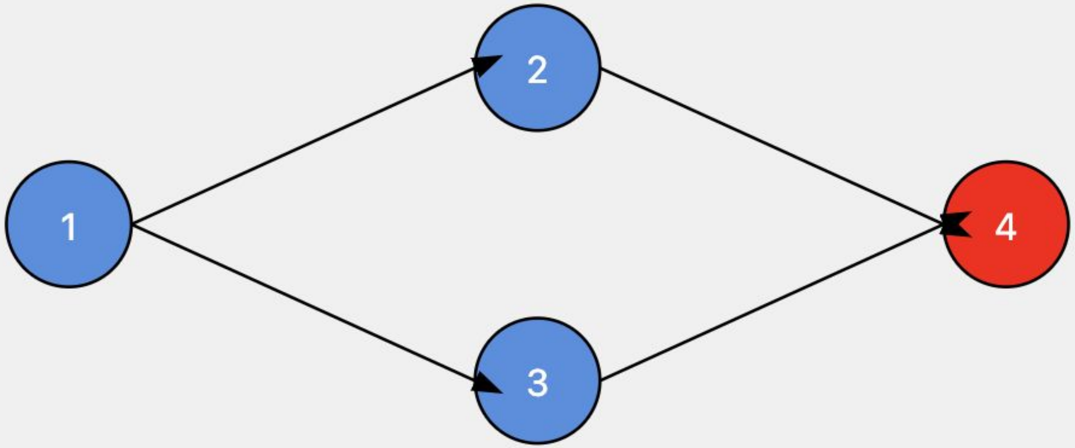
PageRank algoritam s visećim vrhovima

Marina Matešić
Tomislav Novak
Timotej Repak

2. prosinca 2024.


