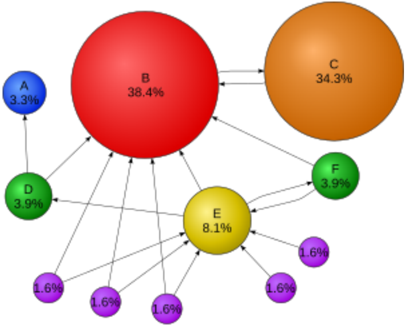
PageRank algoritam s naglaskom na viseće čvorove

Seminarski rad iz MTMAP

Doris Đivanović Karlo Gjogolović Vana Glumac

Uvod

CILJ: izračunati "*vrijednost* " kojom je definirana važnost stranice



Uvod

Problem povezivanja web-stranica moguće definirati kao:

$$a_{ij} = \begin{cases} 1 & \text{ako postoji poveznica sa stranice } i \text{ na stranicu } j, \\ 0 & \text{inače.} \end{cases}$$

Ako za neki redak i $\sum_{j=1}^{n} a_{ij} = 0$
$\Rightarrow$ na stranici i ne postoje poveznice na iduću stranicu
$\Rightarrow$ viseći čvorovi (eng. *dangling nodes*)

Google matrica

Pretpostavimo da imamo k visećih čvorova
Permutirajmo matricu A $\Rightarrow \tilde{A} = PAP^T$ + normirajmo

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

gdje su $H_{11} \geq 0$ i $H_{12} \geq 0$ redom matrice dimenzije $k \times k$ i $k \times (n-k)$ koje zadovoljavaju

$$H_{11}e + H_{12}e = e.$$

Želimo: H stohastička

Google matrica

Umjesto svakog nul-retka stavimo w^T gdje je $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$, $w \geq 0$ i $||w|| = 1$

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

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

Za $0 \leq \alpha \leq 1$ dobili smo *Google* matricu

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

Zbog dodavanja matrice ev^T ranga jedan, matrica G ima jedinstvenu stacionarnu distribuciju π koju nazivamo *PageRank*, tj. vrijedi $\pi^T G = \pi^T$

Metoda potencija

Za proizvoljnu početnu distribuciju π_0^T pri čemu vrijedi $\pi_0^T \geq 0$ i $||\pi_0^T|| = 1$ k-tu distribuciju π_k^T dobivamo iteracijom $\pi_k^T = \pi_{k-1}^T G = ... = \pi_0^T G^k$.

$$\Rightarrow \lim_{k \to \infty} \pi_k^T = \lim_{k \to \infty} \frac{\pi_{k-1}^T G}{||\pi_{k-1}^T G||} = \pi^T$$

π je traženi *PageRank* vektor.

lumping

$$G = \begin{bmatrix} G_{11} & G_{12} \\ eu_1^T & eu_2^T \end{bmatrix} \text{ pri čemu je } u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} = \alpha w + (1-\alpha)v$$

$$G_{11}e + G_{12}e = e$$

Neka je $X \equiv \begin{bmatrix} I_k & 0 \\ 0 & L \end{bmatrix}$, gdje je $L \equiv I_{n-k} - \frac{1}{n-k}\hat{e}\hat{e}^T$

lumping

$\Rightarrow XGX^{-1} = \begin{bmatrix} G^{(1)} & * \\ 0 & 0 \end{bmatrix}$, gdje je $G^{(1)} = \begin{bmatrix} G_{11} & G_{12}e \\ u_1^T & u_2^T e \end{bmatrix}$
stohastička matrica s istim ne-nul svojstvenim vrijednostima kao i G.
Ako vrijedi i $\sigma^T G^{(1)} = \sigma^T$, $\sigma \geq 0$ i $||\sigma|| = 1$ i
$\sigma^T = [\sigma_{1:k}^T \quad \sigma_{k+1}]$ pri čemu je desna strana particija od σ^T

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

.

Algoritam

Input: H, v, w, α.

1: **Inicijalizacija** σ: Odaberi početni vektor $\sigma = [\sigma_{1:k}^T \quad \sigma_{k+1}]$ td. $\sigma \geq 0$ i $\|\sigma\| = 1$.
2: **while** se ne postigne konvergencija
Ažuriraj prvi blok:

$$\sigma_{1:k}^T = \alpha\sigma_{1:k}^T H_{11} + (1-\alpha)v_1^T + \alpha\sigma_{k+1}w_1^T$$

Ažuriraj završni blok:

$$\sigma_{k+1} = 1 - \sigma_{1:k}^T e$$

endwhile
3: **Rekonstrukcija vektora** π:

$$\pi^T = \left[\sigma_{1:k}^T \quad \alpha\sigma_{1:k}^T H_{12} + (1-\alpha)v_2^T + \alpha\sigma_{k+1}w_2^T\right].$$

Output: Aproksimacija PageRank vektora π.

