MA1202/6202

PAGE RANK

FORELESNING E16

Vi åpner en nettleser og søker på "NTNU".

Hvordan avgjør Google hvilke nettsider den skal vise oss?

↳ Alle nettsider som nevner NTNU.

I hvilken rekkefølge vises disse søketreffene?

↳ Hva om den sida som skriver NTNU flest ganger, blir rangert øverst?

↳ Gir irrelevante treff!

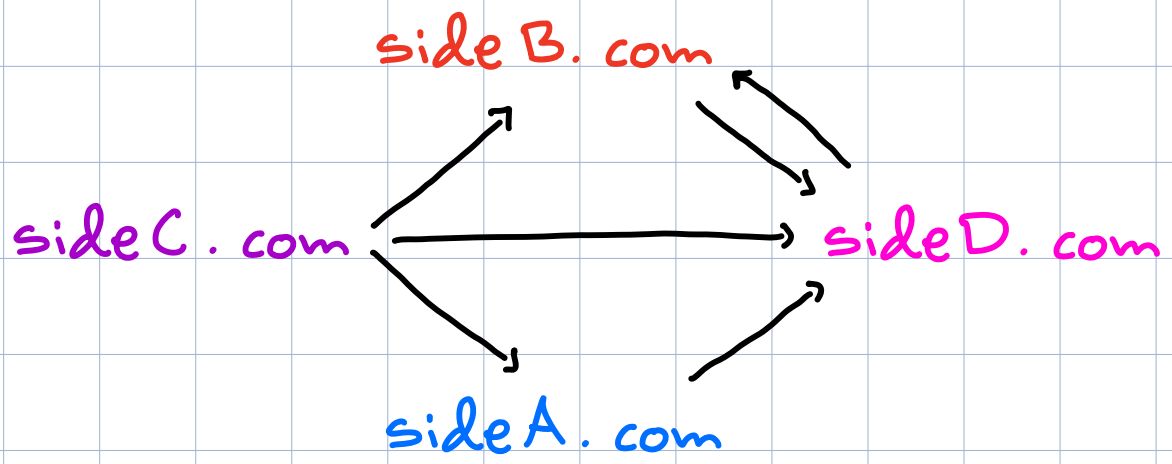
KONGSTANKEN

Viktige sider blir ofte lenket til!

EKSEMPEL

Vi åpner en nettleser og søker på "NTNU".
Søket gir fire treff, nemlig nettsidene
sideA.com, sideB.com, sideC.com og sideD.com.

Anta at lenkestrukturen ser slik ut:
(En pil x → y betyr at x inneholder en peker til y.)

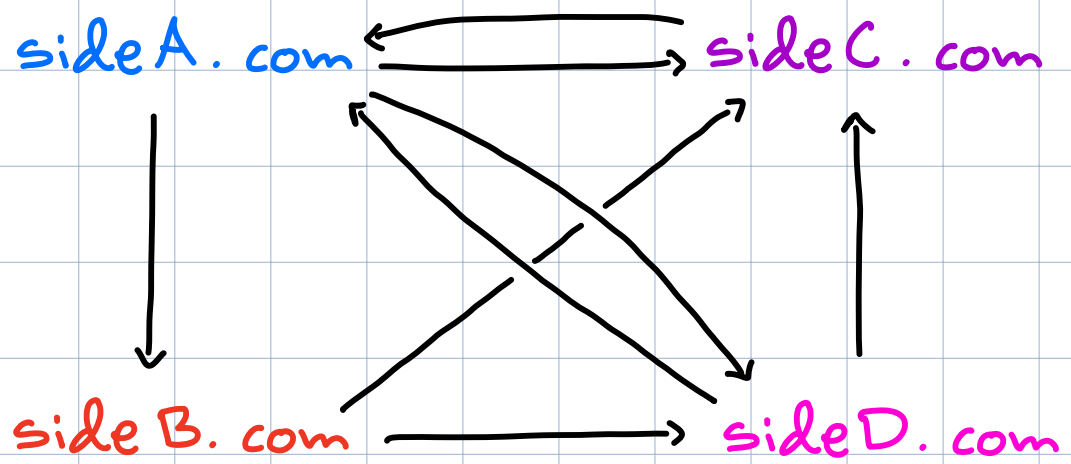


RANGERING:

1. sideD.com
2. sideB.com
3. sideA.com
4. sideC.com

EKSEMPEL

Anta at lenkestrukturen ser slik ut:



Hvordan skal vi nå rangere sidene etter "viktighet"?

RANDOM SURFER - MODELLEN

2.1.2. Intuitive justification

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a Web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank.

ALTSÅ

Hver nettside "overfører" sin viktighet jevnt til alle nettsidene den peker til.

REVIDERT KONGSTANKE

En nettside er viktig hvis mange viktige nettsider peker til den.

