Adatbányászat

Webbányászat
PageRank, Hubs and Authorities

Szegedi Tudományegyetem

Miért akarjuk rangsorolni a Weboldalakat?

- Mert "tudásra szomjazunk"
- Mert a Google-nak megéri. Pontosan hogy is?
- Mert állatorvost keresünk, pizzázni akarunk, meg akarjuk találni a szállásunkat, köszi esszét akarunk írni,...

Milyen a Web felépítése?

- Nagy *erősen* összefüggő, központi komponens
- Kisebb erősen összefüggő komponensek, melyek a központi EÖK-hoz be,-vagy kiélekkel csatlakoznak
- A központi EÖK-hoz beélekkel csatlakozó komponens elemei és az ahhoz kiélekkel csatlakozó komponens elemei közötti linkek
- Leszakadó komponensek

Mi kell a hatékony információvisszakereséshez?

- Ismerni kell a dokumentumok tartalmát ⇒ indexelés problémaköre
- (Esetleg többszavas) keresőkifejezésre vissza kell tudni adni, mely dokumentumokban szerepeltek
  - Szótövezés
  - Termek relevancia szerinti súlyozása, pl. *tf-idf* súlyozás (és annak számtalan variánsa)
- Keresőkifejezést tartalmazó találatok rangsorolása várható hasznosság tekintetében

Weboldalak rangsorolása

- Miben mérjük egy oldal fontosságát?
  - Látogatottság?
  - Linkek?
- Rekurzív definíciónkban egy oldal fontossága legyen arányban a rá mutató oldalak fontosságával

$$rang(j) = \sum_{i \to j} \frac{rang(i)}{\text{fokszám}(i)}$$

- Mit okoz, ha magas egy csúcs befoka? És ha a kifoka?

Sztochasztikus mátrixok

- M mátrix sor-sztochasztikus, amennyiben $\forall m_{i,j} \geq 0$, továbbá $\forall i \in \{1, \ldots, n\} \sum_{j=1}^{n} m_{i,j} = 1$
- Hasonlóan adható meg az oszlop-sztochasztikusság definíciója
- M mátrix $m_{i,j}$ elemének jelentése? $\Rightarrow$ M átmenet/tranzíciós mátrix, mely segítségével Markov folyamatok írhatók le
- Milyen jelentés társítható $p_1^\mathsf{T} = p_0^\mathsf{T} M$ és $p_1 = M p_0$ szorzatokhoz?
- És milyen $p_i^\mathsf{T} = p_{i-1}^\mathsf{T} M = p_0^\mathsf{T} M^i$-hez?

Sztochasztikus mátrixok stacionárius eloszlása

- Irreducibilitás: bármely pontból bármely másikba vezessen irányított út
- Aperiodicitás: $\forall i \exists n' : P(\sigma_n = i | \sigma_0 = i, n > n') > 0$
  $\Rightarrow \exists p^{*\mathsf{T}}$ sztochasztikus vektor, mely M stacionárius eloszlása
- Stacionárius eloszlás: $p^{*\mathsf{T}} = \lim_{t \to \infty} p_0^\mathsf{T} M^t$
  - Kicsit másképp: azon p_t^T, melyre $p_t^\mathsf{T} \approx p_t^\mathsf{T} M$
- Hatványiteráció: szorozzuk p^T-t M-el, amíg nem konvergál
- A konvergencia vonatkozhat az oldalak rangbéli változásának (valamilyen norma szerinti) összességére, de a sorrendjükre is

$$M = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

$$\begin{tikzcd} 2 \arrow[r, leftrightarrow] \arrow[d, leftrightarrow] & 4 \arrow[d] \\ 1 \arrow[r, leftrightarrow] \arrow[ur] & 3 \end{tikzcd}$$

Ergodikus Markov lánc – Hatványiteráció

$$M^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{5}{12} & \frac{5}{12} & \frac{1}{6} \\ 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{3}{4} & 0 & 0 & \frac{1}{4} \end{pmatrix}$$

$$M^3 = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{5}{8} & \frac{1}{12} & \frac{1}{12} & \frac{5}{24} \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ 0 & \frac{3}{8} & \frac{3}{8} & \frac{1}{4} \end{pmatrix}$$

p_0	p_1	p_2	p_3	...	p_6	...	p_9
0,25	0,375	0,313	0,344	...	0,332	...	0,333
0,25	0,208	0,229	0,219	...	0,224	...	0,222
0,25	0,208	0,229	0,219	...	0,224	...	0,222
0,25	0,208	0,229	0,219	...	0,224	...	0,222

Mennyire meglepő, hogy 3 pont fontossága is végig teljesen egyező?