Algoritam

Kako vrijedi

$$G_{11} = \alpha H_{11} + (1-\alpha) e v_1^T \qquad \text{i} \qquad u_1^T = \alpha w_1^T + (1-\alpha) v_1^T$$

te

$$G_{12} = \alpha H_{12} + (1-\alpha) e v_2^T \qquad \text{i} \qquad u_2^T = \alpha w_2^T + (1-\alpha) v_2^T,$$

ekvivalentni su izrazi u algoritmu i:

$$\begin{bmatrix} \sigma_{1:k}^T & \sigma_{k+1} \end{bmatrix} = \begin{bmatrix} \sigma_{1:k}^T & \sigma_{k+1} \end{bmatrix} \begin{bmatrix} G_{11} & G_{12}e \\ u_1^T & u_2^t e \end{bmatrix},$$

tj.

$$\sigma^T = \sigma^T G^{(1)}.$$

Algoritam

Iako bi trebalo pisati

$$\sigma_{k+1} = \sigma_{1:k}^T G_{12} e + \sigma_{k+1} u_2^T e,$$

odnosno

$$\sigma_{k+1} = \sigma_{1:k}^T \left[\alpha H_{12} + (1-\alpha) e v_2^T\right] e + \sigma_{k+1} \left[\alpha w_2^T + (1-\alpha) v_2^T\right] e,$$

znamo da vrijedi

$$||\sigma|| = 1,$$

pa za element σ_{k+1} vektora σ vrijedi

$$\sigma_{k+1} = 1 - \sum_{i=1}^{k} \sigma_i = 1 - \sigma_{1:k}^T e.$$

Algoritam

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

Množenjem blokova, te uvrštavanjem izraza za G_{12} i u_2, dobivamo

$$\sigma^T \begin{pmatrix} G_{12} \\ u_2^T \end{pmatrix} = \alpha \sigma_{1:k}^T H_{12} + (1-\alpha) v_2^T + \alpha \sigma_{k+1} w_2^T.$$

PREDNOSTI ALGORITMA: jednostavna implementacija, manje računskih operacija jer je manja matrica, predvidivo ponašanje konvergencije

GENERALIZACIJA - više klasa visećih čvorova

Uvodimo međusobno različite vektore $w_1, w_2, \ldots, w_m$, pri čemu svaki vektor odgovara jednoj od m klasa.
Odredimo fiksan poredak klasa i označimo ih rednim brojevima $1, \ldots, m$.
Potrebno je rasporediti $n - k$ visećih čvorova u tih m klasa. Dakle, $k_1 + k_2 + \cdots + k_m = n - k$.

$$H \equiv \begin{bmatrix} H_{11} & H_{12} & \cdots & H_{1,i+1} & \cdots & H_{1,m+1} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{0} \end{bmatrix},$$

Definiramo

$$w_j^T = \begin{bmatrix} w_{j1}^T & w_{j2}^T & \cdots & w_{j,i+1}^T & \cdots & w_{j,m+1}^T \end{bmatrix},$$

Vektor-retkom w_j^T zamijenimo sve nul-retke u matrici H koji odgovaraju visećim čvorovima iz klase j.

GENERALIZACIJA - više klasa visećih čvorova

Google matrica je konveksna kombinacija matrica

$$F \equiv \alpha S + (1 - \alpha) e v^T, \qquad 0 \leq \alpha \leq 1,$$

Po konstrukciji je retčano stohastička matrica koja ima jedinstvenu stacionarnu distribuciju. Dakle, *PageRank* vektor π će zadovoljavati

$$\pi^T F = \pi^T, \qquad \pi \geq 0, \quad ||\pi|| = 1.$$

GENERALIZACIJA - više klasa visećih čvorova

Označimo

$$u_i = \alpha w_i + (1-\alpha)v \qquad i = 1, \ldots, m,$$

Iz toga dobivamo

$$u_{i1} = \alpha w_{i1} + (1-\alpha)v_1 \quad \text{i} \quad u_{i,j+1} = \alpha w_{i,j+1} + (1-\alpha)v_{j+1}$$

