Filosofien og matematikken bag Google

– Med fokus på PageRank

Jakob Lindblad Blaavand, Oxford University

Indledning

En internetsøgemaskine er god, hvis den først og fremmest kan søge blandt al information på internettet. Derudover skal den være hurtig til at finde resultaterne, og måske vigtigst af alt skal den vise de mest relevante resultater først. Sat lidt på spidsen er de to første krav blot et spørgsmål om nok computerkraft. Ordningen af søgeresultaterne er mere subtil og var det, der gjorde Google så speciel i forhold til andre søgemaskiner, da Sergey Brin og Lawrence Page lancerede Google tilbage i 1997. Hvis jeg på Google søger på *matematik* og *København*, får jeg 554.000 resultater, og de kommer svimlende hurtigt – efter kun 0.23 sekunder. Når jeg søger på disse to ord, vil jeg naturligvis have fat i *Institut for Matematiske Fag* på *Københavns Universitets* hjemmeside. Google giver mig også denne hjemmeside som det første søgeresultat. Men hvordan kan Google gætte det fra blot *matematik* og *København*? Dette mindre mysterium skal vi undersøge i denne artikel.

Søgeresultaterne skal sorteres efter relevans – men hvilke parametre ligger til grund for dette kriterie, og hvordan implementeres de? Det ville være mest demokratisk, hvis man kunne sætte en stor gruppe mennesker til at bestemme, hvilke hjemmesider der er de mest troværdige og vigtigste. Man kunne så sortere søgeresultaterne efter deres placering på denne universelle liste. Der er mange problemer med denne fremgangsmåde. Først og fremmest vil listen naturligvis være afhængig af gruppens sammensætning. Derudover er det en uoverskuelig tidshorisont. Hvis mennesker skal tage stilling til vigtigheden og troværdigheden af 255 millio-

ner websites, er en sådan liste altså underlagt en subjektiv vurdering, som er uhensigtsmæssig. Løsningen er at få internettet til selv at ordne siderne efter vigtighed.

Størstedelen af denne artikel er en forklaring af filosofien bag Google, krydret med en smule matematik. Vi vil kun se på den afgørende ide, der fik Google til at være så meget bedre end sine daværende konkurrenter, og ikke diskutere de mange andre tiltag Google i dag bruger til at ordne søgeresultater. I afsnittet 'Matematiske beviser' ser vi på den konkrete matematik. Her bliver alle påstande fra artiklen bevist. Beviserne er alene baseret på lineær algebra, og den eneste forudsætning for at kunne forstå beviserne er at kende til egenværdier og egenvektorer for en matrix. Hovedteksten er stærkt inspireret af [1], mens det matematiske afsnit er baseret på [2]. Hvis du vil vide mere om lineær algebra, er [3] et godt udgangspunkt.

Hvilke sider er vigtigst?

Hvis du har lavet en hjemmeside, så har du også lavet links til andre hjemmesider. Det har du gjort, fordi du synes de hjemmesider, du linker til, indeholder vigtig information af den ene eller den anden karakter. Når du derfor laver et link til en hjemmeside, siger du samtidig, at du mener, siden er vigtig. Hvis vi kan kortlægge hele internettets linkstruktur, så kan vi få alle hjemmesideforfatteres mening om, hvilke sider der er vigtige. På den måde kan vi forfølge vores første demokratiske tanke om, at en stor gruppe mennesker skal være med til at bestemme, hvad der er vigtigt: Hjemmesidens placering på vigtighedslisten er bestemt af *hvor mange* og *hvilke* hjemmesider, der linker til den. Hvis listen også skal afspejle en form for troværdighed, er det også nødvendigt at tage i betragtning, hvilke sider der linker til hvilke. F.eks. er det

meget mere troværdigt for en hjemmeside, at en stor hjemmeside som *jp.dk* linker til den, end, at jeg som privatperson linker til hjemmesiden.

