

Googles PageRank-algorithm

26 november 2021

Början (~1995-1996)

- ▶ Larry Page och Sergey Brin



Början (~1995-1996)

- ▶ Larry Page och Sergey Brin



- ▶ Doktorander i datavetenskap på Stanford University.

Internet är en riktad graf

Pages doktorandprojekt var att studera strukturen hos Internet, modellerad som en riktad graf:

Internet är en riktad graf

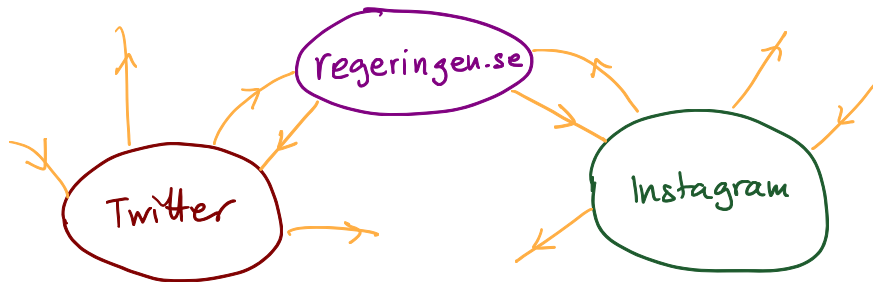
Pages doktorandprojekt var att studera strukturen hos Internet, modellerad som en riktad graf:

- ▶ Noder är **hemsidor**
- ▶ En riktad kant från X till Y är en **länk** på sidan X till sidan Y

Internet är en riktad graf

Pages doktorandprojekt var att studera strukturen hos Internet, modellerad som en riktad graf:

- ▶ Noder är **hemsidor**
- ▶ En riktad kant från X till Y är en **länk** på sidan X till sidan Y



Sökmotorer

Gamla sökmotorer som Altavista och Yahoo hade svårt att sortera ut relevanta träffar från icke-relevanta. Ofta dök de bästa träffarna upp på sidan 8 eller senare.

Sökmotorer

Gamla sökmotorer som Altavista och Yahoo hade svårt att sortera ut relevanta träffar från icke-relevanta. Ofta dök de bästa träffarna upp på sidan 8 eller senare.

PageRanks idé

Viktiga sidor blir länkade till från andra sidor.

Inspirerad av idén att viktiga vetenskapliga artiklar har många citeringar.

Hur gör man i praktiken?

Om jag surfar på måfå genom att klicka på länkar, hur storlek är sannolikheten att hamnar på sidan X ?

Hur gör man i praktiken?

Om jag surfar på måfå genom att klicka på länkar, hur storlek är sannolikheten att hamnar på sidan X ?

Ju större sannolikheten är, desto viktigare är hemsidan enligt PageRank.

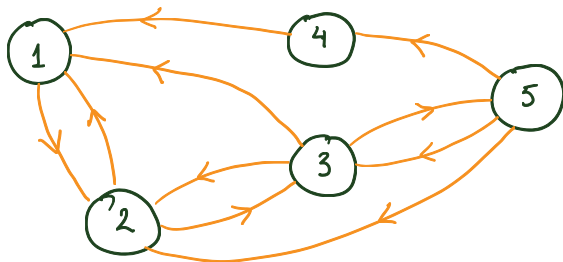
Hur gör man i praktiken?

Om jag surfar på måfå genom att klicka på länkar, hur storlek är sannolikheten att hamnar på sidan X ?

Ju större sannolikheten är, desto viktigare är hemsidan enligt PageRank.

Men hur räknar man ut sannolikheten?

Slumpvandring på en riktad graf, I



Viktad grannmatrix:

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

På plats (i, j) sätter vi $1/d$ om det finns en riktad kant från i till j , där d är antalet riktade kanter med startnod i , annars sätter 0.

Slumpvandring på en graf, II

Tolkning av viktade grannmatriser

Talet på plats (i, j) i Q är sannolikheten att jag går från sida i till sida j om jag slumpmässigt klickar på en länk på sida i .

