencija matrice $G^{(1)}$: brži algoritam, jednostavan, minimalna pohrana

Nekoliko vektora visećih vrhova

- Ideja
- $m \geq 1$ različitih klasa visećih vrhova $\Rightarrow$ svakoj dodjeljujemo w_i,

$i = 1, \dots, m$
- Općenita Google matrica:

$$F = \begin{array}{c} \\ k \\ k_1 \\ \vdots \\ k_m \end{array} \begin{array}{c} \begin{array}{cccc} k & k_1 & \dots & k_m \end{array} \\ \begin{bmatrix} F_{11} & F_{12} & \dots & F_{1,m+1} \\ eu_{11}^T & eu_{12}^T & \dots & eu_{1,m+1}^T \\ \vdots & \vdots & & \vdots \\ eu_{m,1}^T & eu_{m,2}^T & \dots & eu_{m,m+1}^T \end{bmatrix} \end{array}$$

Nekoliko vektora visećih vrhova

- gdje je $u_i = \begin{bmatrix} u_{i,1} \\ \vdots \\ u_{i,m+1} \end{bmatrix} = \alpha w_i + (1-\alpha)v$
- za PageRank $\tilde{\pi}$ matrice F,

$$\tilde{\pi}^T F = \tilde{\pi}^T \ , \quad \tilde{\pi} \geq 0,$$

$$||\tilde{\pi}|| = 1.$$

Nekoliko vektora visećih vrhova

Theorem 2.: Neka je ρ stacionarna distribucija "lumped" matrice

$$F^{(1)} = \begin{bmatrix} F_{11} & F_{12}e & \dots & F_{1,m+1}e \\ u_{11}^T & u_{12}^T e & \dots & u_{1,m+1}^T e \\ \vdots & \vdots & & \vdots \\ u_{m,1}^T & u_{m,2}^T e & \dots & u_{m,m+1}^T e \end{bmatrix},$$

$$\rho^T F^{(1)} = \rho^T, \quad \rho \geq 0, \quad \|\rho\| = 1.$$

Nekoliko vektora visećih vrhova

S particijom $\rho^T = \begin{bmatrix}\rho_{1:k}^T & \rho_{k+1:k+m}^T\end{bmatrix}$, gdje je $\rho_{k+1:k+m}$ $m \times 1$, PageRank matrice F je

$$\tilde{\pi}^T = \left[\rho_{1:k}^T \quad \rho^T \begin{pmatrix} F_{12} & \dots & F_{1,m+1} \\ u_{12}^T & \dots & u_{1,m+1}^T \\ \vdots & & \vdots \\ u_{m,2}^T & \dots & u_{m,m+1}^T \end{pmatrix}\right].$$

PageRank visećih i nevisećih vrhova

- ideja
- $\pi_1 =$ PageRank nevisećih vrhova
- $\pi_2 =$ PageRank visećih vrhova
- utjecaj w na π_1, π_2

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \end{bmatrix}$$

$$w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}, \quad v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

- H_{11}, H_{12}

PageRank visećih i nevisećih vrhova

- iz **Algoritma**:

$$\pi_2 \nrightarrow \pi_1$$
$$\pi_1 \rightarrow \pi_2$$

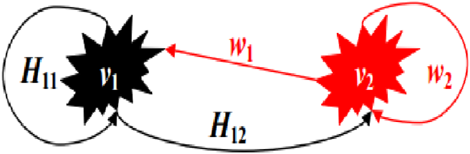
- Imamo:

$$\pi_1^T = ((1-\alpha)v_1^T + \rho w_1^T)(I - \alpha H_{11})^{-1} \tag{5}$$

$$\pi_2^T = \alpha \pi_1^T H_{12} + (1-\alpha)v_2^T + \alpha(1 - \|\pi_1\|)w_2^T \tag{6}$$

$$\rho = \alpha \frac{1-(1-\alpha)v_1^T(I-\alpha H_{11})^{-1}e}{1+\alpha w_1^T(I-\alpha H_{11})^{-1}e} \geq 0$$

PageRank visećih i nevisećih vrhova



- π_1 ne ovisi o...

