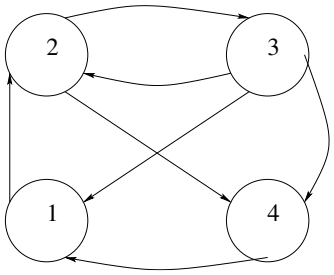
21.6. Hur Linjär algebra skapade Google

Linjär algebra ligger bakom den idag mest överlägsna sökmotorn av alla; Google.
Dagens teknik för sökmotorer använder algoritmer för länkanalys. Dessa algoritmer bygger på teorin för matriser och en av dem mest populära och mest användbara algoritmer för länkanalys är The PageRank algorithm.
Sergey Brin och Larry Page (den senare har gett namn till The Pageranking algoritmen), två doktorander vid Stanford universitet, med sin idé om hur en sökning kan gå till lade 1998 grunden för att Google skulle få det övertag och den dominans den idag har bland alla sökmotorer.
Brin and Page definierade vad som menas med em "*viktig*" sida, dvs hur rankning av en sida ska gå till. Deras idé var att

en sida betraktas som viktig om andra viktiga sidor länkar till den.

Vi illustrerar grundidéerna med ett enkelt exempel. Antag att vi har 4 sidor som länkar till varandra enligt figuren nedan

Figur 21.12.



Låt x_j, $j = 1, 2, 3, 4$, vara rankningen hos sidan j. Ranknigen x_j för sida j definieras som en linjärkombination i rankningen av dem sidor som länkar till just sidan j. Ur figur följer att

$$\begin{cases} x_1 = & & \lambda_3 x_3 + \lambda_4 x_4 \\ x_2 = \lambda_1 x_1 + & \lambda_3 x_3 & \\ x_3 = & \lambda_2 x_2 & \\ x_4 = & \lambda_2 x_2 + \lambda_3 x_3 & \end{cases} \tag{21.21}$$

För att linjärkombinationen ska uppfylla dem krav som The Pageranking algoritm ställer låter vi

$$\lambda_j = \frac{1}{\text{antal länkar ut från sidan } j}, \qquad j = 1, 2, 3, 4.$$

Några av dem viktiga idéerna bakom The PageRanking algoritmen i systemet (21.21) är

- En sida ska inte bli viktig genom att länka till sig själv. Därför ser vi till att x_1 inte återfinns i högra ledet i första ekvationen. Detsamma gäller för x_2, x_3 och x_4 som inte återfinns i högra ledet i respektive ekvation.
- En sida ska inte bli viktig genom att länka till så många andra sidor. Detta ser vi till genom att i linjärkombinationen dividerar vi med antal länkar en sida gör.

- Om sida j inte länkar ut till någon annan sida sätter vi $\lambda_j = 0$.

I vårt exempel följer ur figuren att är $\lambda_1 = \frac{1}{1} = 1$, $\lambda_2 = \frac{1}{2}$, $\lambda_3 = \frac{1}{3}$ och $\lambda_4 = \frac{1}{1} = 1$. Sätter vi in dessa värden i systemt (21.21) ovan så får vi att

$$\begin{cases} x_1 &= x_3/3 + x_4 \\ x_2 &= x_1 + x_3/3 \\ x_3 &= x_2 \\ x_4 &= x_2/2 + x_3/3 \end{cases} \tag{21.22}$$

Vi skriver nu ekvationssystemt (21.22) ovan på matrisform och får

$$\begin{pmatrix} x_3/3 + x_4 \\ x_1 + x_3/3 \\ x_2 \\ x_2/2 + x_3/3 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \Leftrightarrow \begin{pmatrix} 0 & 0 & 1/3 & 1 \\ 1 & 0 & 1/3 & 0 \\ 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 1/3 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

$$\Leftrightarrow$$

$$H\boldsymbol{v} = \boldsymbol{v}. \tag{21.23}$$

Matrisen H kallar vi för hyperlänksmatrisen. Eftersom elemneten h_{ij} i matrisen H ligger mellan 0 och 1 samt att kolonnsumman är 1, så är H en stokastisk matris (se Övning 7.30). Dessutom kan elemneten h_{ij} tolkas som sannolikheten att sidan i länkar till sidan j.
Rankningen hos varje sida bestäms alltså av egenvektorn $\boldsymbol{v}$ som hör ihop med egenvärdet $\lambda = 1$ i ekvation (21.23), dvs ekvationen $H\boldsymbol{v} = \lambda\boldsymbol{v}$.
Vi bestämmer nu egenvektorn i (21.22) och får