Altså, hvis vi har en side P, lader vi et tal, sidens *PageRank*, $I(P)$, være en samlet beskrivelse af sidens troværdighed og vigtighed. Når søgeresultaterne i en søgning skal vises, bliver hjemmesiderne sorteret efter denne PageRank. Men hvordan skal vi bestemme PageRank fra de ovenstående betragtninger? Antag, at siden P_j har l_j links. Hvis et af disse links peger på siden P_i, så overfører P_j en $1/l_j$'te del af dens PageRank til P_i. Så PageRanken af P_i, $I(P_i)$, er summen af bidrag fra alle siderne, der linker til P_i. Altså, hvis mængden af sider, der linker til P_i, betegnes B_i, er PageRanken givet ved

$$I(P_i) = \sum_{P_j \in B_i} \frac{I(P_j)}{l_j}. \tag{1}$$

Har vi ikke et problem? En sides PageRank bestemmes af andre siders PageRank. Men hvis vi nu har to sider, der linker til hinanden - hvordan får vi så bestemt hver deres PageRank? Er PageRank virkelig veldefineret? Det er som spørgsmålet om, hvad der kom først: hønen eller ægget? Skal vi så til at skrotte vores håb om at lave en vægtning af sider efter alle internettets links? Vi har endnu ikke brugt noget matematik på vores situation, og hvis vi gør det, så har vi heldigvis en løsning.

Hyperlinkmatricen

Vi definerer nu en matrix $H = (H_{ij})$, som kaldes *hyperlinkmatricen*. Hver række og søjle repræsenterer en webside. Hvis den i'te række er websiden P_i så er den i'te søjle det også. Altså er

H en kvadratisk matrix. I den ij'te indgang står der $1/l_j$ hvis $P_j \in B_i$ og 0 ellers. Dvs. hvis P_j linker til P_i står der $1/l_j$ i den ij'te indgang og 0 hvis P_j ikke linker til P_i. Hvis man ser på den j'te *søjle*, så viser denne vektor, hvilke sider der linkes til fra P_j. Hvis man derimod ser på den j'te *række* kan man aflæse hvilke sider, der linker til P_j. Det skal bemærkes, at hvis en hjemmeside ikke linker til nogen andre hjemmesider, så er alle indgange i den tilhørende søjle 0.

Hyperlinkmatricen er altså helt speciel i den forstand, at alle indgangene er større end eller lig nul, og summen af alle indgangene i en søjle er 1, medmindre den side søjlen repræsenterer ikke har nogen links.

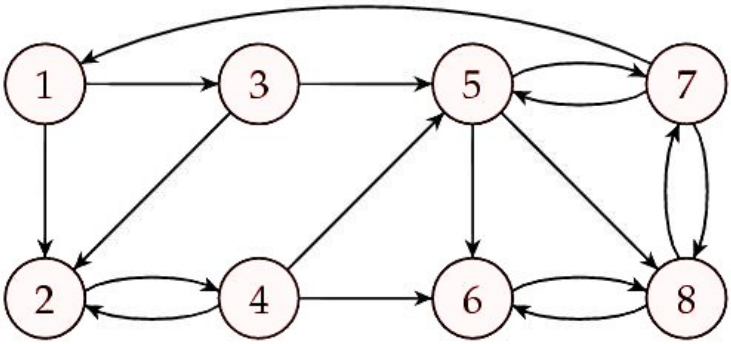
Vi kan nu lave vektoren $I = [I(P_i)]$, hvis indgange er PageRanks. Husker vi tilbage på definition (1) af en sides PageRank, så har vi, at vores PageRank-vektor opfylder følgende ligning:

$$I = HI,$$

og vi kan tage dette som værende definitionen på I. I virkeligheden har vi her n ligninger med n ubekendte, hvor n er antallet af websider på nettet. Altså *rigtig* mange ligninger, der skal løses for at finde alle websiders PageRank. Per definition er I en egenvektor for H med egenværdi 1, og den slags vektorer er helt specielle. Der findes teknikker til at finde dem, og forhåbentlig bliver det overskueligt at løse ligningerne.