PageRank visećih i nevisećih vrhova

- ovisnost preko norma (a ne individualnih elemenata):

$$\begin{aligned} \|v_1\| + \|v_2\| &= 1 \\ \|w_1\| + \|w_2\| &= 1 \\ H_{11}e + H_{12}e &= e \end{aligned}$$

- π_1 ne ovisi o π_2! (niti o broju visećih vrhova)

PageRank visećih i nevisećih vrhova

- π_1 ovisi o w_1 i v_1, distribuirani preko linkova H_{11}
- π_2 dolazi od w_2, v_2 i π_1 preko linkova H_{12}
- $\overset{(6)}{\Rightarrow}$ utjecaj w_2 na π_2 se smanjuje kako $\|\pi_1\|$ raste

PageRank visećih i nevisećih vrhova - utjecaj od w

- uzmimo normu $\|w\| = w^T e$, $w \geq 0$ u (5):

$$\pi_1^T = ((1-\alpha)v_1^T + \rho w_1^T)(I - \alpha H_{11})^{-1}$$

(e je vektor jedinica, I jedinična matrica)

- označimo $\|w\|_H = w^T (I - \alpha H_{11})^{-1} e$

PageRank visećih i nevisećih vrhova - utjecaj od w

$$\pi_1^T e = ((1-\alpha)v_1^T + \rho w_1^T)(I - \alpha H_{11})^{-1} e$$

$$\|\pi_1\| = (1-\alpha)v_1^T(I - \alpha H_{11})^{-1} e + \rho w_1^T(I - \alpha H_{11})^{-1} e$$

$$\|\pi_1\| = (1-\alpha)\|v_1\|_H + \rho\|w_1\|_H$$

$$\|\pi_1\| = (1-\alpha)\|v_1\|_H + \alpha \frac{1-(1-\alpha)v_1^T(I-\alpha H_{11})^{-1}e}{1+\alpha w_1^T(I-\alpha H_{11})^{-1}e}\|w_1\|_H$$

$$\|\pi_1\| = (1-\alpha)\|v_1\|_H + \alpha \frac{1-(1-\alpha)\|v_1\|_H)^{-1}e}{1+\alpha\|w_1\|_H}\|w_1\|_H$$

$$\|\pi_1\| = \frac{(1-\alpha)\|v_1\|_H(1+\alpha\|w_1\|_H) + \alpha\|w_1\|_H - \alpha\|w_1\|_H(1-\alpha)\|v_1\|_H}{1+\alpha\|w_1\|_H}$$

$$\Rightarrow \boxed{\|\pi_1\| = \frac{(1-\alpha)\|v_1\|_H + \alpha\|w_1\|_H}{1+\alpha\|w_1\|_H}}$$

PageRank visećih i nevisećih vrhova - utjecaj od w

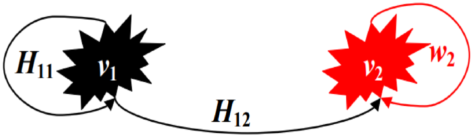
- kombinirani PageRank $\|\pi_1\|$ nevisećih vrhova je rastuća funkcija od $\|w_1\|$ jer

$$(1-\alpha)\|w_1\| \leq \|w_1\|_{\leq} \frac{1}{1-\alpha}\|w_1\|$$

- $w \geq 0 \Rightarrow \|\pi_1\|$ minimalan za $w_1 = 0$

PageRank visećih i nevisećih vrhova - utjecaj od w

- za $w_1 = 0$:

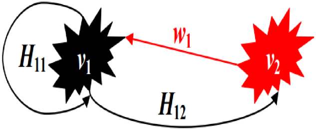


- neviseći vrhovi primaju PageRank π_1 od...
- viseći vrhovi primaju PageRank π_2 od...

$\Rightarrow$ w_2 ima jači utjecaj π_2

PageRank visećih i nevisećih vrhova - utjecaj od w

- za $w_2 = 0$:



- neviseći vrhovi primaju PageRank π_1 od...
- viseći vrhovi primaju PageRank π_2 od...

