LinAlgDat

Google's page rank

Resumé

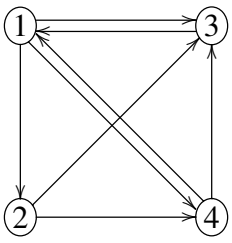
Vi viser hvordan lineære ligninger naturligt optræder i forbindelse med en simpel udgave af Google's algoritme for at vise de mest interessante links først i en søgning på internettet.

1 Web og linkmatricer

Når man foretager en søgning på internettet med Google vises de sider som indeholder søgeordene i en given rækkefølge, og det er ret ofte, at de første links der vises, faktisk er de mest brugbare. Vi skal se lidt nærmere på hvilken strategi der kan anvendes for at få en "god" rangordning af siderne. Vi vil i det væsentlige følge artiklen [1].

Eksempel 1 *Vi betragter et (meget) lille web med 4 sider, kaldet 1,2,3,4. Disse sider refererer (linker) til hinanden på følgende måde:*

Side 1 refererer til side 2, 3 og 4
Side 2 refererer til side 3 og 4
Side 3 refererer til side 1
Side 4 refererer til side 1 og 3



Figuren viser en orienteret graf, med $\{1,2,3,4\}$ *som de fire knuder (punkter) og med pile mellem knuderne. Retningen af en pil afgør hvilken side der refererer til hvilken, fx betyder pilen fra* 1 *til* 2*, at side* 1 *har et link til side* 2*. Man kan tænke på, at side* 2 *er vigtig, fordi der er en side der henviser til den.*

Vi vil nu definere en "score" for hver side, der viser hvor vigtig den er. Vi opridser to muligheder.

- *Første mulighed går ud på følgende: Vi kunne sætte scoren* x_k *for side* k *til antallet af sider der henviser til side* k*. Dermed er* $x_1 = 2$*,* $x_2 = 1$*,* $x_3 = 3$ *og* $x_4 = 2$*. Altså er side* 3 *den højst rangerende. Denne metode tager imidlertid ikke hensyn til at et link fra en vigtig side bør øge sidens vigtighed mere end et link fra en ikke så vigtig side. Så denne mulighed er nok alligevel for simpel.*

- *En anden og bedre mulighed er at sætte scoren x_k for side k til summen af alle scorerne for de sider, der henviser til side k (i stedet for antallet af sider der henviser til side k). Det giver fx, at $x_1 = x_3 + x_4$. De øvrige ligninger udtrykker x_2, x_3 og x_4 ved kombinationer af x_1, x_2, x_3 og x_4 .*

  Et problem ved denne tilgang er, at en webside med mange udgående links får større inflydelse på de andre siders score end en side med få links. Dette omgås ved en "normalisering", hvor den samlede score som en side kan give til andre sættes til 1: hvis side j indeholder et link til side k og N_j links i alt, så øger vi scoren for side k med x_j/N_j (i stedet for x_j). Vi vil benytte denne metode for at bestemme rangordenen af siderne i webbet.

Overvejelserne oven for giver $N_1 = 3$, $N_2 = 2 = N_4$ og $N_3 = 1$ og dermed følgende ligninger for tallene x_1, x_2, x_3 og x_4:

$$\begin{aligned} x_1 &= x_3 + 1/2 x_4 \\ x_2 &= 1/3 x_1 \\ x_3 &= 1/3 x_1 + 1/2 x_2 + 1/2 x_4 \\ x_4 &= 1/3 x_1 + 1/2 x_2. \end{aligned}$$

Dette ligningssystem kan skrives på formen $\mathbf{Ax} = \mathbf{x}$, hvor

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix} \quad og \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}.$$

Vi omskriver systemet til $\mathbf{Ax} - \mathbf{x} = 0$ og opskriver den tilsvarende totalmatrix:

$$\left[\begin{array}{cccc|c} -1 & 0 & 1 & 1/2 & 0 \\ 1/3 & -1 & 0 & 0 & 0 \\ 1/3 & 1/2 & -1 & 1/2 & 0 \\ 1/3 & 1/2 & 0 & -1 & 0 \end{array}\right].$$

