

Föreläsning 10

Anteckning... TMV200 Föreläsningar

Skapad: 2020-11-27 08:01

Ändrad: 2020-11-27 09:48

Författare: hchristian.johansson@gmail.com

TMV200 27/11-20

Sist: Matriser, grannmatriser för riktade grafer.

Idag: Googles PageRank-algorithm

Böjan: Larry Page & Sergey Brin
(doktorander i datavetenskap på Stanford)

Page + Winograd: Idé att Page skulle studera strukturen hos Internet som en riktad graf (andra halvår av 90'-talet)

Internet är en riktad graf!

Noder = hemsidor, ^{riktad} (kant från hemsida x till hemsida y om det finns en länk till y på x).

PageRank: En idé för att ranka hemsidors relevans.

Gamla sökmotorer (AltaVista, Yahoo)

Klart vanligt att mest relevanta sökningen nämner
dyka upp på sida 8, eller mycket längre ned, eller
"inte alls".

Idé: Nära besläktet med en idé för att avgöra
vilken forskning som är relevant.

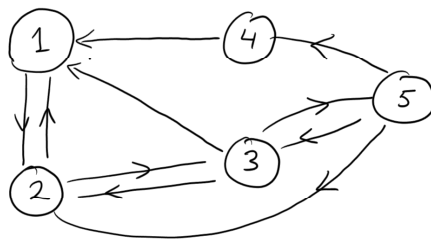
"Viktiga artiklar citeras ofta".

För internet: "Viktiga" hemsidor länkas till från
många andra hemsidor.

Hur skall vi göra i praktiken?

Modellera en "random surfer" - surfar på
webben genom att slumpmässigt klicka på länkar.

Ex



"Fem hemsidor +
+ länkar"

Senast:

Grannmatris:

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

Variant:

$$B = \begin{pmatrix} b_{11} & \dots & b_{1n} \\ \vdots & \ddots & \vdots \\ b_{n1} & \dots & b_{nn} \end{pmatrix}$$

$$\left(\frac{1}{n} \right)$$

om det finns en kant från
... ..

$$b_{ij} = \begin{cases} 1 & \text{till } j, \text{ där } v_i = \text{nummer} \\ & \text{kanter som börjar på } i. \\ 0 & \text{annars.} \end{cases}$$

För grafen ovan: $P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$

"viktad gränsmatris".

Slumpvandring på en graf: (exempel på en Markovkedja, se kursen i "Matematisk statistik och diskret matematik" nästa år).

Börjar på en hemsida x_0 (steg 0). I steg 1 väljer vi en länk från x_0 till en annan hemsida slumpmässigt, säg x_1 . Från x_1 väljer vi en länk från x_1 till en annan hemsida, säg x_2 , osv. (Här x_i är hemsidor).

|| Vad är sannolikheten att jag, efter k steg, befinner mig på hemsida y ?

Hemsida nr m kan representeras av radvektorn

$$(0 \dots 0 \underset{\substack{\uparrow \\ \text{plats } m}}{1} 0 \dots 0)$$

Ex Hemsidan nr 2 för vårt mini-internet motsvarar

$$(0 \ 1 \ 0 \ 0 \ 0)$$

Mer generellt (i exemplet): $(p_1 \ p_2 \ p_3 \ p_4 \ p_5)$

med $p_i \geq 0$ & $p_1 + p_2 + p_3 + p_4 + p_5 = 1$

- 0 1 ... 0

representerar situationen där jag är på nersidan -
med sannolikhet p_i .

Beräkning:

$$\underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \end{pmatrix}}_{\text{"hemsida 2"}} \cdot \underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}}_{P = \text{övergångsmatrix}} =$$

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

$$= \begin{pmatrix} 1/2 & 0 & 1/2 & 0 & 0 \end{pmatrix}, \text{ dvs andra raden.}$$

sannolikheterna för var jag befinner mig
om jag tar ett steg från hemsida 2.

Mer generellt: $(p_1 \ p_2 \ p_3 \ p_4 \ p_5) \cdot P =:$

$$=:(p'_1 \ p'_2 \ p'_3 \ p'_4 \ p'_5)$$

sannolikheterna för var jag befinner mig i
steg $k+1$, om sannolikheterna för var
jag befann mig efter steg k är $(p_1 \ p_2 \ p_3 \ p_4 \ p_5)$.
(oberoende av k).

Så om jag startar på hemsida m , motsvarande en
radvektor $e_m = (0 \dots 0 \underset{\substack{\uparrow \\ \text{plats } m}}{1} 0 \dots 0)$, så

ger radvektorn $e_m P^k$ sannolikheterna för var jag
befinner mig efter k steg.

Vad händer när k är "stort"?

Sannolikheterna i $e_m P^k$ "stabiliserar" (man säger
att de konvergerar, se kursen i Matematisk Analys nästa år).

Det finns en radvektor $\pi = (\pi_1 \ \pi_2 \ \pi_3 \ \pi_4 \ \pi_5)$

• skillnaden mellan $e_m P^k$ och π blir mindre och

sa att summan ...
mindre" ju större k är.

π är oberoende av vilken hemsida man starter på.

(Har egenskapen $\pi P = \pi$, kallas för en eigenvektor till P , se linjär algebra nästa termin).

Hemsida i rankas efter storleken på π_i , ju större desto bättre.

| vårt exempel är

$$\pi \approx \begin{pmatrix} 0,293 & 0,390 & 0,220 & 0,024 & 0,073 \end{pmatrix}$$

↑ ↑
näst viktigaste
viktigast hemsidan

Rankingen är 2, 1, 3, 5, 4.

Det här är nästan PageRank.

PageRank är modifikationen av den här algoritmen där surfaren med 15% procentens chans väljer en hemsida på måfå istället för att följa en länk.

| vårt exempel betyder det att man byter ut P mot

$$P' = 0,85 \cdot P + 0,15 \cdot Q, \quad \text{där alla tal i } Q \text{ är } 1/5.$$

Sen gör man samma analys.

En sökmotor består av många komponenter, och PageRank är bara en del, men den var avgörande delen för Googles framgång i början.

... måste man indexera "alla" hemsidor och deras

lexikon
innehåll (ord, fraser, länkar)

Originalkoncept (förenklat!) Gör PageRank på "hela internet".
(det går faktiskt!)

Sen när jag söker på en fras / ord
så plockar man ut de dokument där
fras/ordet förekommer och rankar enligt
PageRank.