"DEFINISJON"

Ved "random surfing"

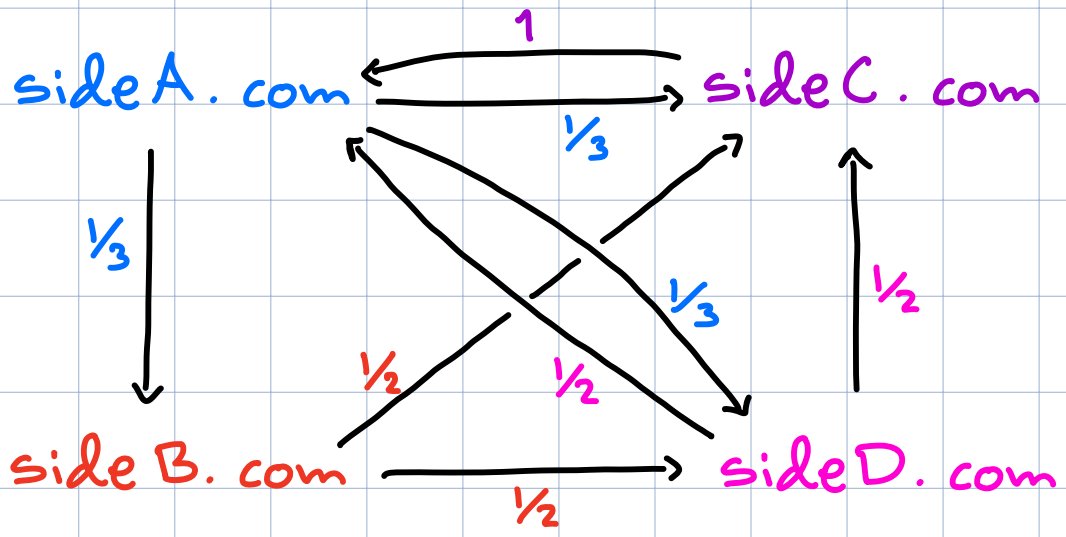
$$\text{viktigheten}(k) \overset{(*)}{=} \sum_j \left(\begin{array}{c}\text{sannsynligheten for å gå}\\ \text{fra } j \text{ til } k\end{array}\right) \cdot \text{viktigheten}(j).$$

MERK

La P være matrisa som inneholder "sannsynlighetene for å gå fra j til k" og la $\bar{x}$ være en vektor som inneholder viktigheten til hver nettside. Da blir $(*)$ til $\bar{x} = P\bar{x}$, så det å bestemme nettsidenes viktighet er et egenvektorproblem.

EKSEMPEL (IGJEN)

Anta at lenkestrukturen ser slik ut:



Overgangsmatrise:

Regulær!

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

EKSEMPEL (IGJEN)

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

Altså er "viktighetsvektoren" $\bar{x}$ slik at $P\bar{x} = \bar{x}$, i.e. $(P - I_{4\times 4})\bar{x} = \bar{0}$. Generell løsning er på formen

$$s \cdot \begin{pmatrix} 12 \\ 4 \\ 9 \\ 6 \end{pmatrix}, \quad s \in \mathbb{R},$$

så vi velger $s = 1/31$ og får

$$\bar{x} \approx \begin{pmatrix} 0.38 \\ 0.12 \\ 0.29 \\ 0.19 \end{pmatrix}.$$

RANGERING:

1. sideA.com
2. sideC.com
3. sideD.com
4. sideB.com

PROBLEMER

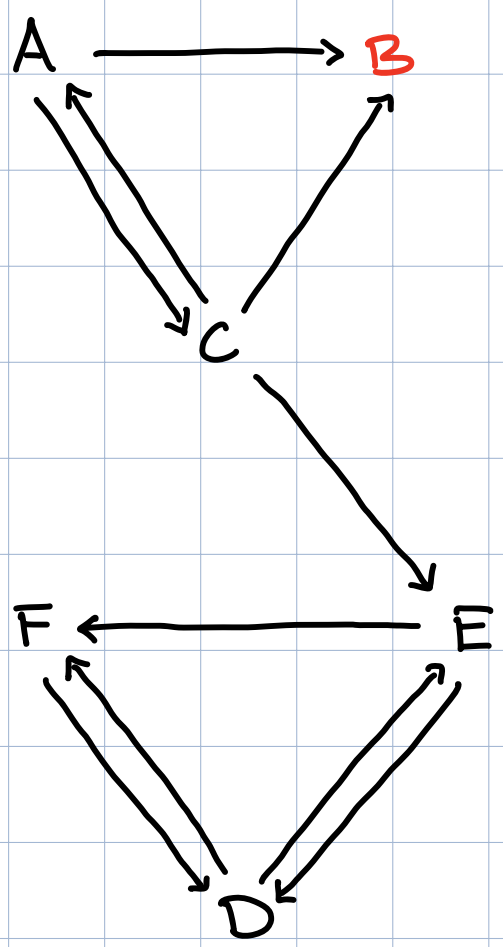
I Hva om en side har null lenker ut?

II Surferen vil av og til kjede seg og "teleportere" til en helt tilfeldig side.

2.1.2. Intuitive justification

PageRank can be thought of as a model of user behavior. We assume there is a "random surfer" who is given a Web page at random and keeps clicking on links, never hitting "back" but eventually gets bored and starts on another random page. The probability that the random surfer visits a page is its PageRank.