Miniatureweb

Lad os se på et eksempel, hvor vi har otte websider, der linker til hinanden og udgør et internet i legetøjsstørrelse. Links er repræsenteret ved pile.



Den tilhørende matrix og egenvektor er

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 0 \\ 1/2 & 0 & 1/2 & 1/3 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1/3 & 1 & 1/3 & 0 \end{bmatrix} \quad \text{og} \quad I = \begin{bmatrix} 0.0600 \\ 0.0675 \\ 0.0300 \\ 0.0675 \\ 0.0975 \\ 0.2025 \\ 0.1800 \\ 0.2950 \end{bmatrix}$$

Da side 1 linker til side 2 og 3 står der $\frac{1}{2}$ på plads 2 og 3 i første søjle og 0'er på resten af pladserne. Side 2 linker kun til side 4, og derfor står der 1 på plads 4 i anden søjle og 0'er på resten af pladserne, osv.

I dette tilfælde viser PageRank-vektoren, I, at side 8 er den vigtigste, fordi det største tal står på plads 8. Det kan til dels forklares med, at der er tre sider, der linker til side 8. Der bliver

også linket til side 2, 5 og 6 tre gange, men 8 er den vigtigste, fordi de websider, der peger på 8, selv har mange sider, der peger på dem. Her er det altså igen troværdigheden af siderne, der er med til afgøre PageRanken, og ikke alene antallet af links.

Beregning af I

Der findes mange måder at finde egenvektorer til kvadratiske matricer på. Men når vi i dette tilfælde har en hyperlinkmatrix med 50 milliarder rækker og 50 milliarder søjler, virker alle beregningsmetoder håbløse og tidskrævende.[10] Dog er vi så heldige, at der i gennemsnit kun er 10 links per side, så langt størstedelen af indgangene i H er nul. Derfor bruger man det, der hedder *potensmetoden*, til at finde egenvektoren I til egenværdien 1.

For at bruge potensmetoden skal man vælge sig en vektor I^0, som man mener kunne være en kandidat til I, og så laver man følgen af vektorer I^k givet ved

$$I^{k+1} = HI^k = H^k I^0.$$

Hvis H er en særlig pæn matrix, vil I^k vektorerne nærme sig egenvektoren I, når k bliver stor. Selv for Googles store matrix skal man kun op på ca. $k = 60$ for at få en god approksimation til I.

I eksemplet er de første elementer i følgen beregnet for $k = 60$ og $k = 61$.

[10] Det tidligere tal 255 millioner er antallet af websites, mens de 50 milliarder er det totale antal, hvor undersider tælles med.

I^0	I^1	I^2	I^3	I^4	$\cdots$	I^{60}	I^{61}
1	0	0	0	0.0278	$\cdots$	0.06	0.06
0	0.5	0.25	0.1667	0.0833	$\cdots$	0.0675	0.0675
0	0.5	0	0	0	$\cdots$	0.03	0.03
0	0	0.5	0.25	0.1667	$\cdots$	0.0675	0.0675
0	0	0.25	0.1667	0.1111	$\cdots$	0.0975	0.0975
0	0	0	0.25	0.1806	$\cdots$	0.2025	0.2025
0	0	0	0.0833	0.0972	$\cdots$	0.18	0.18
0	0	0	0.0833	0.3333	$\cdots$	0.295	0.295

Vigtige spørgsmål

Nu har vi altså fundet en relativt effektiv måde at beregne I på. Med denne måde vælger vi os et startpunkt for en følge, der så tilnærmer sig egenvektoren. Men på dette tidspunkt skal vi stille os selv tre centrale spørgsmål:

- Vil følgen I^k altid konvergere?
- Er grænsevektoren af I^k uafhængig af startvektoren I^0?
- Indeholder I rent faktisk den informationen, vi vil have?

