PageRank Algoritmus

Jana Katreniaková
katreniakova@dcs.fmph.uniba.sk

Katedra Informatiky
Fakulta Matematiky, Fyziky a Informatiky
Univerzita Komenského

1.12.2010

Motivácia

Potrebujeme ohodnotiť stránky - nejaké skóre dôležitosti

Model

Orientovaný graf: dvojica (V, E), kde

- V je množina vrcholov
- E je množina hrán: $E \subseteq V \times V$ – usporiadané dvojice vrcholov

Chceme získať nejaké ohodnotenie vrcholov

Základy z grafov

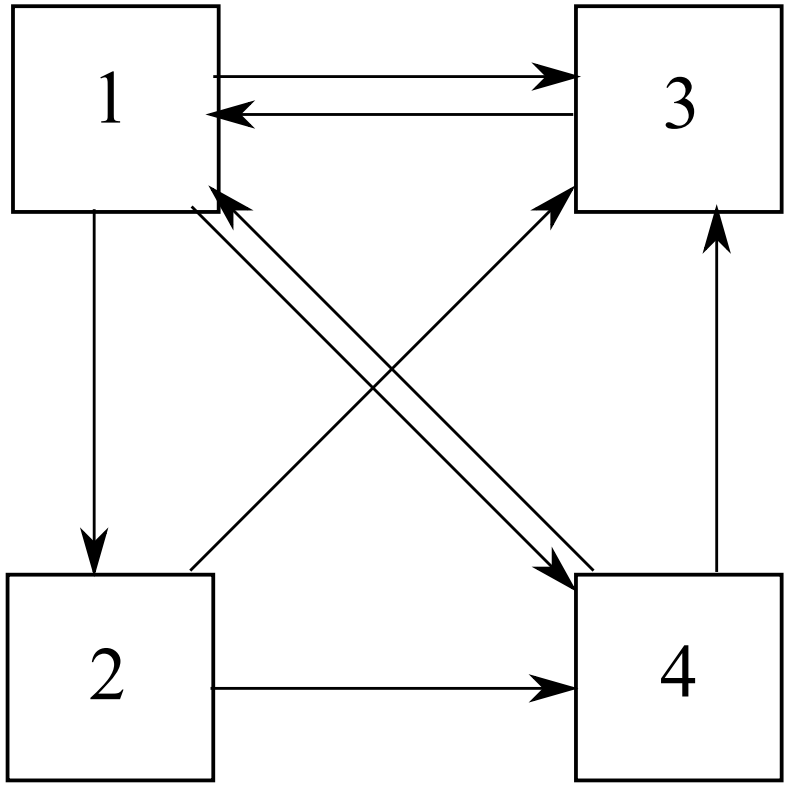
Graf možno reprezentovať viacerými spôsobmi

- Pre každý vrchol zoznam hrán
- Binárna matica veľkosti $|V| \times |V|$, kde $A_{i,j} = 1$ iff $(i, j) \in E$
- Špeciálne v prípade, že máme ohodnotené hrany – hodnota $A_{i,j}$ je
  - 0 ak $(i, j) \notin E$
  - hodnota hrany (i, j), ak $(i, j) \in E$

Pozor

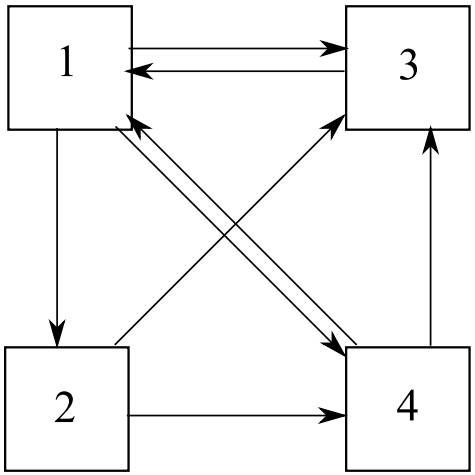
jedná sa o orientovaný graf, treba rozlišovať medzi (i, j) a (j, i)