I Hva om en side har null lenker ut?



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

$\}$ I

$$\begin{pmatrix} 0 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 1/2 & 1 \\ 0 & 1/6 & 1/3 & 1/2 & 0 & 0 \\ 0 & 1/6 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

I Hva om en side har null lenker ut?

Generelt: Hvis vi har n nettsider, så blir overgangsmatrisa ei $(n \times n)$-matrise. Hvis en kolonne i denne matrisa består av kun 0-ere, erstatt hver av dem med $1/n$.

Dette skaffer oss ei stokastisk matrise!

(Men ikke nødvendigvis ei regulær matrise.)

II Surferen vil av og til kjede seg og "teleportere" til en helt tilfeldig side.

Generelt: La α være sannsynligheten for at surferen "teleporterer" til en helt tilfeldig side.

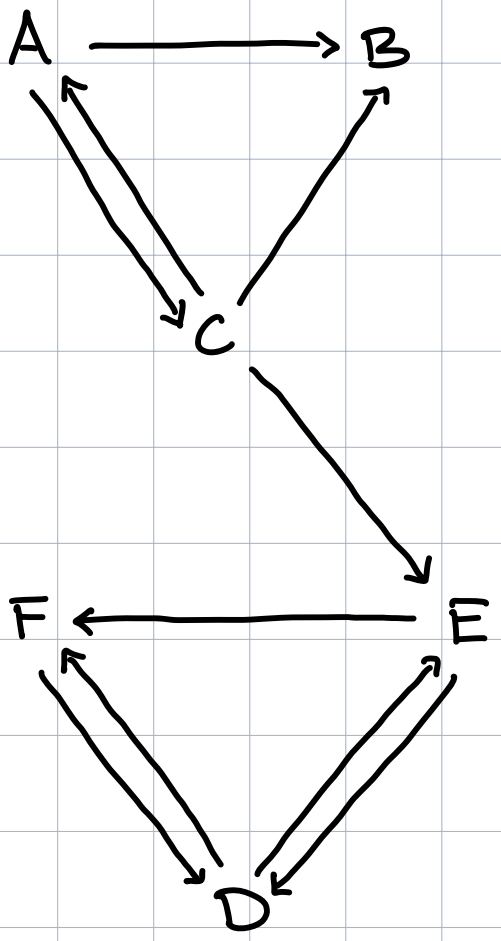
Hva om vi legger til α i hver posisjon i overgangsmatrisa?

↳ Da får vi ikke lenger ei stokastisk matrise.

Løsning: I hver posisjon i overgangsmatrisa multipliserer vi med $1-\alpha$ og legger til α/n

Dette skaffer oss ei regulær matrise!

II Surferen vil av og til "teleportere" til en helt tilfeldig side.

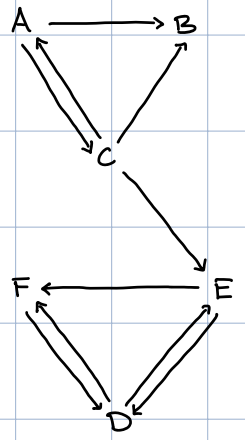


I ⇝

$$\begin{pmatrix} 0 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 1/2 & 1 \\ 0 & 1/6 & 1/3 & 1/2 & 0 & 0 \\ 0 & 1/6 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

⇝ II med $\alpha = 1/10$

$$\begin{pmatrix} 1/60 & 1/6 & 19/60 & 1/60 & 1/60 & 1/60 \\ 7/15 & 1/6 & 19/60 & 1/60 & 1/60 & 1/60 \\ 7/15 & 1/6 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/6 & 1/60 & 1/60 & 7/15 & 11/12 \\ 1/60 & 1/6 & 19/60 & 7/15 & 1/60 & 1/60 \\ 1/60 & 1/6 & 1/60 & 7/15 & 7/15 & 1/60 \end{pmatrix}$$



$$P = \begin{pmatrix} 1/60 & 1/6 & 19/60 & 1/60 & 1/60 & 1/60 \\ 7/15 & 1/6 & 19/60 & 1/60 & 1/60 & 1/60 \\ 7/15 & 1/6 & 1/60 & 1/60 & 1/60 & 1/60 \\ 1/60 & 1/6 & 1/60 & 1/60 & 7/15 & 11/12 \\ 1/60 & 1/6 & 19/60 & 7/15 & 1/60 & 1/60 \\ 1/60 & 1/6 & 1/60 & 7/15 & 7/15 & 1/60 \end{pmatrix}$$

Viktighetsvektoren (a.k.a. PageRank-vektoren) er den sannsynlighetsvektoren $\bar{x}$ som oppfyller $P\bar{x} = \bar{x}$, altså $(P - I_{6\times 6})\bar{x} = \bar{0}$. En datamaskin vil fortelle oss at

$$\bar{x} \approx \begin{pmatrix} 0.037 \\ 0.054 \\ 0.041 \\ 0.375 \\ 0.206 \\ 0.286 \end{pmatrix} \Rightarrow$$

RANGERING:

D, F, E, B, C, A

(fra høy til lav viktighet/relevans).