Google

Definicije



$$G = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

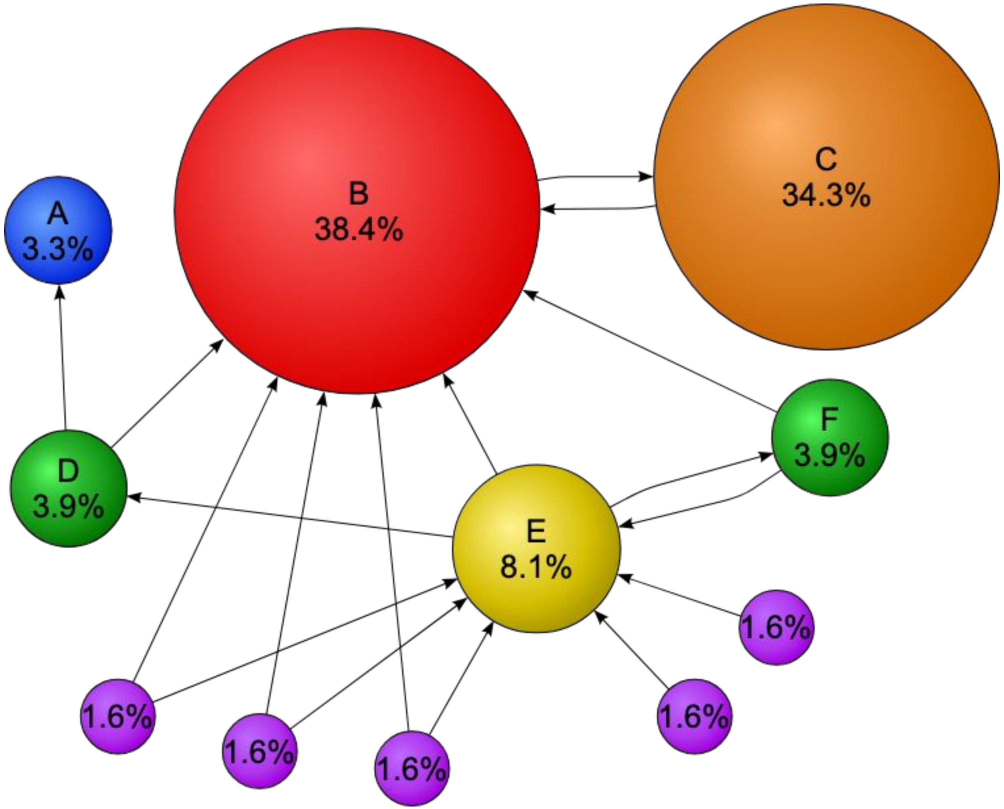
Definicije

$$x_j = \sum_{i \in L_j} \frac{1}{n_i} x_i$$

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{pmatrix}, \quad A_{ij} = \begin{cases} \frac{1}{n_i}, & \text{ako } i \to j, \\ 0, & \text{inače.} \end{cases}$$

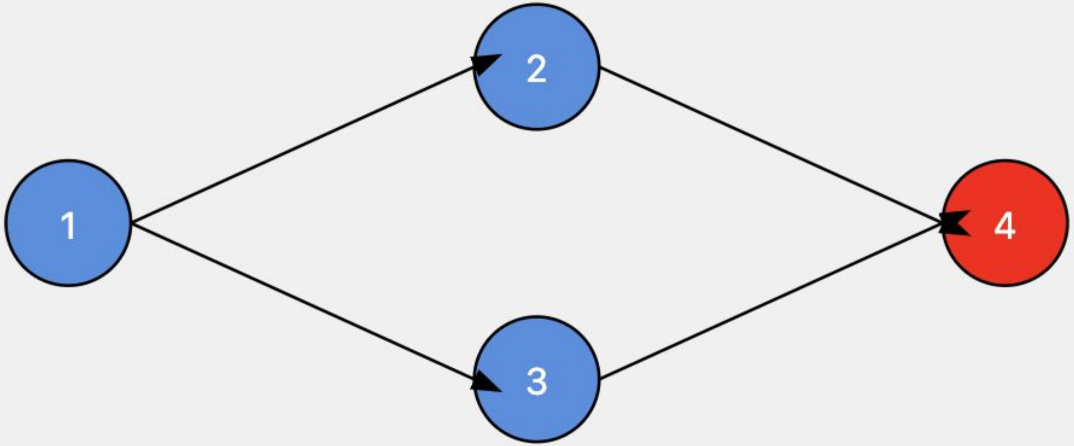
$$x = Ax, \quad tj. \quad Ax = 1 \cdot x$$

Markovljevi lanci i PageRank



Viseći vrhovi

- udio visećih vrhova je izuzetno velik u praksi



$$G = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$M_\alpha = (1-\alpha)A + \alpha S \text{ za neki } \alpha \in [0,1] \text{ gdje je } [S_{ij}] = 1/n$$

Sažimanje

Neka je, od n vrhova, k broj nevisećih ($1 \leq k < n$). Tih k vrhova fiksirajmo kao prvih k. Sada polazna $n \times n$ matrica izgleda ovako:

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

$$S := \begin{bmatrix} H_{11} & H_{12} \\ ew_1^T & ew_2^T \end{bmatrix}, \text{ gdje je } w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \text{ dimenzije } n \times 1$$

$$G := \alpha S + (1-\alpha)ev^T,\ 0 \leq \alpha < 1)$$

Konačno, matrica G izgleda ovako:

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

gdje je $u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$ konveksna kombinacija v i w.

Klase visećih vrhova

- personalizacija pretrage prema različitim temama, jezicima ili domenama, uzimanje u obzir različitih vrsta stranica (npr. tekstualne datoteke, slike videozapisi)
- svakoj klasi dodjeljujemo jedinstveni vektor

Klase visećih vrhova

$$F \equiv \begin{array}{c} \\ k \\ k_1 \\ \vdots \\ k_m \end{array} \begin{array}{c} \begin{array}{cccc} k & k_1 & \dots & k_m \end{array} \\ \begin{pmatrix} F_{11} & F_{12} & \cdots & F_{1,m+1} \\ eu_{11}^T & eu_{12}^T & \cdots & eu_{1,m+1}^T \\ \vdots & \vdots & & \vdots \\ eu_{m1}^T & eu_{m2}^T & \cdots & eu_{m,m+1}^T \end{pmatrix} \end{array},$$

$$u_i \equiv \begin{bmatrix} u_{i1} \\ \vdots \\ u_{i,m+1} \end{bmatrix} \equiv \alpha w_i + (1-\alpha)v.$$