Základy z grafov – príklad



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

Jednoduchý nápad



- Linky na nejakú stránku (tzv. backlinks) sú hlasy za túto stránku
- Zjavne pri reprezentácii maticou nám stačí spočítať čísla v riadkoch
- $$\begin{pmatrix} 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix} \quad potom\ x_1 = 2\ x_2 = 1\ x_3 = 3\ x_4 = 2$$

Jednoduchý nápad - vylepšenie

- Sú naozaj stránky 1 a 4 rovnako zaujímavé?
  - na 1 ukazuje 3 (má hodnotu 3) a 4 (má hodnotu 2)
  - na 4 ukazuje 2 (má hodnotu 1) a 1 (má hodnotu 2)
- Asi by sme chceli aby lepšia stránka mala silnejší hlas

$$
\begin{aligned}
x_1 &:= x_3 + x_4 \\
x_2 &:= x_1 \\
x_3 &:= x_1 + x_2 + x_4 \\
x_4 &:= x_1 + x_2
\end{aligned}
$$

Ešte vylepšenie

- A čo ak máme stránku, čo ukazuje na príliš veľa iných? Sú jej linky hodnotné?
- Nech má teda stránka len jeden hlas a je na nej, na koľko ho rozdelí

$$\begin{aligned} x_1 &:= x_3/1 + x_4/2 \\ x_2 &:= x_1/3 \\ x_3 &:= x_1/3 + x_2/2 + x_4/2 \\ x_4 &:= x_1/3 + x_2/2 \end{aligned}$$

Formálne:

- n_j počet linkov zo stránky j
- L_k množina stránok ukazujúcich na k
- potom $x_k = \sum_{j \in L_k} \frac{x_j}{n_j}$

A čo s tým? Veď to sa nedá

- Prvá možnosť - simulácia
- Môžeme povedať, že tie hodnoty $1/n_j$ sú hodnoty hrán, potom máme maticu

$$\begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

A teraz teoria! Opakovanie algebry.

Sú tam nejaké problémy?

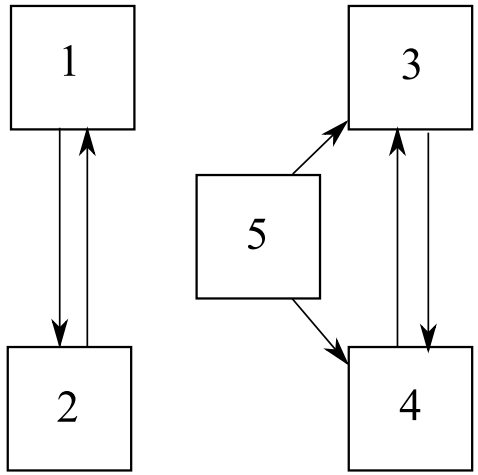
"Dangling nodes"

- dangling node = sink = vrchol s výstupným stupňom 0
- v čom je problém? Matica nie je stochastická, ale iba substochastická (v matici sú aj nulové stĺpce $\rightarrow$ nemá vlastné číslo 1)
- ale má aj tak nejaký vlastný vektor (najväčší - tzv. Perronov) $\lambda \leq 1$ a k nemu prislúchajúci Perronov vlastný vektor (vlastný vektor s nezápornými hodnotami)

Nesúvislosť grafu

- Ak web je zložený z nejakých nesúvislých podwebov (podgrafov) ako vieme "porovnať"dôležitosť častí?
- Všeobecne, ak graf nie je silne súvislý, tak máme viac vlastných vektorov, ktoré sú nezávislé

A čo teraz s tým?



$$\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Nesúvislosť grafu ešte raz

- Po vypočítaní zistíme, že sú dva bázové vektory $[1/2, 1/2, 0, 0, 0]$ a $[0, 0, 1/2, 1/2, 0]$
- Vlastným vektorom je každá ich lineárna kombinácia

Vadí nám to?

Všeobecne platí, že matica vyzerá nasledovne

$$\begin{pmatrix} A_1 & 0 & \dots & 0 \\ 0 & A_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & A_r \end{pmatrix}$$

A ak v_i je vlastný vektor matice A_i, tak potom vektory

$$w_1 \begin{pmatrix} v_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, w_2 \begin{pmatrix} 0 \\ v_2 \\ \vdots \\ 0 \end{pmatrix} \dots w_r \begin{pmatrix} 0 \\ 0 \\ \vdots \\ v_r \end{pmatrix}$$ (a ich lin. kombinácie)

sú vlastné vektory A

OK, vadí! A čo teda s tým?

