Mikael Smedbäck

F12: PageRank

Hur sker sökning på internet?

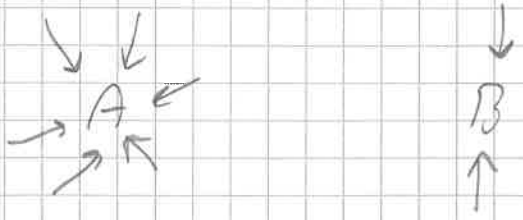
* Webcrawlers besöker hemsidor och skapar en indexeringsfil
* Indexeringsfilens utseende:

ord	förekomst på hemsidor
ost	2, 4, 10
potatis	4, 8
bil	1, 2

exempel:

Hemsida 2 innehåller orden "ost" och "bil". Ordet "bil" förekommer på hemsida 1 och 2.

* Sök på ordet "ost" ger två träffar.
* Ranking före Google: popularitetsranking: En sida rankas högre om den har många pekare (länkar) till sig:



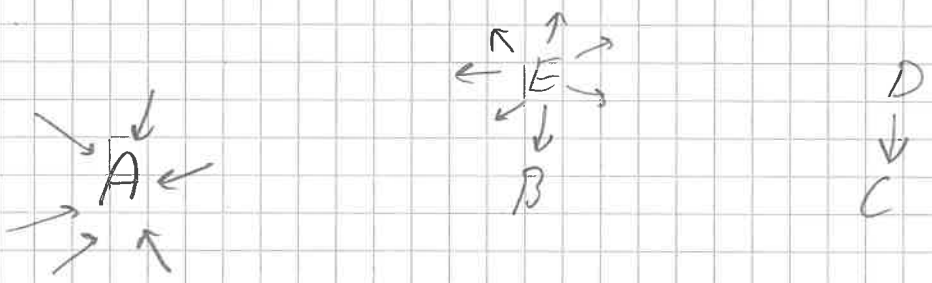
1. Hemsida A
2. Hemsida B
:

* problem: länkar kan "köpas"

PageRank

* Google använder PageRank, skapat av Sergey Brin och Larry Page
* Idén är att även ta hänsyn till kvaliteten på de rekommenderande länkarna!

Exempel: Person A, B och C söker ett jobb:

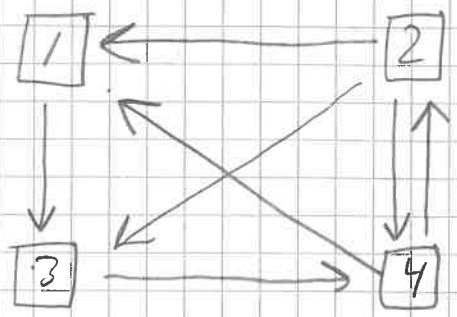


Person A får 1000 rekommendationsbrev - bra
Person B får ett brev, men av Einstein - bättre
Person C får ett brev, av Dirac, som bara skriver ett enda brev - bäst!

Matematisk formulering

Exempel:

$|H_1| = 1$ $\quad$ $|H_2| = 3$



$|H_3| = 1$ $\quad$ $|H_4| = 2$

* Figuren visar fyra hemsidor och länkarna (pekarna) mellan dem.
* $|H_2| = 3$ anger att tre länkar utgår från hemsida 2, o.s.v.
* Rankingen (relevansen) för hemsida i ges av en summa av alla länkar in till sidan, d.v.s. som i popularitetsrankingen
* Men nu viktas varje länk via de länkande sidornas ranking och antalet utgående länkar:

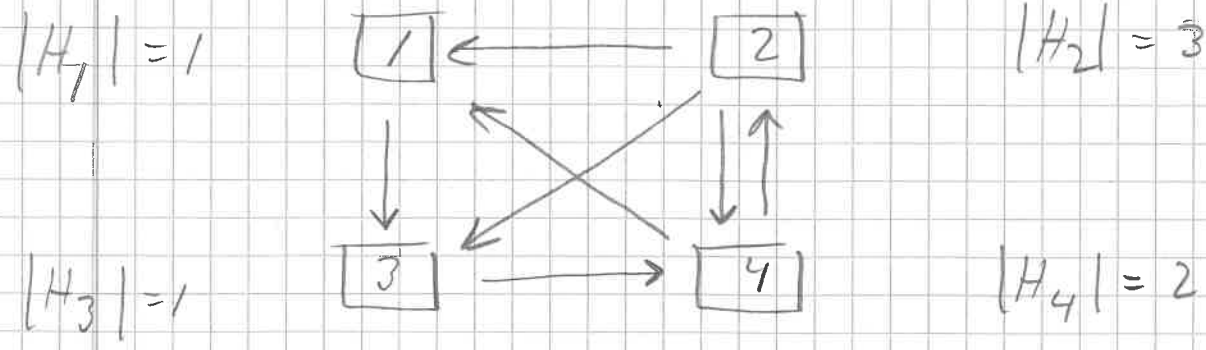
$$r(1) = r(H_1) = \sum_{k \to H_1} \frac{r(H_k)}{|H_k|} = \frac{r(2)}{3} + \frac{r(4)}{2}$$

$$r(2) = r(H_2) = \frac{r(4)}{2}$$

$$r(3) = r(H_3) = \frac{r(1)}{1} + \frac{r(2)}{3}$$

$$r(4) = r(H_4) = \frac{r(2)}{3} + \frac{r(3)}{1}$$

Hyperlänkmatrisen



* Definition av hyperlänkmatrisen A:

$$A_{ik} = \begin{cases} \frac{1}{|H_k|} & \text{om länk } \boxed{k} \to \boxed{i} \text{ finns} \\ 0 & \text{annars} \end{cases}$$

* I exemplet ovan:

$\frac{1}{|H_4|}$ etc.

$$A = \begin{bmatrix} 0 & \frac{1}{3} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & \frac{1}{2} \\ 1 & \frac{1}{3} & 0 & 0 \\ 0 & \underbrace{\frac{1}{3}}_{\text{antal länkar ut från } \boxed{2}} & 1 & 0 \end{bmatrix}$$

← alla länkar till 1
← alla länkar till 2
← alla länkar till 3
← alla länkar till 4

* $\frac{1}{3}(1+1+1) = 1$

alla ingående länkar vars ursprung är 2

d.v.s. A är en stokastisk matris!

Rankingvektorn

* Vi söker rankingvektorn:

$$\vec{r} = \begin{bmatrix} r(1) & r(2) & r(3) & r(4) \end{bmatrix}$$

* Rankingvektorn är en utfallsvektor, om vi kräver $r(1) + r(2) + r(3) + r(4) = 1$.

* Ekvation för rankingvektorns komponenter:

$$r(i) = \sum_{k \to H_i} \frac{r(k)}{|H_k|}$$

(term bidrar endast om länk $k \to i$ finns)

* På matrisform: (om alla länkar finns, annars noll)

$$A = \begin{bmatrix} 0 & \frac{1}{|H_2|} & \frac{1}{|H_3|} & \frac{1}{|H_4|} \\ \frac{1}{|H_1|} & 0 & \frac{1}{|H_3|} & \frac{1}{|H_4|} \\ \frac{1}{|H_1|} & \frac{1}{|H_2|} & 0 & \frac{1}{|H_4|} \\ \frac{1}{|H_1|} & \frac{1}{|H_2|} & \frac{1}{|H_3|} & 0 \end{bmatrix}$$

ger

$$A\vec{r} = \vec{r}$$

Tillbaka till exemplet

* Vår ekvation: $(I - A)\vec{r} = \vec{r} - A\vec{r} = \vec{0}$, d.v.s.

$$\begin{bmatrix} 1 & -1/3 & 0 & -1/2 \\ 0 & 1 & 0 & -1/2 \\ -1 & -1/3 & 1 & 0 \\ 0 & -1/3 & -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