Vi løser ligningssystemet ved at lave rækkeoperationer på totalmatricen! Resultatet bliver (se begrundelse nedenfor)

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = t \begin{bmatrix} 2 \\ 2/3 \\ 3/2 \\ 1 \end{bmatrix}.$$

Heraf ses, at rangordningen bør være: side 1 er den vigtigste, så side 3, derefter side 4 og endelig side 2.

Resultatet kan opnås ved følgende rækkeoperationer: først multipliceres rækkerne med hhv 2,6,6 og 6;

$$\left[\begin{array}{cccc|c} -1 & 0 & 1 & 1/2 & 0 \\ 1/3 & -1 & 0 & 0 & 0 \\ 1/3 & 1/2 & -1 & 1/2 & 0 \\ 1/3 & 1/2 & 0 & -1 & 0 \end{array}\right] \xrightarrow[6\mathbf{r}_4 \to \mathbf{r}_4]{\begin{array}{c} 2\mathbf{r}_1 \to \mathbf{r}_1 \\ 6\mathbf{r}_2 \to \mathbf{r}_2 \\ 6\mathbf{r}_3 \to \mathbf{r}_3 \end{array}} \left[\begin{array}{cccc|c} -2 & 0 & 2 & 1 & 0 \\ 2 & -6 & 0 & 0 & 0 \\ 2 & 3 & -6 & 3 & 0 \\ 2 & 3 & 0 & -6 & 0 \end{array}\right].$$

Derefter foretages følgende kæde af rækkeoperationer

$$\left[\begin{array}{cccc|c} -2 & 0 & 2 & 1 & 0 \\ 2 & -6 & 0 & 0 & 0 \\ 2 & 3 & -6 & 3 & 0 \\ 2 & 3 & 0 & -6 & 0 \end{array}\right] \xrightarrow[\mathbf{r}_1+\mathbf{r}_4\to\mathbf{r}_4]{\substack{\mathbf{r}_1+\mathbf{r}_2\to\mathbf{r}_2\\ \mathbf{r}_1+\mathbf{r}_3\to\mathbf{r}_3}} \left[\begin{array}{cccc|c} -2 & 0 & 2 & 1 & 0 \\ 0 & -6 & 2 & 1 & 0 \\ 0 & 3 & -4 & 4 & 0 \\ 0 & 3 & 2 & -5 & 0 \end{array}\right]$$

$$\xrightarrow[-\mathbf{r}_3+\mathbf{r}_4\to\mathbf{r}_4]{2\mathbf{r}_3+\mathbf{r}_2\to\mathbf{r}_2} \left[\begin{array}{cccc|c} -2 & 0 & 2 & 1 & 0 \\ 0 & 0 & -6 & 9 & 0 \\ 0 & 3 & -4 & 4 & 0 \\ 0 & 0 & 6 & -9 & 0 \end{array}\right]$$

$$\xrightarrow[\mathbf{r}_2\leftrightarrow\mathbf{r}_3]{\mathbf{r}_2+\mathbf{r}_4\to\mathbf{r}_4} \left[\begin{array}{cccc|c} -2 & 0 & 2 & 1 & 0 \\ 0 & 3 & -4 & 4 & 0 \\ 0 & 0 & -6 & 9 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\xrightarrow{-1/3\mathbf{r}_3\to\mathbf{r}_3} \left[\begin{array}{cccc|c} -2 & 0 & 2 & 1 & 0 \\ 0 & 3 & -4 & 4 & 0 \\ 0 & 0 & 2 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\xrightarrow[2\mathbf{r}_3+\mathbf{r}_2\to\mathbf{r}_2]{-\mathbf{r}_3+\mathbf{r}_1\to\mathbf{r}_1} \left[\begin{array}{cccc|c} -2 & 0 & 0 & 4 & 0 \\ 0 & 3 & 0 & -2 & 0 \\ 0 & 0 & 2 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right]$$

