Göteborgs Universitet
Institutionen för Matematiska Vetenskaper
Anders Södergren

MMG200 - Linjär algebra, 7.5hp
Höstterminen 2017

Googles PageRank-algoritm

Det här är en kort beskrivning av en tillämpning av egenvärden och egenvektorer som är hämtad från vår vardag. Frågan som jag är intresserad av är *Hur rankas hemsidor i en googlesökning*? Till att börja med vill jag beskriva en typisk sökning. Den består av följande tre steg:

1. Användaren skriver ett ord (eller en mening) X.
2. Sökmotorn hittar *alla* sidor som innehåller X.
3. Sökmotorn rankar och ordnar dessa sidor på något sätt.

Söktjänstföretaget *Google* revolutionerade steg tre med hjälp av deras *PageRank*-algoritm. Deras huvudidé var att rankingen av hemsidor ska baseras på hur sidorna är länkade till varandra istället för deras innehåll. Motiveringen är att innehållet på en hemsida lätt kan manipuleras för att rankas högt i vanliga sökningar, men det är svårare att påverka den storskaliga bilden av vilka (framför allt viktiga och populära) sidor som länkar till en given hemsida.

Nu följer en förenklad beskrivning av hur PageRank fungerar och vad algoritmen har att göra med egenvärden och egenvektorer. Vi betraktar internet som en riktad graf $G = (V,E)$, där

$$\begin{cases} V \text{ är grafens noder. En nod för varje hemsida. Dessa numreras } 1,\dots,N. \\ E \text{ är grafens (riktade) kanter. En kant för varje länk från en hemsida till en annan.} \end{cases}$$

(Se figuren på nästa sida.) Rankingen av hemsidorna i en viss sökning bildar en vektor $\boldsymbol{v} = (v_1,v_2,\dots,v_N) \in \mathbb{R}^N_{\geq 0}$, där v_i står för rankingen (eller värdet) av hemsida i. Rankingen v_i beror på alla länkar till hemsida i. Länkarna viktas enligt principen att vikten på en länk från sida j till sidan i ska

(i) vara omvänt proportionell mot n_j:= totala antalet länkar som utgår från sidan j (om detta antal är positivt; annars sätter vi $n_j = N$).

(ii) bero på rankingen av den länkande sidan j.

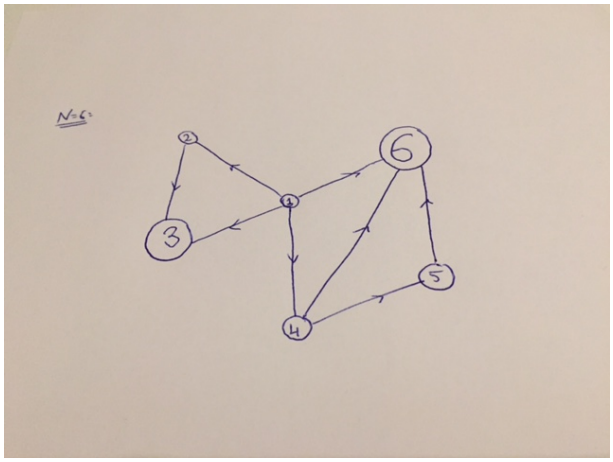
Det anses alltså värdefullt att vara länkad från en populär sida som inte har alltför många länkar till andra sidor. En motivering till detta är att det annars skulle vara lätt att manipulera sökresultat genom att skapa (många) nya sidor som är länkade till alla de sidor som du vill ska rankas högt.

För att formulera detta i termer av en ekvation introducerar vi

$$b_{ij} = \begin{cases} 1, & \text{om } j \text{ länkar till } i \text{ (eller om } j \text{ inte länkar till någon),} \\ 0, & \text{annars.} \end{cases}$$

Observera nu att beskrivningen av PageRank ovan säger att

$$v_i = \sum_{j=1}^{N} \frac{b_{ij}}{n_j} v_j \qquad (1 \leq i \leq N)$$



Figur 1: Figuren illustrerar hur internet kan beskrivas som en graf (om vi antar att det totala antalet hemsidor är 6).

uppfyller villkoren i) och ii) ovan. Om vi dessutom låter P vara den matris som uppfyller

$$P_{ij} = \frac{b_{ij}}{n_j}, \qquad 1 \leq i,j \leq N$$

(dvs. elementet på plats (i,j) i matrisen P är b_{ij}/n_j), så följer det att

$$\boldsymbol{v} = P\boldsymbol{v}. \tag{1}$$

Alltså är rankingsvektorn $\boldsymbol{v}$ en egenvektor till matrisen P med egenvärde 1.

Google undersöker med jämna mellanrum hur internets sidor är länkade till varandra så matrisen P kan betraktas som en känd (och för Google explicit) matris. För att ranka sidorna i en sökning är det alltså 'bara' för Google att lösa egenvärde/egenvektor-problemet i (1) tillsammans med bivillkoret att $\boldsymbol{v} \in \mathbb{R}^N_{\geq 0}$. Hur är det med existens och entydighet av lösningarna till den här ekvationen? P är en så kallad *stokastisk matris*, vilket betyder att alla element är ickenegativa och att för varje kolumn är summan av elementen i den kolumnen lika med 1. För sådana matriser gäller Frobenius sats som säger att $\lambda = 1$ är ett egenvärde till P och att det finns en egenvektor $\boldsymbol{x}$ hörande till detta egenvärde som dessutom uppfyller att alla dess koordinater är större än eller lika med 0. I allmänhet är dock egenrummet hörande till λ inte endimensionellt.

Det enda kvarvarande problemet är nu att matrisen P är väldigt stor, så det är inte trivialt att hitta lösningen till (1).