- proces sažimanja uključuje transformacije sličnosti koje iterativno smanjuju veličinu matrice uz očuvanje njezinih stohastičkih svojstava i svojstvenih vrijednosti

Za dvije klase visećih vrhova:

- X1 sažima redove i stupce koji odgovaraju w2, dok ostavlja nepromijenjen vodeći blok veličine k + k1:

$$X_1 = \begin{bmatrix} I & 0 \\ 0 & L_1 \end{bmatrix}, \quad L_1 = I - \frac{1}{k_2}\hat{e}\hat{e}^T,$$

gdje je $\hat{e} = e - e_1$, a e_1 jedinični vektor.

Primjenom X_1 modificira se matrica:

$$F' = X_1 F X_1^{-1},$$

a postupak se ponavlja za sljedeću klasu visećih vrhova s transformacijom $\mathbf{X}_2$.

Konačna sažeta matrica ima oblik:

$$P_2X_2P_1X_1FX_1^{-1}P_1^TX_2^{-1}P_2^T = \begin{bmatrix} F_{11} & F_{12}e & F_{13}e & * \\ u_{11}^T & u_{12}^Te & u_{13}^Te & * \\ u_{21}e & u_{22}e & u_{23}e & * \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} F^{(1)} & 0 \\ 0 & 0 \end{bmatrix}$$

Algoritam

Teorem S oznakama kao u prethodnom poglavlju i G particioniranom kao u (4), neka je

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

te neka je $\sigma^T = \begin{bmatrix} \sigma_{1:k}^T & \sigma_{k+1} \end{bmatrix}$, takav da je σ_{k+1} broj. Onda je PageRank upravo

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

Algoritam

$$G^{(1)} \equiv \begin{bmatrix} G_{11} & G_{12}e \\ u_1^T & u_2^T e \end{bmatrix}$$

Ulaz: H, v, w, α
Izlaz: $\hat{\pi}$

1. Izabrati početni vektor $\hat{\sigma}^T = \begin{bmatrix} \hat{\sigma}_{1:k}^T & \hat{\sigma}_{k+1} \end{bmatrix}$ takav da je $\hat{\sigma} \geq 0$, $\|\hat{\sigma}\| = 1$.
2. Sve dok nije zadovoljen uvjet zaustavljanja:

$$\hat{\sigma}_{1:k}^T = \alpha \hat{\sigma}_{1:k}^T H_{11} + (1-\alpha) v^T + \alpha \hat{\sigma}_{k+1} w_1^T$$

$$\hat{\sigma}_{k+1} = 1 - \hat{\sigma}_{1:k}^T e$$

3. Završiti

Procjena PageRanka:

$$\hat{\pi}^T = \begin{bmatrix} \hat{\sigma}_{1:k}^T & \alpha \hat{\sigma}_{1:k}^T H_{12} + (1-\alpha) v^T + \alpha \hat{\sigma}_{k+1} w_2^T \end{bmatrix}$$

Analiza složenosti

- složenost jedne iteracije: $O(NNZ(S))$
- broj iteracija za konvergenciju ovisi o damping faktoru α i zadanoj toleranciji τ
- Za $\alpha = 0.85$ i toleranciju $\tau = 10-6$: približno $k = O(\log(\tau)/\log(\alpha))$ iteracija
- ukupna složenost algoritma s visećim vrhovima: $O(k \cdot NNZ(S))$
- sažimanjem se broj redaka i stupaca matrice prijelaza
  smanjuje s $n \times n$ na $(k + 1) \times (k + 1)$

Tehnički detalji algoritma i empirijski rezultati

- Python; moduli numpy i scipy
- KOD