Svaret på alle tre ovenstående spørgsmål er desværre nej. Men det, der redder os, er, at svaret på tredje spørgsmål er nej. Vi har nemlig ikke helt fundet frem til den information, som vi satte os for at finde – men vi er meget tæt på.

Googlematricen

Tidligere har vi opfattet PageRank som et mål for, hvor vigtig en side er, beregnet ud fra vigtigheden af de sider, der peger på den. Det var et forsøg på at bruge internettets linkstruktur til på demokratisk vis at afgøre hvilke sider, der er de vigtigste og

mest troværdige. Vi har dog kun taget højde for hjemmesidernes fprfattere. Hvis vi skulle være helt demokratiske, skulle vi også have brugerne af internettet med – altså surferne. Selvfølgelig er der et stort overlap mellem forfattere og brugere, men selvom man synes, at en bestemt hjemmeside er vigtig, er det jo ikke sikkert, man vil linke til den, hvis nu ens hjemmeside har et meget konkret formål.

Vi vil nu tage hyperlinkmatricen som udgangspunkt og ændre den en smule, så vi får en matrix, hvor vi kan svare ja på alle spørgsmålene fra ovenstående, og hvor vi altså får indkodet websurfernes stemme i afgørelsen af, hvilke hjemmesider er vigtige. Vores udgangspunkt er, at vi ud fra H vil konstruere en matrix G, og løse ligningssystemet $GI = I$. Dette I er vores rigtige PageRank-vektor.

Vi forestiller os, at vi tager en tilfældig surftur på nettet. Vi starter på siden P_j, som har l_j links. Vi vælger så et tilfældigt af de l_j links. Et af linksene er til siden P_i. Dermed er sandsynligheden for at vi rammer P_i, når vi står på P_j, altså $1/l_j$.

Forfølger vi denne tanke, så kan PageRank $I(P_j)$ ses som sandsynligheden for, at man tilfældigvis surfer forbi siden, hvis man bare klikker rundt på må og få på nettet. Det giver ganske god mening, for hvis du surfer efter noget bestemt information, så vil du uværgeligt ende på de samme sider flere gange. Altså er de sider vigtigere end andre, og disse siders PageRank skal være højere. Dog giver denne fortolkning af PageRank os et problem: Hvis vi på vores surftur støder på en side, der ikke linker til andre websider, hvad gør vi så? For at kunne fortsætte forestiller vi os, at en side uden links til andre sider rent faktisk linker til alle sider på hele nettet. Hvis vi tænker tilbage på vores hyperlinkmatrix, så betyder det, at den søjle, der repræsenterede en side uden links, ville have lutter 0'er. Nu bliver hele denne søjle erstattet

med en vektor med $1/n$ på hver plads, hvor n er antallet af sider på nettet. Denne nye matrix kalder vi S. Det betyder, at vi kan skrive S som $S = H + A$, hvor A er en matrix, med lutter 0'er - bortset fra de søjler, der repræsenterer websider uden links, hvor der i stedet står $1/n$ i hver indgang.

Har vi nu fået simuleret, hvordan en websurfer opfører sig på nettet? I det store hele, ja, fordi man følger links på de sider, man besøger, og hvis der ikke er nogen links, så vælges en tilfældig af nettets mange websider. Men på en surftur vælger man ofte en anden side på nettet fremfor bare at vælge en af de sider, der linkes til. Vi kan formulere dette matematisk ved at vælge et tal, α, mellem 0 og 1, som er sandsynligheden for, at websurferen gør, som vi har forudsagt med matricen S: han vælger en af de sider, der linkes til, eller hvis han kommer til en side uden links, så vælges en tilfældig af nettets mange sider. Dermed er der sandsynlighed $1 - \alpha$ for, at websurferen gør noget andet: vælger en tilfældig af nettets websider. Det hele samler vi i *Googlematricen*:

$$G = \alpha S + (1 - \alpha)\frac{1}{n}1$$

hvor 1 er en matrix af samme dimension som S, med lutter 1-taller i alle indgangene.