$$\xrightarrow[1/3\mathbf{r}_2\to\mathbf{r}_2]{\substack{-1/2\mathbf{r}_1\to\mathbf{r}_1\\ 1/2\mathbf{r}_3\to\mathbf{r}_3}} \left[\begin{array}{cccc|c} 1 & 0 & 0 & -2 & 0 \\ 0 & 1 & 0 & -2/3 & 0 \\ 0 & 0 & 1 & -3/2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

hvoraf resultatet aflæses.

Vi vil nu definere *linkmatricen* for et generelt web.

Definition 2 *Lad W være et web med n sider, som nummereres $1, 2, \ldots, n$. For $k \in \{1,\ldots,n\}$ sætter vi*

$$L_k = \{\textit{de sidenumre, der indeholder et link til side } k\}.$$

Lad N_j være det totale antal links som udgår fra side j. Vi bemærker, at $N_j > 0$ for alle $j \in L_k$, fordi der i hvert fald udgår et link til side k. Scoren x_k for side k udregnes da vha formlerne

$$x_k = \sum_{j\in L_k} \frac{x_j}{N_j}, \quad k = 1,\ldots,n.$$

Disse ligninger kan skrives på matrixfom som $\mathbf{Ax} = \mathbf{x}$, hvor $\mathbf{A}$ er en $n \times n$ matrix med elementerne

$$a_{ij} = \begin{cases} 1/N_j & \textit{hvis der er et link fra side } j \textit{ til side } i, \\ 0 & \textit{ellers.} \end{cases}$$

Denne matrix $\mathbf{A}$ kalder vi linkmatricen *for webbet.*

I Eksempel 1 var det muligt at løse ligningssystemet $\mathbf{Ax} = \mathbf{x}$. Det er ikke altid tilfældet og det skyldes tilstedeværelsen af "hængende" sider i et web: En "hængende" side i et web er en side uden links til andre sider. Hvis side j er en hængende side vil der ikke udgå nogen links fra siden og $N_j = 0$. Dermed vil matricen $\mathbf{A}$ have udelukkende 0'er i søjle j.

Betragter vi et web W uden hængende sider er alle tallene $N_1, \ldots, N_n$ positive, og der gælder, at summen af elementerne i hver af $\mathbf{A}$'s søjler er 1: faktisk indeholder den j'te søjle i $\mathbf{A}$ præcis N_j elementer hver med værdi $1/N_j$ og de resterende elementer er 0. Med andre ord er matricen $\mathbf{A}$ en overgangsmatrix, hvis webbet ikke har nogen hængende sider.

2 Egenværdier og -vektorer

Lad os vende tilbage til ligningen $\mathbf{Ax} = \mathbf{x}$ oven for. Denne ligning udtrykker, at den rangordning $\mathbf{x}$ vi søger faktisk er en egenvektor for matricen $\mathbf{A}$ hørende til egenværdien $\lambda = 1$. Vi var succesfulde: det var faktisk muligt at finde en sådan egenvektor! Det er ikke nogen tilfældighed: Matricen fra Eksempel 1 er en såkaldt *overgangsmatrix*.