PageRank visećih i nevisećih vrhova - utjecaj od w

- iz (6) imamo:

$$\pi_2^T = \alpha \pi_1^T H_{12} + (1-\alpha) v_2^T$$

- π_1 utječe na izračun π_2, no ima samo pozitivni utjecaj jer nema dijela sa 1-$\|\pi_1\|$

PageRank visećih i nevisećih vrhova - utjecaj od w

- za $w = v$:

$\Rightarrow \pi$ je višekratnik od $v^T(I - \alpha H)^{-1}$:

$$\pi^T = \frac{1-\alpha}{1 - \alpha v^T (I - \alpha H)^{-1} d} v^T (I - \alpha H)^{-1}$$

gdje je $d = \begin{bmatrix} 0 \\ e \end{bmatrix}$

Web sastavljen od visećih vrhova

- $n=$ broj web stranica, odnosno visećih vrhova
- $n \times n$ matrica H predstavlja strukturu weba

- Želimo: H stohastička matrica; problem: n nul-redaka $=$ visećih vrhova
- Rješenje: dangling node vector w

Web sastavljen od visećih vrhova

$\Rightarrow S = ew^T$, S je stohastička i ranga 1

$\Rightarrow G$ je ranga 1:

$$G = \alpha ew^T + (1-\alpha)ev^T, \quad \text{za } \alpha \in [0,1\rangle$$

- v=personalization vector

Web sastavljen od visećih vrhova

- G zapišemo kao:

$$G = eu^T$$

gdje je

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \alpha w + (1 - \alpha) v.$$

Web sastavljen od visećih vrhova

- $e_j = j$-ti stupac matrice I
- neka je $u_j \neq 0$ ne-nul element od u

$\Rightarrow 1 = u^T e \neq 0$, dakle $G = eu^T$ je dijagonalizabilna:

$$X^{-1} G X = e_j e_j^T$$

gdje je $X^{-1} = I - e_j e_j^T - eu^T + 2e_j u^T$.

Web sastavljen od visećih vrhova

- množenje zdesna sa X^{-1} i slijeva sa e_j^T
- koristimo $XX^{-1} = I$ i $e_j^T e_j = I$:

$$e_j^T X^{-1} G = e_j^T X^{-1}$$

(znamo: $\pi^T G = \pi^T, \quad \pi \geq 0, \quad \|\pi\| = 1$)

$$\Rightarrow \pi^T = e_j^T X^{-1} = u^T$$

Problemi s pohranom

- problem pohrane matrice P u memoriji
- dekompozicija

$$P = D^{-1}A \tag{7}$$

$\Rightarrow$ štedi se prostor za pohranu i radimo s manjim brojem operacija u metodi potencija prilikom izračuna $x^T P$

Problemi s pohranom

- $nnz(P)$ = broj elemenata u P različitih od 0
- bez dekompozicije iz (7): metoda potencija zahtjeva $nnz(P)$ množenja i $nnz(P)$ zbrajanja

- sa dekompozicijom iz (7):

$$x^T P = x^T D^{-1} A = (x^T) \cdot (diag(D^{-1}))A$$

- množenje po komponentama vektora x^T i $diag(D^{-1})$ $=n$ množenja
- A matrica susjedstva $\Rightarrow$ još $nnz(P)$ zbrajanja $\Rightarrow$ ušteda operacija!

Problemi s pohranom

- ideja: sprema se P ili $A \rightarrow$ popis susjednosti stupaca matrice

$\Rightarrow$ ubrzava se PageRank metoda potencija (u svakoj iteraciji k, zahtjeva se množenje $x^{(k-1)T}P$)

- stupac i sadrži informacije tko pokazuje na i

Problemi s pohranom

- $E = \frac{1}{n} ee^T$ ili $E = ev^T \to$ problem kod pohrane

$\Rightarrow$ vektor a:

$$a_i = 1 \text{ ako red } i \text{ od } P \text{ predstavlja viseći vrh}$$
$$a_i = 0 \text{ inače}$$

Imamo:

$$S = P + av^T$$
$$G = \alpha P + (\alpha a + (1 - \alpha)e)v^T$$

Problemi s pohranom - metoda potencija

- za bilo koju početnu iteraciju $x^{(0)T}$, metoda potencija primjenjena na G

$$
\begin{aligned}
x^{(k)T} = x^{(k-1)T}G &= \alpha x^{(k-1)T}S + (1-\alpha)x^{(k-1)T}ev^T \\
&= \alpha x^{(k-1)T}S + (1-\alpha)v^T \\
&= \alpha x^{(k-1)T}P + (\alpha x^{(k-1)T}a + (1-\alpha))v^T
\end{aligned}
$$

Problemi s pohranom - metoda potencija

- ne moramo spremati G ni S $\rightarrow$ radimo pomoću P
- P rijetko popunjena $\rightarrow$ $nnz(P)$ flopsa
- minimalna pohrana, sprema se samo trenutna iteracija