Variablen α er ret vigtig, for den styrer hvor stor indfyldelse, internettets hyperlinkstruktur skal have i Googlematricen. F.eks. vil $\alpha = 1$ give den oprindelige hyperlinkstruktur, og $\alpha = 0$ giver en webstruktur helt uden links.

Googlematricen giver den hidtil mest realistiske beskrivelse af en websurfers adfærd på nettet. Det er naturligvis afhængigt af, at vi finder en rimelig værdi til α. Google bruger $\alpha = 0.85$, der betyder, at der er 85% sandsynlighed for, at en websurfer følger et

link på en hjemmeside, og 15% sandsynlighed for, at han vælger en tilfældig hjemmeside.

Beregning af I

Så langt så godt. Vi har fået lavet os en matrix G, der kan simulere en websurfers adfærd. Hvis vi nu kan finde egenvektoren I med egenværdien 1, altså finde en vektor så $GI = I$, så har vi fundet vores PageRank-vektor. I afsnittet 'Matematiske beviser' ser vi, at der findes uendeligt mange løsninger til $GI = I$. Men hvis vi vil fortolke PageRank som en sandsynlighed for at en hjemmeside bliver besøgt på en tilfældig surftur, så skal PageRank jo være et positivt tal, og hvis vi lægger PageRanks for alle hjemmesider sammen, skal vi have 1. Med disse ekstra antagelser kan vi vise, at der findes præcist en løsning til $GI = I$, og denne løsning kan findes med potensmetoden. For at potensmetoden virker, skal vi vise, at følgen I^k vil konvergere mod vores PageRank-vektor I, der opfylder $GI = I$ og den ekstra antagelse. Derudover skal vi vise, at grænsevektoren for I^k ikke afhænger af valget af startvektor. Det vil blive gjort i afsnittet 'Matematiske beviser'.

Som sagt vil vi bruge potensmetoden, og husker vi på, at $S = H + A$, bliver

$$G = \alpha H + \alpha A + \frac{1-\alpha}{n}1,$$

og dermed er

$$GI^k = \alpha H I^k + \alpha A I^k + \frac{1-\alpha}{n}1I^k.$$

Da de fleste af indgangene i H er nul, skal der i gennemsnit kun summeres 10 tal i hver af indgangene i produktet HI^k. Desuden er alle rækker i A ens, så AI^k er en vektor med samme tal i hver

indgang, og prikproduktet mellem I^k og en række i A skal kun beregnes en enkelt gang. Det samme er gældende for 1, der også har ens rækker.

Hastigheden af I^k's konvergens afhænger af størrelsen af α. Konvergensen er hurtig, hvis α er lille, og langsom, når den er tæt på 1. Med valget af $\alpha = 0.85$ har Google indgået et kompromis mellem at få så meget som muligt af internettets hyperlinkstruktur med, og hastigheden hvormed I kan beregnes. Det viser sig, at k skal ligge mellem 50 og 100, for at vi kan få tilpas god approksimation til I. Google siger selv, at det tager dem et par dage at beregne I.[11]

I og med, at nettet er en dynamisk størrelse, hvor der hele tiden bliver tilføjet og slettet indhold, vil en PageRank vektor være forældet sekundet efter, at beregningen af den er startet. Rygterne vil derfor vide, at Google for nogle år siden beregnede en ny PageRank-vektor en gang i måneden. I dag er det ikke klart, hvordan det fungerer.

