MA1487 Matematisk modellering
Google PageRank

Henrik Fredriksson
henrik.fredriksson@bth.se

12 december 2018

TIMN

Introduktion

PageRank

Hyperlink-matrisen

Introduktion

Introduktion

- Vad är det som gör att att Google oftast ger relevanta resultat för det sökord som använts?
- De tidigare sökmotorerna räknade antalet förekomster av sökordet och rankade sökresultaten efter detta.
- Ett annat sätt att rangordna sökresultaten är låta antalet sidor som länkar till en viss sida vara ett mått på hur viktig den är. Dock så kan sidor vara viktiga även om det är så många andra sidor som länkar till den.

En sökmotor vill

1. Låta spindlar (web crawler) hitta webbsidor som är publika
2. Indexera webbsidorna så att man kan hitta de sidor som innehåller sökordet eller sökfrasen
3. Ranka sidorna så att användaren får upp sökresult med önskad information.

PageRank

PageRank

Hur funkar Googles rankingalgoritm PageRank?

Varje sida p har ett mått $r(p)$ som beskriver hur viktig sidan är (sidans PageRank).

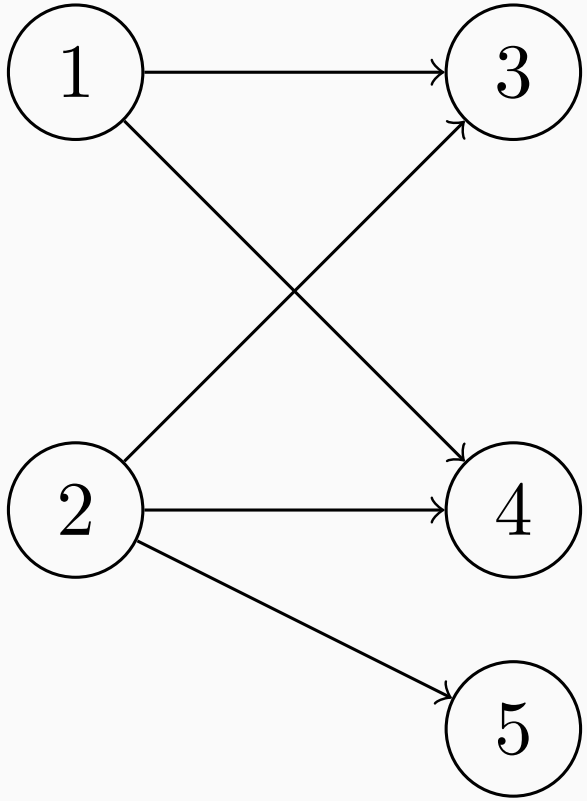
Antag att sidan p_j har l_j länkar till andra sidor. Om någon av dessa länkar till sidan p_i så lämnar sidan p_j över $\frac{1}{l_j}$ av sin PageRank till sidan p_i.

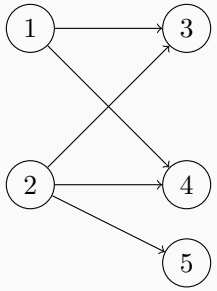
PageRanken för sidan p_i blir summan av alla bidragen som länkar till sidan p_i.

Låt M_i vara mängden av alla sidor som har länkar till sidan p_i. PageRanken för sidan p_i är alltså

$$r(p_i) = \sum_{p_j \in M_i} \frac{r(pj)}{l_j}$$

Betrakta följade miniwebbnätverk. Pilarna innebär att det finns en länk mellan sidorna.





Antag att sidan p_1 har PageRank $r(p_1)$ och sidan p_2 har PageRank $r(p_2)$. PageRanken för sidorna p_3, p_4 och p_5 blir

$$r(p_3) = \frac{r(p_1)}{2} + \frac{r(p_2)}{3}$$

$$r(p_4) = \frac{r(p_1)}{2} + \frac{r(p_2)}{3}$$

och

$$r(p_5) = \frac{r(p_2)}{3}$$

Tolkning av PageRank

En slösurfare besöker en slumpmässig sida och klickar slumpmässigt på länkar tills hen tröttnar och börjar om på en ny slumpmässig sida.

PageRanken för en sida p är sannolikheten att slösurfaren besöker sidan p.

En annan tolkning är att en sida har hög PageRank om många andra sidor länkar till den för att den har relevatant innehåll.

Beräkning av Hyperlink-matrisen

För att bestämma PageRank för en sida p_i behöver vi veta PageRank hos alla sidor som länkar till den. Cirkulerar inte detta?

Detta går att lösa med linjär algebra.