Upravíme mierne vzťah: $x_k = (1-d) \cdot \frac{1}{n} + d \cdot \left(\sum_{j \in L_k} \frac{x_j}{n_j} \right)$
resp. výsledná matica bude

$$M = (1-d) \cdot S + d \cdot A$$

$$(1-d) \cdot \begin{pmatrix} .2 & 0 & 0 & 0 & 0 \\ 0 & .2 & 0 & 0 & 0 \\ 0 & 0 & .2 & 0 & 0 \\ 0 & 0 & 0 & .2 & 0 \\ 0 & 0 & 0 & 0 & .2 \end{pmatrix} + d \cdot \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

kde $d \in (0,1)$ a bežne sa používa 0.85

Máme výsledok! Maticu M

pre maticu M

- M je kladná – t.j. $M_{i,j} > 0 \forall i, j$
- M je stlpcovo stochastická
- z toho vyplýva M (teória) báza vl. vektorov je jednorozmerná a každý vlastný vektor má buď všetky členy kladné alebo všetky záporné
- M má vlastný vektor, ktorý je rozumným rankingom

Už len vypočítať TEN vlastný vektor

Použijeme teóriu

A to fakt takto počítajú aj s webom?

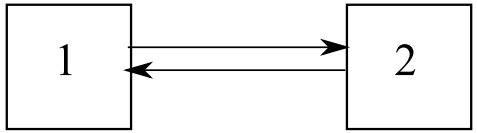
Reálne sa používa...

V podstate simulácia, ale podložená tzv. power metódou

- začneme nejakým vhodným vektorom x_0
- vytvárame postupne $x_k = M.x_{k-1} = M^k.x_0$
- resp. aby sme zaručili rozumné ohraničenie $x_k = \frac{M.x_{k-1}}{||M.x_{k-1}||}$
- môžeme použiť ľubovoľnú normu, napríklad

$$||v|| = \sum_i |v_i|$$

Tak si kúsok zasimulujeme



Začneme napríklad $PR(A_1) = 1$ a $PR(A_2) = 1$
Potom $PR(A_1) = (1 - d) + d.PR(A_2) = 0.15 + 0.85 * 1 = 1$ a
$PR(A_2) = (1 - d) + d.PR(A_2) = 0.15 + 0.85 * 1 = 1$
A keby sme začali inde? $PR(A_1) = 0$ a $PR(A_2) = 0$

$PR(A_1) = (1 - d) + d.PR(A_2) = 0.15 + 0.85 * 0 = 0.15$ a

$PR(A_2) = (1 - d) + d.PR(A_2) = 0.15 + 0.85 * 0 = 0.15$

$PR(A_i) = (1 - d) + d.PR(A_j) = 0.15 + 0.85 * 0.15 = 0.2775$

$PR(A_i) = (1 - d) + d.PR(A_j) = 0.15 + 0.85 * 0.2775 = 0.385875$ $PR(A_i) \longrightarrow 1$

A čo zhora? $PR(A_1) = 40$ a $PR(A_2) = 40$

$PR(A_1) = (1 - d) + d.PR(A_2) = 0.15 + 0.85 * 40 = 34.15$ a

$PR(A_2) = (1 - d) + d.PR(A_2) = 0.15 + 0.85 * 40 = 34.15$

$PR(A_i) = (1 - d) + d.PR(A_j) = 0.15 + 0.85 * 34.15 = 29.1775$

$PR(A_i) = (1 - d) + d.PR(A_j) = 0.15 + 0.85 * 29.1775 = 24.950875$ $PR(A_i) \longrightarrow 1$

Čo ešte dodať?

- Aj stránka na ktorú nič neukazuje má nejakú hodnotu a vie hlasovať
- S použitím váženého priemeru matíc S a A sme vyriešili problém s nesúvislosťou webu
- Dangling nodes sa dajú
  - buď odignorovať (len bude kúsok inak fungovať vlastný vektor
  - alebo odstrániť dangling links, vypočítať vl. vektor a potom zase pridať a upraviť