Matematiske beviser

I dette afsnit vil vi bevise påstandene omkring potensmetoden og eksistensen af en PageRank-vektor. Først og fremmest det mest essentielle spørgsmål: Findes der overhovedet en løsning til $GI = I$, der samtidig opfylder, at $\sum_{i=1}^{n} I(P_i) = 1$ og $I(P_i) > 0$? Dette er et ligningssytem med $n + 1$ ligninger og n ubekendte. Som udgangspunkt er det ikke sikkert, at vi kan finde en sådan løsning. I tilfælde af, at vi kan finde en løsning, er vi så sikre på, at den

[11] Disse oplysninger er nogle år gamle. Det er derfor meget usikkert, hvad de korrekte er i dag, og om det stadig er denne metode, der bruges. Men det er forretningshemmeligheder i dag.

er entydig? Desuden skal vi se, at potensmetoden virker og er uafhængig af valget af startvektor.

I det følgende vil jeg bruge betegnelsen $||x||_1 = \sum_{i=1}^{n} |x_i|$ for summen af den numeriske værdi af indgangene i en vektor. $||x||_1$ kaldes for 1-længden af x.

Proposition 1 *Hvis $G = (g_{ij})$ er en matrix med $g_{ij} > 0$ for alle i, j og alle søjler summer til 1, da er 1 en egenværdi for G og ligningen $Gx = x$ har en løsning, hvor alle indgange i x er positive, særligt findes en løsning med $||x||_1 = \sum_{i=1}^{n} x_i = 1$.*

Bevis. Rækkerne i G^T summer alle til 1, så vektoren v givet ved $(1, 1, \ldots, 1)^T$ er en egenvektor for G^T med egenværdien 1, så $G^T v = v$. Dermed er 1 en rod i $\det(G^T - \lambda Id)$, og da $\det(G^T - \lambda Id) = \det(G - \lambda Id)$ er 1 altså også en egenværdi for G.

Helt generelt gælder der, at $|\sum_i y_i| \leq \sum_i |y_i|$ og hvis y_i'erne har blandede fortegn er uligheden skarp.

Lad $x \in V_1(G)$ være en egenvektor i egenrummet tilhørende egenværdien 1 for G. Antag, at der er forskellige fortegn i x's indgange.

Da $x = Gx$ er $x_i = \sum_{j=1}^{n} g_{ij} x_j$ og da x_i'erne har blandede fortegn $g_{ij} > 0$, har $g_{ij} x_j$'erne blandede fortegn. Vi har altså en streng ulighed $|x_i| < \sum_{j=1}^{n} g_{ij} |x_j|$, og derfor har vi

$$\sum_{i=1}^{n} |x_i| < \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} |x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} g_{ij} = \sum_{j=1}^{n} |x_j|,$$

hvilket er en modstrid.

Altså har alle indgange i x samme fortegn. Særligt findes en vektor i $V_1(G)$ med alle indgangene positive, og dermed også en med 1-længde 1. □

Ovenstående proposition viser altså eksistensen af en løsning til ligningssystemet $GI = I$, der opfylder $||I||_1 = 1$ og alle indgange i I er positive. Hvis $\dim V_1(G) = 1$ findes der kun én løsning. Denne løsning er vores PageRank-vektor.

For at kunne vise at $\dim V_1(G) = 1$ skal vi først have vist følgende generelle lemma.

Lemma 2 *Lad v og w være lineært uafhængige vektorer i $\mathbb{R}^m$, $m \geq 2$. Der findes reelle tal s, t så $x = sv + tw$ har både positive og negative indgange.*

Bevis. Da v, w er lineært uafhængige er $v \neq 0 \neq w$. Lad $d = \sum v_i$. Hvis $d = 0$ indeholder v forskellige fortegn og $s = 1$ og $t = 0$ opfylder det ønskede.

Hvis $d \neq 0$ defineres $s = \frac{-\sum_i w_i}{d}$ og $t = 1$. Da v, w er lineært uafhængige er $x = sv + tw \neq 0$, men det er samtidig klart at $\sum_i x_i = 0$, så x_i'erne må have blandede fortegn. □

Nu er vi i stand til at bevise entydigheden af PageRank-vektoren.