Egenvärde och egenvektor

Låt $\boldsymbol{A}$ vara en matris och $\boldsymbol{v}$ en vektor sådana att

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}.$$

Man säger då att $\boldsymbol{v}$ är en *egenvektor* till $\boldsymbol{A}$ med *egenvärde* λ

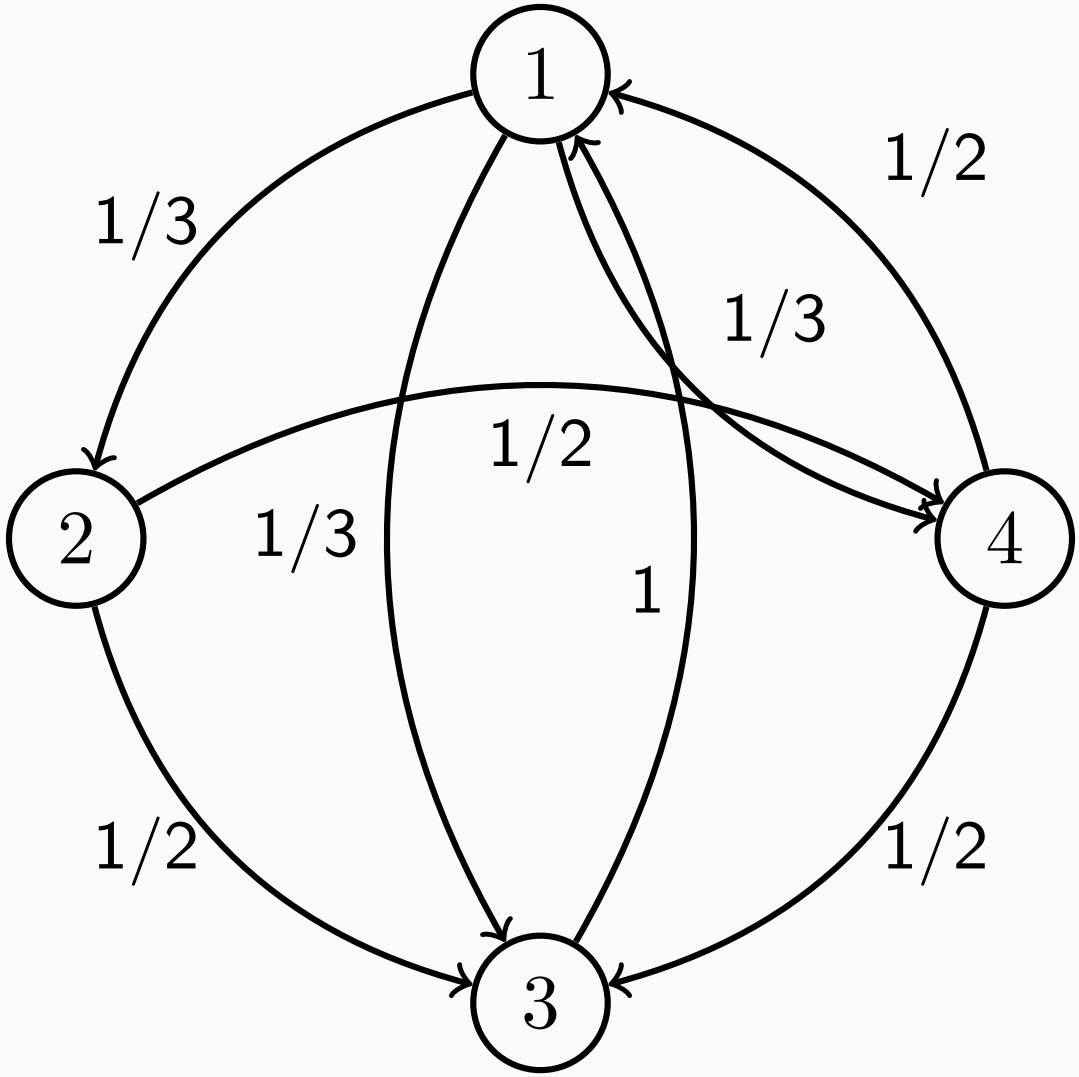
Hyperlink-matrisen

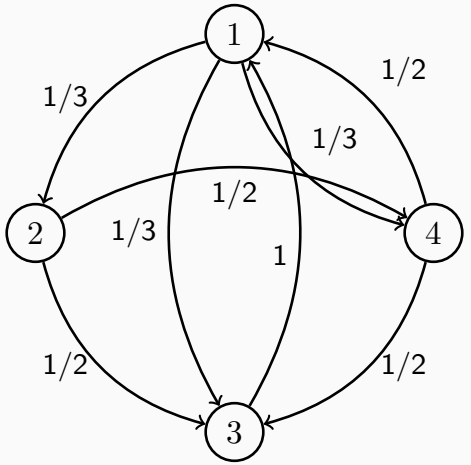
Hyperlink-matrisen

Vi inför den så kallade *hyperlink-matrisen*, $H = (h_{ij})$ som defineras som

$$h_{ij} = \begin{cases} \frac{1}{l_j} \text{ om sidan } p_j \text{ länkar till sidan } p_i \\ 0 \text{ annars.} \end{cases}$$