* Genomför Gausseliminationen: (efter multiplikation 6,2,3,3)

$$\begin{bmatrix} 6 & -2 & 0 & -3 \\ 0 & 2 & 0 & -1 \\ -3 & -1 & 3 & 0 \\ 0 & -1 & -3 & 3 \end{bmatrix} \sim \begin{bmatrix} -3 & -1 & 3 & 0 \\ 0 & 2 & 0 & -1 \\ 6 & -2 & 0 & -3 \\ 0 & -1 & -3 & 3 \end{bmatrix} \textcircled{2} \sim$$

$$\sim \begin{bmatrix} -3 & -1 & 3 & 0 \\ 0 & 2 & 0 & -1 \\ 0 & -4 & 6 & -3 \\ 0 & -1 & -3 & 3 \end{bmatrix} \sim \begin{bmatrix} -3 & -1 & 3 & 0 \\ 0 & -1 & -3 & 3 \\ 0 & -4 & 6 & -3 \\ 0 & 2 & 0 & -1 \end{bmatrix} \textcircled{-4} \; \textcircled{2} \sim$$

$$\begin{bmatrix} -3 & -1 & 3 & 0 \\ 0 & -1 & -3 & 3 \\ 0 & 0 & 18 & -15 \\ 0 & 0 & -6 & 5 \end{bmatrix} \textcircled{1/3} \sim \begin{bmatrix} -3 & -1 & 3 & 0 \\ 0 & -1 & -3 & 3 \\ 0 & 0 & 6 & -5 \\ 0 & 0 & -6 & 5 \end{bmatrix} \textcircled{1} \sim$$

→

$$\rightarrow \sim \begin{bmatrix} (-3) & -1 & 3 & 0 \\ 0 & (-1) & -3 & 3 \\ 0 & 0 & (6) & -5 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \\ \times 2 \\ \times 2 \\ \\ \end{matrix} \sim \begin{bmatrix} -6 & -2 & 6 & 0 \\ 0 & -2 & -6 & 6 \\ 0 & 0 & 6 & -5 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \leftarrow \\ \leftarrow \\ (1) \\ (-1) \end{matrix} \quad (7)$$

(trappstegsform)

$$\begin{bmatrix} -6 & -2 & 0 & 5 \\ 0 & -2 & 0 & 1 \\ 0 & 0 & 6 & -5 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \leftarrow \\ (-1) \\ \\ \\ \end{matrix} \sim \begin{bmatrix} (-6) & 0 & 0 & 4 \\ 0 & (-2) & 0 & 1 \\ 0 & 0 & (6) & -5 \\ 0 & 0 & 0 & 0 \\ x & y & z & w \end{bmatrix}$$

* Lösningen är: $(t \in \mathbb{R})$

$$\begin{cases} -6x = -4t \\ -2y = -t \\ +6z = 5t \\ w = t \end{cases} \Leftrightarrow \begin{cases} x = \frac{2}{3}t \\ y = \frac{1}{2}t \\ z = \frac{5}{6}t \\ w = t \end{cases}$$

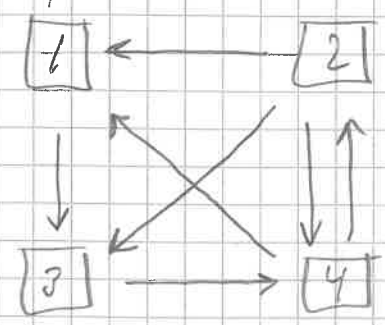
Sätt $t = 6$ ger ($t = \frac{1}{2}$ ger utfallsvektor)

$x = 4 \quad y = 3 \quad z = 5 \quad w = 6$

$r(1) = 4 \quad r(2) = 3 \quad r(3) = 5 \quad r(4) = 6$

* Googles sökresultat blir:

1.	Hemsida	4
2.	Hemsida	3
3.	Hemsida	1
4.	Hemsida	2



Kommentarer

* Att hitta exakta "egenvektorer" $\vec{r}$ till ekvationen $A\vec{r} = \vec{r}$ kan bli komplicerat när matrisen A är stor.
* Om en slutfördelning existerar, d.v.s. om gränsvärdet $\lim_{n\to\infty} A^n \vec{r}_0$ existerar, för någon startgissning $\vec{r}_0$: Då gäller:

$$A^{n+1}\vec{r}_0 \approx A^n \vec{r}_0,$$

d.v.s. en approximativ "lösning" är

$$\vec{r} = A^n \vec{r}_0$$

för n tillräckligt stort.

* Google använder PageRank + finjusteringar.
* PageRank rankar enbart med hjälp av länkar (pekare), utan hänsyn till innehåll. D.v.s. hela sidan får samma ranking, oavsett vilket av sidans ord man sökte på.
Slutsats: bättre att söka "specifikt", så att PageRank kan ranka relevanta sidor.