Proposition 3 *Hvis $G = (g_{ij})$ er en kvadratisk matrix med $g_{ij} > 0$ og $\sum_i g_{ij} = 1$ for alle j er $\dim V_1(G) = 1$.*

Bevis. Antag, at der findes to lineært uafhængige vektorer $v, w \in V_1(G)$. Pr. lemma 2 findes s, t så komponenterne af $x = sv + tw$ har blandede fortegn. Men pr. Proposition 1 vil ethvert $x \in V_1(G)$ have enten kun positive eller kun negative komponenter, hvilket er en modstrid.

Altså vil en basis for $V_1(G)$ kun bestå af en enkelt vektor, og $\dim V_1(G) = 1$. □

Vi har nu bevist, at der findes en entydig PageRank-vektor. Spørgsmålet er nu blot, om vi kan finde den med potensmetoden beskrevet ovenfor.

Proposition 4 *Lad $G = (g_{ij})$ være en kvadratisk matrix med $g_{ij} > 0$ og $\sum_{i=1}^{n} g_{ij} = 1$ for alle j, og lad V være underrummet i $\mathbb{R}^n$ bestående af vektorer v så $\sum_i v_i = 0$. Da er G matrixrepræsentationen af en lineær transformation $G : V \to V$ og der findes et $0 \leq c < 1$ så $||Gv||_1 \leq c||v||_1$ for alle $v \in V$.*

Bevis. Lad os først se, at G tager elementer i V til elementer i V. Lad $w = Gv$, så $w_i = \sum_{j=1}^{n} g_{ij} v_j$ og

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \sum_{j=1}^{n} g_{ij} v_j = \sum_{j=1}^{n} v_j \sum_{i=1}^{n} g_{ij} = \sum_{j=1}^{n} v_j = 0.$$

Dermed er $w \in V$ og $G : V \to V$.

Nu til vurderingen, som er en smule besværlig at vise.

$$||w||_1 = \sum_{i=1}^{n} e_i w_i = \sum_{i=1}^{n} e_i \sum_{j=1}^{n} g_{ij} v_j,$$

hvor $e_i = \operatorname{sgn}(w_i)$ og e_i'erne er ikke alle ens, da $\sum_i w_i = 0$, og $w \in V$ – medmindre $w = 0$, hvor uligheden klart gælder.

$$||w||_1 = \sum_{j=1}^{n} v_j \sum_{i=1}^{n} e_i g_{ij} = \sum_{j=1}^{n} a_j v_j,$$

hvor $a_j = \sum_{i=1}^{n} e_i g_{ij}$.

Da e_i'erne ikke alle er ens, $\sum_{i=1}^{n} g_{ij} = 1$ og $0 < g_{ij} < 1$, er det klart, at

$$-1 < -1 + \min_i g_{ij} \leq a_j \leq 1 - \min_i g_{ij} < 1.$$

Vi har altså, at $|a_j| \leq |1 - \min_i g_{ij}| < 1$. Lad derfor $c := \max_j |1 - \min_i g_{ij}|$ for så er $|a_j| \leq c < 1$ for alle j. Dermed har vi nu, at

$$||w||_1 = \sum_{j=1}^{n} a_j v_j = |\sum_{j=1}^{n} a_j v_j| \leq \sum_{j=1}^{n} |a_j||v_j| \leq c \sum_{j=1}^{n} |v_j| = c||v||_1,$$

hvormed det ønskede er vist. □

Vi kan nu afslutte dette matematiske afsnit med en sætning, der indeholder svar på alle spørgsmål stillet under afsnittet *Vigtige Spørgsmål*.

Theorem 5 *Enhver kvadratisk matrix $G = (g_{ij})$ med $0 < g_{ij} < 1$ og $\sum_{i=1}^{n} g_{ij} = 1$ for alle j, har en entydig egenvektor I tilhørende egenværdien 1, der yderligere kun har positive indgange og $||I||_1 = 1$. Vektoren I kan beregnes ved $I = \lim_{k\to\infty} G^k x_0$, hvor x_0 er en vektor med positive indgange og $||x_0||_1 = 1$.*

Bevis. Vi ved allerede fra de ovenstående propositioner, at G har 1 som egenværdi, og at $\dim V_1(G) = 1$. Det gav os som ønsket, at I eksisterer og er entydig. Vi mangler blot at bevise, at potensmetoden virker. Det vil sige at følgen $G^k x_0$ konvergerer mod I for et vilkårligt valg af x_0 med ovenstående egenskaber.