Der gælder faktisk, at en vilkårlig overgangsmatrix $\mathbf{A}$ har $\lambda = 1$ som egenværdi, se Afsnit 4. Det tilsvarende egenrum E_1 er defineret ved

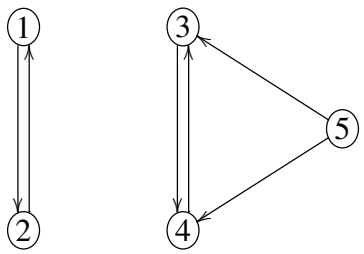
$$E_1 = E_1(\mathbf{A}) = \{\mathbf{x} \,|\, \mathbf{Ax} = \mathbf{x}\}.$$

I Eksempel 1 foretog vi en rangordning af websiderne via en ordning af elementerne i en (normaliseret) egenvektor efter størrelse. Der er i det generelle tilfælde en del problemer, som kan opstå, fx:

- Matricen har muligvis ikke $\lambda = 1$ som egenværdi (det kan være tilfældet hvis der er hængende sider).
- Måske er det ikke muligt naturligt at ordne egenvektorerne hørende til egenværdien $\lambda = 1$; dette sker i hvert fald hvis dimensionen af egenrummet E_1 er større end 1.
- Måske er der både positive og negative koordinater i en egenvektor hørende til egenværdien 1.

Eksempel 3 *Vi ser på webbet hvor links er givet som følger:*

Side 1 linker til side 2
Side 2 linker til side 1
Side 3 linker til side 4
Side 4 linker til side 3
Side 5 linker til side 3 og side 4



(Faktisk er der to dele af webbet som ikke linker til hinanden.) Den tilsvarende matrix er i dette tilfælde givet ved

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Denne matrix er også en overgangsmatrix så $\lambda = 1$ er en egenværdi. Udregninger viser, at egenrummet E_1 udspændes af de to lineært uafhængige vektorer

$$\mathbf{x}_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 0 \\ 1/2 \\ 1/2 \\ 0 \end{bmatrix}.$$

Se vi udelukkende på $\mathbf{x}_1$ kunne vi tro, at side 1 eller side 2 er de vigtigste, mens side 3 og side 4 kunne vær de vigtigste hvis vi alene ser på $\mathbf{x}_2$. Vi kunne også se på linearkombinationen

$$1/2\mathbf{x}_1 + 1/2\mathbf{x}_2 = \begin{bmatrix} 1/4 \\ 1/4 \\ 1/4 \\ 1/4 \\ 0 \end{bmatrix},$$

og nu ser hver af siderne 1, 2, 3 og 4 lige vigtige ud.

Eksempel 4 *Vi modificerer webbet fra Eksempel 1 idet side 3 nu ikke længere linker til side 1. I det tilfælde bliver linkmatricen*

$$\mathbf{A} = \begin{bmatrix} 0 & 0 & 0 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{bmatrix}$$

Planen er at bestemme en egenvektor hørende til egenværdien $\lambda = 1$. Men allerede nu er situationen anderledes: matricen $\mathbf{A}$ har ikke 1 som en egenværdi. Man kan vise, at matricen har følgende fire egenværdier (afrundede værdier):

$$\lambda_1 = 0.5614, \quad \lambda_2 = -0.2807 + 0.2640i, \quad \lambda_3 = -0.2807 - 0.2640i, \quad \lambda_4 = 0.$$

Vi ser også, at egenværdierne λ_2 og λ_3 ikke er reelle tal, men komplekse. Udregninger giver, at

$$\mathbf{x}_1 = \begin{bmatrix} 0.8907 \\ 0.5289 \\ 1.8907 \\ 1.0000 \end{bmatrix}$$

er en egenvektor hørende til egenværdien λ_1. Hvis man bruger en egenvektor hørende til den numerisk største egenværdi (her λ_1) for at rangordne websiderne er side 3 den vigtigste. Dette er ikke hensigtsmæssigt, idet siderne 3 og 4 optræder symmetrisk i webbet.

3 Entydig rangordning i web uden hængende sider

Vi så i Eksempel 3, at et web med to adskilte "delweb" leder til et egenrum E_1 af dimension større end 1, og dermed, at der ikke var nogen strategi for rangordningen. Dette problem kan omgås på en snedig måde, som beskrives neden for.