A Web azonban nem ergodikus – Zsákutcák

- Léteznek olyan oldalak, amelyekről nem mutat kimenő link
- Az ilyen oldalak "elnyelik" a hálózat szereplőinek fontosságát
- Legegyszerűbb ilyen gráf

$$M = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ 0 & 0 \end{pmatrix} \Rightarrow M^i = \frac{1}{2^{i-1}} M \Rightarrow \forall v, \lim_{i \to \infty} v^\intercal M^i = \vec{0}$$

A zsákutcák lehetséges feloldása – példa

- "Fejtsük vissza" a gráfot, amíg zsákutcamentes nem lesz
- A zsákutca jelenséget okozó csúcsok eltávolításával sokszor újabb zsákutca csúcso(ka)t generálunk
- Határozzuk meg a zsákutcamentesített gráf csúcsainak rangját, és abból származtassuk a korábban eltávolított csúcsok rangját
- Ezzel az eljárással a rangok összege meghaladja 1,0-t (persze normalizálhatunk utólag)

$$M = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

A Web azonban nem ergodikus – Pókháló struktúrák

- "Kijártat" nélküli "csapda" komponens a hálózatban, ami magába gyűjti az összes oldal fontosságát
- Legegyszerűbb megjelenése: hurokélen túli éllel nem rendelkező csúcs (persze elképzelhetők nagyobb csapdák is)

$$M = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix}$$

p_0	p_1	p_2	p_3	...	p_6	...	p_9
0,25	0,125	0,104	0,073	...	0,029	...	0,011
0,25	0,208	0,146	0,108	...	0,042	...	0,016
0,25	0,458	0,604	0,712	...	0,888	...	0,957
0,25	0,208	0,146	0,108	...	0,042	...	0,016

A pókháló struktúrák lehetséges feloldása – példa

- A weben történő véletlen bolyongás modellezésébe vegyük bele a "teleportálás" lehetőségét
- A véletlen böngészés során $\beta(\approx 0,8 - 0,9)$ valószínűséggel kövessük az adott tartózkodási helyünkből kimenő élek egyikét
- $(1-\beta)$ valószínűséggel találjuk magunkat egy random weboldalon $\Rightarrow$ megszüntetjük a beragadást
- $p^{\mathsf{T}}_{(i+1)} = p^{\mathsf{T}}_{(i)}\beta M + (1-\beta)\frac{\vec{1}}{\text{csúcsok száma}}$
- Esetleg $p^{\mathsf{T}}_{(i+1)} = p^{\mathsf{T}}_{(i)}(\beta M + \frac{(1-\beta)}{n}\vec{1}\vec{1}^{\mathsf{T}})$, ahol n a csúcsok száma, ha az jobban tetszik

A pókháló struktúrák lehetséges feloldása – példa

$$\beta M = \begin{pmatrix} 0 & \frac{4}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{2}{5} & 0 & 0 & \frac{2}{5} \\ 0 & 0 & \frac{4}{5} & 0 \\ 0 & \frac{2}{5} & \frac{2}{5} & 0 \end{pmatrix}$$

Mekkora volt β a példában?

p_0	p_1	p_2	p_3	...	p_6	...	p_9
0,25	0,150	0,137	0,121	...	0,105	...	0,101
0,25	0,217	0,177	0,157	...	0,134	...	0,130
0,25	0,417	0,510	0,565	...	0,627	...	0,639
0,25	0,217	0,177	0,157	...	0,134	...	0,130

Perszonalizált PageRank

- Mennyire objektív a PageRank által kapott sorrendje a weboldalaknak?
- Eltérő emberek számára ugyanarra a keresőkifejezésre más számít releváns találatnak
- Minden ember számára legyen eltérő súlyozása a weboldalaknak?
- Sőt, ugyanannak az embernek is eltérő időpontokban eltérő találatok kellhetnek
- Minden ember minden lelkiállapotához legyen eltérő súlyozása a weboldalaknak, amiket tárolunk is?