Tada *Google* matricu možemo pisati ovako:

$$F = \begin{bmatrix} F_{11} & F_{12} & \cdots & F_{1,m+1} \\ eu_{11}^T & eu_{12}^T & \cdots & eu_{1,m+1}^T \\ \vdots & \vdots & \ddots & \vdots \\ eu_{m1}^T & eu_{m2}^T & \cdots & eu_{m,m+1}^T \end{bmatrix},$$

GENERALIZACIJA - više klasa visećih čvorova

Provodimo niz transformacija počevši od posljednje klase prema prvoj. Nakon m puta, dobiva se konačna matrica:

$$\begin{bmatrix} F^{(1)} & * \\ 0 & 0 \end{bmatrix} \qquad \text{gdje je } F^{(1)} = \begin{bmatrix} F_{11} & F_{12}e & \cdots & F_{1,m+1}e \\ u_{11}^T & u_{12}^T e & \cdots & u_{1,m+1}^T e \\ \vdots & \vdots & \ddots & \vdots \\ u_{m1}^T & u_{m2}^T e & \cdots & u_{m,m+1}^T e \end{bmatrix}.$$

Matrica je stohastička, dimenzije $k+m$, te ima iste ne-nul svojstvene vrijednosti kao i prvobitna matrica F.

GENERALIZACIJA - više klasa visećih čvorova

PageRank π izračunava se pomoću stacionarne distribucije ρ matrice $F^{(1)}$ power metodom

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

Particioniranjem $\rho^T = [\rho_{1:k}^T \quad \rho_{k+1:k+m}^T]$ gdje je $\rho_{k+1:k+m}$ dimenzije $m \times 1$, *PageRank* π dobivamo kao:

$$\pi^T = \left[\rho_{1:k}^T \quad \rho^T \begin{pmatrix} F_{12} & \cdots & F_{1,m+1} \\ u_{12}^T & \cdots & u_{1,m+1}^T \\ \vdots & \ddots & \vdots \\ u_{m2}^T & \cdots & u_{m,m+1}^T \end{pmatrix} \right].$$

GENERALIZACIJA - više klasa visećih čvorova

Prethodni algoritam sada generaliziramo na m klasa.
Iz:

$$\rho^T = \rho^T F^{(1)}$$

slijedi:

- $\rho_{1:k}^T = \alpha \rho_{1:k}^T H_{11} + (1-\alpha) v_1^T + \alpha \sum_{i=1}^m \rho_{k+i} w_{i1}^T$
- $\rho_{k+i} = \rho_{1:k}^T \left[\alpha H_{1,i+1} + (1-\alpha) e v_{i+1}^T\right] e + \sum_{j=1}^m \left[\rho_{k+j}(\alpha w_{j,i+1}^T + (1-\alpha) v_{i+1}^T) e\right]$

Iz ovih $1+m$ jednadžbi možemo rekonstruirati cijeli ρ^T jer vrijedi sljedeća particija:

$$\rho^T = \begin{bmatrix} \rho_{1:k}^T & \rho_{k+1} & \dots & \rho_{k+m} \end{bmatrix}$$

GENERALIZACIJA - više klasa visećih čvorova

Teorijski rezultat i u generalnom slučaju kaže da vrijedi:

$$\pi^T = \left[\rho_{1:k}^T \quad \rho^T \begin{pmatrix} F_{12} & \cdots & F_{1,m+1} \\ u_{12}^T & \cdots & u_{1,m+1}^T \\ \vdots & \ddots & \vdots \\ u_{m2}^T & \cdots & u_{m,m+1}^T \end{pmatrix} \right].$$

Iz toga slijedi:

$$\pi^T = \left[\pi_{1:k}^T \quad \pi_{k+1:k+k_1}^T \quad \pi_{k+k_1+1:k+k_1+k_2}^T \quad \cdots \quad \right],$$

$$\pi^T_{\left(k+\sum_{j=1}^{i-1} k_j\right)+1:\left(k+\sum_{j=1}^{i-1} k_j\right)+k_i} = \alpha \rho_{1:k}^T H_{1,i+1} + (1-\alpha) v_{i+1}^T + \alpha \sum_{j=1}^{m} \rho_{k+j} w_{j,i+1}^T$$

$$\text{za svaki} \qquad i = 1, \ldots, m.$$

Korist proširenja na više klasa:

- Pretraživanje ovisno o namjenama visećeg čvora, npr. to može biti slika, videozapis ili neki drugi medijski sadržaj
- Preciznije upravljanje neželjenim poveznicama, poput spam stranica.
- Personalizaciju pretraživanja na temelju interesa korisnika.
- Pretraživanje prema jeziku ili geografskim specifičnostima.