Slumpvandring på en graf, II

Tolkning av viktade grannmatriser

Talet på plats (i, j) i Q är sannolikheten att jag går från sida i till sida j om jag slumpmässigt klickar på en länk på sida i .

Sannolikhetsvektorer

Idag kallar vi en radvektor

$$\mathbf{p} = (p_1 \ p_2 \ \dots \ p_{n-1} \ p_n)$$

med $p_1 + \dots + p_n = 1$ för en sannolikhetsvektor.

Vi tolkar den som “jag är på sida i med sannolikhet p_i ”.

Beräkna sannolikheter

Sannolikhet efter flera klick

Om min sannolikheten för vilken hemsida jag befinner mig på ges av sannolikhetsvektorn $\mathbf{p}$, så är matrisprodukten

$$\mathbf{p}Q^k$$

en sannolikhetsvektor, som ger sannolikheten för var jag är efter ytterligare k klick.

Beräkna sannolikheter

Sannolikhet efter flera klick

Om min sannolikheten för vilken hemsida jag befinner mig på ges av sannolikhetsvektorn $\mathbf{p}$, så är matrisprodukten

$$\mathbf{p}Q^k$$

en sannolikhetsvektor, som ger sannolikheten för var jag är efter ytterligare k klick.

Perron–Frobenius sats

Det finns en unik sannolikhetsvektor π så att $\mathbf{p}Q^k$ kommer närmare och närmare π ju större k är (oavsett vad $\mathbf{p}$ är). π är den unika sannolikhetsvektorn som löser ekvationen

$$\mathbf{x}Q = \mathbf{x}.$$

Kallas för en egenvektor till Q .

Ranking

I vårt exempel är

$$\pi \approx (0,293 \quad 0,390 \quad 0,220 \quad 0,024 \quad 0,073).$$

Vi rankar hemsida i efter hur stor π_i är. Här är sida 2 “viktigast”, följt av 1, 3, 5 och 4.

Ranking

I vårt exempel är

$$\pi \approx (0,293 \quad 0,390 \quad 0,220 \quad 0,024 \quad 0,073).$$

Vi rankar hemsida i efter hur stor π_i är. Här är sida 2 “viktigast”, följd av 1, 3, 5 och 4.

Säg att jag söker på “lunch”, och att ordet lunch förekommer på hemsidor 1, 2 och 4. Då skulle sökmotorn ge mig de sidorna i ordningen 2, 1, och 4.

Grundläggande version av PageRank

För att undvika vissa patologiska fenomen i internetgrafen så modifierar man slumpvandringen ovan. Man använder istället matrisen

$$Q' = 0.85 \cdot Q + 0,15 \cdot M,$$

där M har talet 1 på alla positioner.

Grundläggande version av PageRank

För att undvika vissa patologiska fenomen i internetgrafen så modifierar man slumpvandringen ovan. Man använder istället matrisen

$$Q' = 0.85 \cdot Q + 0,15 \cdot M,$$

där M har talet 1 på alla positioner.

Tolkningen är någon som surfar på nätet genom att med 85% chans välja en länk på måfå, och med 15% chans gå direkt till en hemsida på måfå.

Avslutning

PageRank var en liten men avgörande del i Googles ursprungliga succé. Hur Google fungerar idag är en industrihemlighet, vem Google säger att PageRank (med modifieringar) fortfarande är en del i sökmotorn.

Avslutning

PageRank var en liten men avgörande del i Googles ursprungliga succé. Hur Google fungerar idag är en industrihemlighet, vem Google säger att PageRank (med modifieringar) fortfarande är en del i sökmotorn.

Matematiken som dök upp här kommer ni se mer av i kurserna framöver:

Matriser, egenvektorer — **Linjär algebra**

Konvergens — **Envariabelsanalys**

Sannolikheter, slumpvandringar — **Matematisk statistik och diskret matematik**