Lad $x_0 \in \mathbb{R}^n$ have positive indgange og $||x_0||_1 = 1$. Vi ved som sagt, at I findes, at $I_i = I(P_i) > 0$ og $\sum_i I(P_i) = 1$.

V er underrummet af $\mathbb{R}^n$, hvor indgangene summerer til 0. Definer $v = x_0 - I$. Dermed er $v \in V$, da summen af v's indgange er nul, fordi summen af indgangene i både x_0 og i I er 1. Derfor er $x_0 = I + v$, og $G^k x_0 = G^k I + G^k v = I + G^k v$. altså er $G^k x_0 - I = G^k v$, og et induktionsargument giver nu, at

$$||G^k v||_1 \leq c^k ||v||_1,$$

hvor $0 \leq c < 1$. Samlet set har vi, at $\lim_{k\to\infty} ||G^k v||_1 = 0$, hvorfor $G^k x_0 \to I$ for $k \to \infty$, hvormed det ønskede er vist. □

Opsamling

Da Page og Brin startede Google i 1997, blev internettet forvandlet fra at være en bunke ustrukturerede informationer, som ingen kunne finde rundt i, til at blive en – ikke fuldstændig ordnet – bunke informationer. Men det var blevet meget lettere at finde relevante informationer hurtigt.

Hovedidéen var at få internettet til selv at ordne informationerne efter relevans ved hjælp af dets links. Som det er vist ovenfor er idéen simpel, men meget anvendelig. Resultatet af en søgning bliver bl.a. sorteret efter sidernes PageRank fundet i PageRank-vektoren. Google siger selv, at PageRank er et af mere end 200 kriterier, der bliver sorteret efter. De resterende kriterier er forretningshemmeligheder, som Google ikke siger noget om, ganske som Google ikke offentliggør, hvad en hjemmesides PageRank præcist er. I den seneste tid har Google lanceret et af de mest radikale ekstra sorteringskriterier. Ved at gemme din søgehistorik lærer Google dig og dine præferencer at kende. Dermed kan Google sortere søgeresultater så de er specielt tilpasset dig. Det er ganske smart når matematikartikler for mig får højere prioritet. Søger man f.eks. på 'Hitchin map flat' får man som udgangspunkt et kort over byen Hitchin i Hertfordshire i England, og en masse tilbud om ledige lejligheder. Laver man geometri som jeg, håber man dog at få artikler om Higgs bundter og egenskaber af Hitchin afbildningen. Alt har dog en pris, nemlig at Google ved forfærdeligt meget om dig. Du må så gøre op med dig selv, om de bedre

søgeresultater er det værd – man kan dog slå overvågningen fra, så man stadig kan bruge Google anonymt – heldigvis.

I kølvandet på PageRank-algoritmen er der udviklet andre algoritmer, som også bruger internettets hyperlinkstruktur til at vurdere websiders vigtighed. Et eksempel er HITS-algoritmen, som blev lavet af Jon Kleinberg, der ligger til grund for Teoma søgemaskinen, der driver ask.com. Du kan selv vurdere, hvilken der er bedst ved at sammenligne resultater.

Litteratur

[1] David Austin. How google finds your needle in the web's haystack. http://www.ams.org/samplings/feature-column/fcarc-pagerank.

[2] Kurt Bryan and Tanya Liese. The linear algebra behind google. http://www.rose-hulman.edu/~bryan/googleFinalVersionFixed.pdf.

[3] John B. Fraleigh and Raymond A. Beauregard. *Linear Algebra.* Addison Wesley, 3 edition.