Lad W være et web med n sider, hvoraf ingen er hængende og lad $\mathbf{A}$ være den tilsvarende $n \times n$ linkmatrix. Vi lader $\mathbf{S}$ betegne den $n \times n$ matrix, hvor alle elementer er lig med $1/n$. Denne matrix er en overgangsmatrix, så $\lambda = 1$ er en egenværdi. Man kan vise, at det tilsvarende egenrum E_1 er 1-dimensionalt; faktisk gælder

$$E_1 = \text{span}\left\{ \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right\}. \tag{1}$$

Ideen er nu at arbejde med en kombination af $\mathbf{A}$ og $\mathbf{S}$: Vi definerer

$$\mathbf{M} = (1-m)\mathbf{A} + m\mathbf{S},$$

hvor $0 \leq m \leq 1$. (Man kan tænke på $\mathbf{M}$ som et vægtet gennemsnit af $\mathbf{A}$ og $\mathbf{S}$.) Matricen $\mathbf{M}$ er igen en overgangsmatrix (da $\mathbf{A}$ og $\mathbf{S}$ begge er det) og det viser sig, at egenrummet for $\mathbf{M}$ hørende til $\lambda = 1$ er 1-dimensionalt, for $m \in (0,1]$, og at man finde en egenvektor med udelukkende *positive* koordinater. Se Afsnit 4. Rangordningen af siderne kan dermed findes via numerisk ordning af en (normaliseret) egenvektor.

Hvis $m = 0$ er $\mathbf{M}$ lig med $\mathbf{A}$, mens $\mathbf{M}$ lig med $\mathbf{S}$ for $m = 1$. Matricen $\mathbf{S}$ svarer i øvrigt til et ganske særligt web: et web hvor hver side refererer præcis 1 gang til alle andre sider. I et sådant web er alle sider lige vigtige, og det reflekteres i udseendet af egenvektoren i (1).

Eksempel 5 *Vi betragter linkmatricen* $\mathbf{A}$ *fra Eksempel 1 og lader* $m = 0.15$ *(det har Google faktisk gjort). Det vægtede gennemsnit* $\mathbf{M}$ *af* $\mathbf{A}$ *og* $\mathbf{S}$ *er givet ved (afrundede tal)*

$$\mathbf{M} = \begin{bmatrix} 0.0375 & 0.0375 & 0.8875 & 0.4625 \\ 0.3208 & 0.0375 & 0.0375 & 0.0375 \\ 0.3208 & 0.4625 & 0.0375 & 0.4625 \\ 0.3208 & 0.4625 & 0.0375 & 0.0375 \end{bmatrix}.$$

I dette tilfælde kan man vise, at egenrummet hørende til egenværdien $\lambda = 1$ *lig med*

$$\textit{span}\left\{ \begin{bmatrix} 0.368 \\ 0.142 \\ 0.288 \\ 0.202 \end{bmatrix} \right\}.$$

Igen er side 1 den vigtigste.

Ser vi på matricen fra Eksempel 3 (med de to adskilte delweb) er det vægtede gennemsnit (for $m = 0.15$*) givet ved*

$$\mathbf{M} = \begin{bmatrix} 0.03 & 0.88 & 0.03 & 0.03 & 0.03 \\ 0.88 & 0.03 & 0.03 & 0.03 & 0.03 \\ 0.03 & 0.03 & 0.03 & 0.88 & 0.455 \\ 0.03 & 0.03 & 0.88 & 0.03 & 0.455 \\ 0.03 & 0.03 & 0.03 & 0.03 & 0.03 \end{bmatrix}.$$

I dette tilfælde kan man vise, at egenrummet hørende til egenværdien $\lambda = 1$ lig med

$$span\left\{\begin{bmatrix} 0.2 \\ 0.2 \\ 0.285 \\ 0.285 \\ 0.03 \end{bmatrix}\right\}.$$

Nu har vi mulighed for at sammenligne de to adskilte delweb i samme ramme, og side 3 og 4 er de vigtigste.