Här är i-radindex och j-kolumnindex och l_j antalet länkar från sidan p_j.


1
2
3
4
1/3
1/2
1/3
1/2
1/3
1
1/2
1/2



Hyperlink-matrisen H blir

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

Notera att summan kolumnerna blir 1. Matrisen är s.k. kolumnstokastisk.

Om vi inför vektorn $\boldsymbol{r} = (r(p_1), r(p_2), \ldots)$ så kan definitionen av PageRank skriva som

$$\boldsymbol{r} = \boldsymbol{H}\boldsymbol{r}$$

vilket innebär att vektorn $\boldsymbol{r}$ är en egenvektor till matrisen $\boldsymbol{H}$ med egenvärdet 1.

Potensmetoden

Hur hittar vi vektorn $\boldsymbol{r}$?

Vi betraktar $\boldsymbol{r} = \boldsymbol{H}\boldsymbol{r}$ som ett dynamiskt system.

Sätt $\boldsymbol{r} = \left(\frac{1}{n}, \frac{1}{n}, \frac{1}{n}, \ldots\right)$ där n är antalet sidor vi vill undersöka.

Från början har alltså alla sidor samma PageRank-värde.

Vi kan nu beräkna $\boldsymbol{H}\boldsymbol{r}$, $\boldsymbol{H}^2\boldsymbol{r}$, $\boldsymbol{H}^3\boldsymbol{r}, \ldots$ tills systemet har nått ett jämnviktstillstånd.

$$\boldsymbol{H} = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}, \boldsymbol{r} = \begin{pmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \end{pmatrix}$$

$$\boldsymbol{Hr} = \begin{pmatrix} 0.375000000000000 \\ 0.0833333333333333 \\ 0.333333333333333 \\ 0.208333333333333 \end{pmatrix}, \boldsymbol{H^2 r} = \begin{pmatrix} 0.437500000000000 \\ 0.125000000000000 \\ 0.270833333333333 \\ 0.166666666666667 \end{pmatrix}$$

$$\boldsymbol{H^3 r} = \begin{pmatrix} 0.354166666666667 \\ 0.145833333333333 \\ 0.291666666666667 \\ 0.208333333333333 \end{pmatrix}, \boldsymbol{H^7 r} = \begin{pmatrix} 0.389756944444444 \\ 0.127314814814815 \\ 0.290509259259259 \\ 0.192418981481481 \end{pmatrix}$$

$$\boldsymbol{H^8 r} = \begin{pmatrix} 0.386718750000000 \\ 0.129918981481481 \\ 0.289785879629630 \\ 0.193576388888889 \end{pmatrix} \boldsymbol{H^9 r} = \begin{pmatrix} 0.386574074074074 \\ 0.128906250000000 \\ 0.290653935185185 \\ 0.193865740740741 \end{pmatrix}$$

I det illustrativa webbnätverket så har alltså sidan p_1 högst PageRank. Ovanstående metod funkar ifall det inte finns några *dangling nodes* (sidor som inte länkar till några andra sidor) och nätverket är sammanhängade (består inte av åtskijilda subnätverk).

Vi byter ut matrisen $\boldsymbol{H}$ mot matrisen

$$\boldsymbol{M} = (1 - m)\boldsymbol{H} + m\boldsymbol{S}$$

där $\boldsymbol{S}$ är en matris som har samma dimension som $\boldsymbol{H}$ och samtliga element är $1/n$.

Värdet på m är en vikt och brukar sättas till $m = 0.15$

PageRank är idag en av flera komponenter Google använder för att ranka webbsidor

Vidare läsning

1. https://backlinko.com/google-ranking-factors
2. https://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf
3. http://infolab.stanford.edu/~backrub/google.html
4. http://www.cs.bham.ac.uk/~pxt/IDA/pagerank.pdf
5. https://pi.math.cornell.edu/~levine/4740/DeeperInsidePageRank.pdf