$$\boldsymbol{v} = H\boldsymbol{v} \Leftrightarrow (E - H)\boldsymbol{v} = 0 \Leftrightarrow \left(\begin{array}{cccc|c} 1 & 0 & -1/3 & -1 & 0 \\ -1 & 1 & -1/3 & 0 & 0 \\ 0 & -1/2 & 1 & 0 & 0 \\ 0 & -1/2 & -1/3 & 1 & 0 \end{array}\right)$$

som har lösningen $x_1 = 5t$, $x_2 = 6t$, $x_3 = 3t$ och $x_4 = 4t$. Vi vill att totala rankningssumman skall vara 1, dvs $x_1 + x_2 + x_3 + x_4 = 1$, och väljer därför att sätta $t = 18$. Därmed får vi egenvektorn $\frac{1}{18}\begin{pmatrix} 5 \\ 6 \\ 3 \\ 4 \end{pmatrix}$. Vi ser nu att sida 2 är den som är högst rankad och är därmed den mest viktiga.
Det här var ett litet exempel där vi hade 4 sidor och där matrisen H är typen 4×4. I allmänhet är H av typen $n \times n$, där n är väldigt stort positivt heltal. Detta får till följd att lösa egenvärdesproblemet (21.23), dvs

$$\boldsymbol{v} = H\boldsymbol{v} \Leftrightarrow (E - H)\boldsymbol{v} = \boldsymbol{0}$$

blir ganska svårt. Matrisen $E - H$ kan vara känslig för störningar, t.ex., avrundningar eller division med små tal, så att Gausselimination inte ger noggranna svar.
Därför behövs det andra idéer för att lösa ut egenvektorn i ekvation (21.23). En sådan metod är **potensmetoden** som går ut på att lösa ekvation (21.23) iterativt. Fördelen här är att potensmetoden kräver endast multiplikation och addition. Detta går dessutom ganska

snabbt eftersom matrisen H är "gles", dvs innehåller många nollor. Potensmetoden är en rekursiv metod som med hjälp av en föregående rankningsvektor $\boldsymbol{v}^{k-1}$ och matrisen H bestämmer nästa rankningsvektor $\boldsymbol{v}^k$ via

$$\boldsymbol{v}^k = H\boldsymbol{v}^{k-1}, \qquad k = 1, 2, 3, \cdots . \tag{21.24}$$

Vi behöver dock en startvektor $\boldsymbol{v}^0$ för att tillämpa metoden. Låt oss anta att vid starten så är alla sidor lika rankade, dvs $\boldsymbol{v}^0 = \frac{1}{4}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$. Vi utför ett antal iterationer med potensmetoden och får att

$$\boldsymbol{v}^1 = H\boldsymbol{v}^0 = \frac{1}{24}\begin{pmatrix} 8 \\ 8 \\ 3 \\ 5 \end{pmatrix}, \qquad \boldsymbol{v}^2 = H\boldsymbol{v}^1 = \frac{1}{24}\begin{pmatrix} 6 \\ 9 \\ 4 \\ 5 \end{pmatrix}, \qquad \boldsymbol{v}^3 = H\boldsymbol{v}^2 = \frac{1}{144}\begin{pmatrix} 38 \\ 44 \\ 27 \\ 35 \end{pmatrix},$$

$$\boldsymbol{v}^4 = H\boldsymbol{v}^3 = \frac{1}{24}\begin{pmatrix} 44 \\ 47 \\ 22 \\ 31 \end{pmatrix}, \qquad \cdots \quad , \boldsymbol{v}^{14} = H\boldsymbol{v}^{13} = \begin{pmatrix} 0.274 \\ 0.331 \\ 0.169 \\ 0.225 \end{pmatrix}.$$

Redan efter 14 iterationer ser vi att har två decimaler rätt när vi jämför $\boldsymbol{v}^{14}$ med egenvektorn $\boldsymbol{v} = \frac{1}{18}\begin{pmatrix} 5 \\ 6 \\ 3 \\ 4 \end{pmatrix} \approx \begin{pmatrix} 0.278 \\ 0.333 \\ 0.167 \\ 0.222 \end{pmatrix}$. För att beräkna iteraten $\boldsymbol{v}^k$, så kan beräkningsverktyg som Maple och Matlab användas.
Även potensmetoden ger numeriskt att sida 2 är högst rankad.
Det bör dock påpekas att det uppstår ett antal frågor i samband med att lösa egenvektorn med potensmetoden i ekvationen (21.23). Vi nämner nedan några av dessa som vi inte går in på att svara här.

1. Har hyperlänkmatrisen H ett egenvärde $\lambda = 1$ med en tillhörande egenvektor $\boldsymbol{v}$?
2. Konvergerar potensmetoden? Konvergerar den för varje startvektor $\boldsymbol{v}^0$?
3. Om potensmetoden konvergerar, vad konvergerar den mot?