4 Positive matricer og bestemmelse af egenvektorer ved iteration

Eksempel 1 viser, at det er en relativt krævende opgave at bestemme selv egenvektorer med 4 koordinater ved at løse ligningssystemer. I dette afsnit vil vi give en forsmag på hvordan egenvektorer kan bestemmes uden at løse ligningssystemer. Der er hel teori om bestemmelse af "den numerisk største" egenværdi og tilhørende egenvektorer ved iteration, men her vil vi kun nævne resultater for overgangsatricer.

Definition 6 *En matrix er en overgangsmatrix, hvis alle elementer i matricen er ikke-negative, og hvis summen af alle elementerne i hver søjle er lig med 1.*

Følgende sætning viser, at positive overgangsmatricer altid har "positive" egenvektorer.

Sætning 7 *Lad* $\mathbf{M}$ *være en overgangsmatrix hvor alle elementer er strengt positive. Da er* $\lambda = 1$ *en egenværdi for* $\mathbf{M}$*, og der findes en entydigt bestemt tilhørende egenvektor* $\mathbf{q} = (q_1, \ldots, q_n)^\top$*, som opfylder* $\sum_{j=1}^n q_j = 1$ *og* $q_j > 0$ *for alle* j*. Endvidere gælder*

$$\lim_{k\to\infty} \mathbf{M}^k \mathbf{x}_0 = \mathbf{q},$$

hvor $\mathbf{x}_0 = (x_1, \ldots, x_n)^\top$ *er en vilkårlig vektor, som opfylder* $\sum_{j=1}^n x_j = 1$.

Lad nu $\mathbf{A}$ være en linkmatrix for et web uden hængende sider, og lad som i Afsnit 3

$$\mathbf{M} = (1-m)\mathbf{A} + m\mathbf{S},$$

hvor $0 < m < 1$. Vi har tidligere bemærket, at $\mathbf{M}$ er en positiv overgangsmatrix. Sætningen ovenfor giver så, at egenvektoren $\mathbf{q}$ kan bestemmes ved grænseværdien $\lim_{k\to\infty} \mathbf{M}^k \mathbf{x}_0$, eller rekursivt ved $\mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k$. Udregningen $\mathbf{M}\mathbf{x}_k$ involverer typisk omkring n^2 multiplikationer og additioner (fordi alle elementer i $\mathbf{M}$ er positive), hvilket er en del allerede for $n = 5$, men uoverskueligt mange for

$$n = \text{alverdenswebsider}.$$

Der er dog en væsentlig beregningsmæssig forenkling, som går ud på følgende: hvis der gælder $\mathbf{x}_k = (x_{k,1}, \ldots, x_{k,n})^\top$, hvor $\sum_{j=1}^n x_{k,j} = 1$, så har vi

$$\begin{aligned} \mathbf{x}_{k+1} = \mathbf{M}\mathbf{x}_k &= (1-m)\mathbf{A}\mathbf{x}_k + m\mathbf{S}\mathbf{x}_k \\ &= (1-m)\mathbf{A}\mathbf{x}_k + m\mathbf{s}. \end{aligned}$$

Her har vi brugt, at summen af koordinaterne i $\mathbf{x}_k$ er lig med 1, sådan, at $\mathbf{S}\mathbf{x}_k = (1/n, \ldots, 1/n)^\top = \mathbf{s}$. Beregningen af $\mathbf{M}\mathbf{x}_k$ svarer til beregningen af $\mathbf{A}\mathbf{x}_k$, og denne er langt enklere, idet linkmatricen $\mathbf{A}$ "næsten" udelukkende består af nuller.

Litteratur

[1] Kurt Bryan and Tanya Leise, *The $25,000,000,000 eigenvector, The Linear Algebra Behind Google*, SIAM Rev., 48(3), 569–581. `http://www.rose-hulman.edu/~bryan/google.html`

Henrik Holm (holm@math.ku.dk)
Henrik Laurberg Pedersen (henrikp@math.ku.dk)