Perszonalizált PageRank – "Elfogult" véletlen séták

- Általában egy topikon belül van szükségünk egy találatra (pl. jaguár a sport vagy autózás témakörében)
- Következtetni lehet, hogy mely doménre kíváncsi az ember aktuálisan
  - Keresési előzményekből
  - A keresés helyéből (pl. sportoldal keresődobozából)
  - Az ember meg is adhatja a témakört (implicit vagy explicit)
- Előnye, hogy árnyaltabb képet képes adni a hagyományos PageRank-nál

Perszonalizált PageRank – Példa

- $p^{\mathsf{T}}_{(i+1)} = p^{\mathsf{T}}_i \beta M + (1-\beta)\frac{\vec{1_r}}{|\text{adott témában releváns oldalak}|}$
- $\vec{1_r}$ vektor annyi trükköt tartalmaz, hogy csak azoknak a pontoknak megfelelő pozíciókon tartalmaz 1-eseket, amely oldalak az adott témára nézve relevánsak, más elemei 0-k

$$\beta M = \begin{pmatrix} 0 & \frac{4}{15} & \frac{4}{15} & \frac{4}{15} \\ \frac{2}{5} & 0 & 0 & \frac{2}{5} \\ \frac{4}{5} & 0 & 0 & 0 \\ 0 & \frac{2}{5} & \frac{2}{5} & 0 \end{pmatrix}$$

A PageRank meghackelése

- Linkfarmok létrehozásával az igazi relevanciával nem bíró oldalak fontosabb színben tüntethetők föl
- **TrustRank**: Perszonalizált PageRank alkalmazása, mely során a kedvezményezett "témakört" a (valamilyen kritérium alapján) megbízhatónak titulált oldalak képzik
  - A megbízható oldalak meghatározása történhet emberi erővel, de – várhatóan némileg alacsonyabb hatékonyság mellett – osztályozhatjuk őket automatikusan is

Hubs and Authorities – Gyűjtőlapok és tekintélyek

- PageRank-hez hasonló, de az oldalakat tipizálja is gyűjtőlap minőségük és tekintélyük szerint
- A rekurzivitás itt is megjelenik
- A releváns, magas tekintélyű oldalakra valószínű sok gyűjtőoldal fog mutatni
- Fordítottja is igaz

A gyűjtőlapság és tekintélyesség formalizálása

- A web A (illetve $A^\intercal$) szomszédsági mátrixára támaszkodik
  - Sztochasztikus az A mátrix?
- h és a vektorok i-dik elemei az i-dik weblap gyűjtőlap,-és tekintélyértékeit tartalmazzák
- $\mathrm{h} = \xi A\mathrm{a} : \sum_{i=1}^{n} h_i = 1$ vagy $max(\mathrm{h}) = 1$, illetve $\mathrm{a} = \nu A^\intercal \mathrm{h} : \sum_{i=1}^{n} a_i = 1$ vagy $max(\mathrm{a}) = 1$
- Kicsit másképp: $\mathrm{h} = \xi\nu AA^\intercal \mathrm{h}$, illetve $\mathrm{a} = \nu\xi A^\intercal A\mathrm{a}$
- $AA^\intercal$, illetve $A^\intercal A$ szeretnek "besűrűsödni" $\Rightarrow$ aszinkron update

HITS algoritmus aszinkron megvalósítása

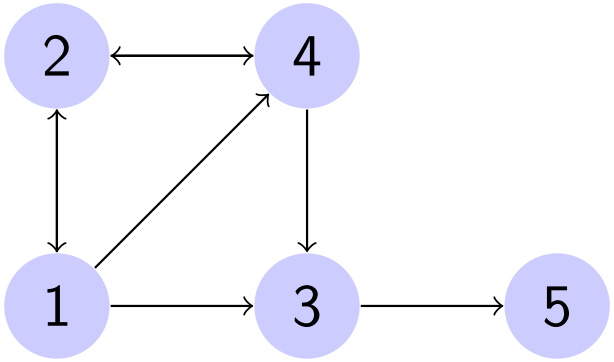
Algorithm 1 HITS algoritmus

Input: A szomszédsági mátrix

Output: a és h vektorok

```
h := 1⃗
while nem konvergál do
    a = A^T h
    a = a/max(a)
    h = Aa
    h = h/max(h)
end while
return a, h
```

HITS algoritmus – példa



$$A = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

h_0	a_1	h_1	a_2	h_2	a_3	h_3	...	a_{10}	h_{10}
1	0,5	1	0,3	1	0,24	1	...	0,21	1
1	1	0,5	1	0,41	1	0,38	...	1	0,36
1	1	0,17	1	0,03	1	0,007	...	1	0
1	1	0,67	0,9	0,69	0,84	0,71	...	0,79	0,72
1	0,5	0	0,1	0	0,02	0	...	3,5e-07	0