Föreläsning 12: linjär algebra för gymnasister

Förra gången: Markovkedjor.

Kom ihåg:

Def. Låt $\{x_1, x_2, \dots\}$ vara en sekvens av utfallsvektorer i $\mathbb{R}^m$. Om det finns en stokastisk matris A så att

$$x_{n+1} = A x_n$$

för alla $n \geq 1$ så kallas sekvensen för en Markovkedja.

Vi lärde oss även vad en stokastisk matris är. Detta behövs även idag:

Def. En kvadratisk $(m \times m)$-matris $A = (a_{ij})$ kallas stokastisk om

(i) alla element i A är icke-negativa,

$$a_{ij} \geq 0 \quad \forall\, i,j = 1, \dots m.$$

(iii) Kolumnsummorna är alla lika med ett, dvs

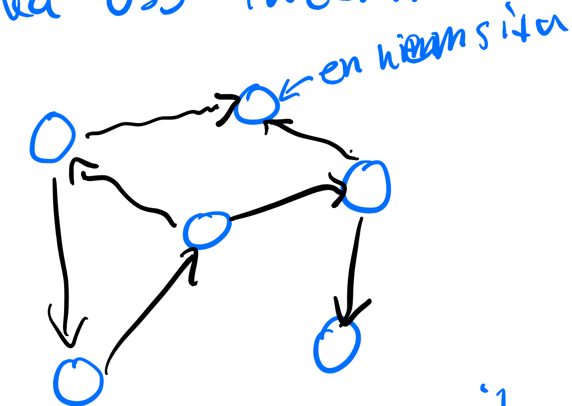
$$\sum_{j=1}^{m} a_{ij} = 1 \quad \forall i = 1, \dots, m.$$

IDAG: Google och Informationssökning

Sökmotor / Söktjänst

Mål: Reda ut vilka hemsidor som är mest relevant för en given sökning.

Vi kan tänka oss internet som en graf.



Problem: ~ 10 miliarder hemsidor.

Vissa söktjänster söker igenom webben efter innehåll och skapar ett index med nyckelord för att underlätta för webbanvändaren.

(Ex: Webcrawler (1994, Brian Pinkerton)

Rangordnar hemsidorna 1, 2, 3 .... ~10 miljarder

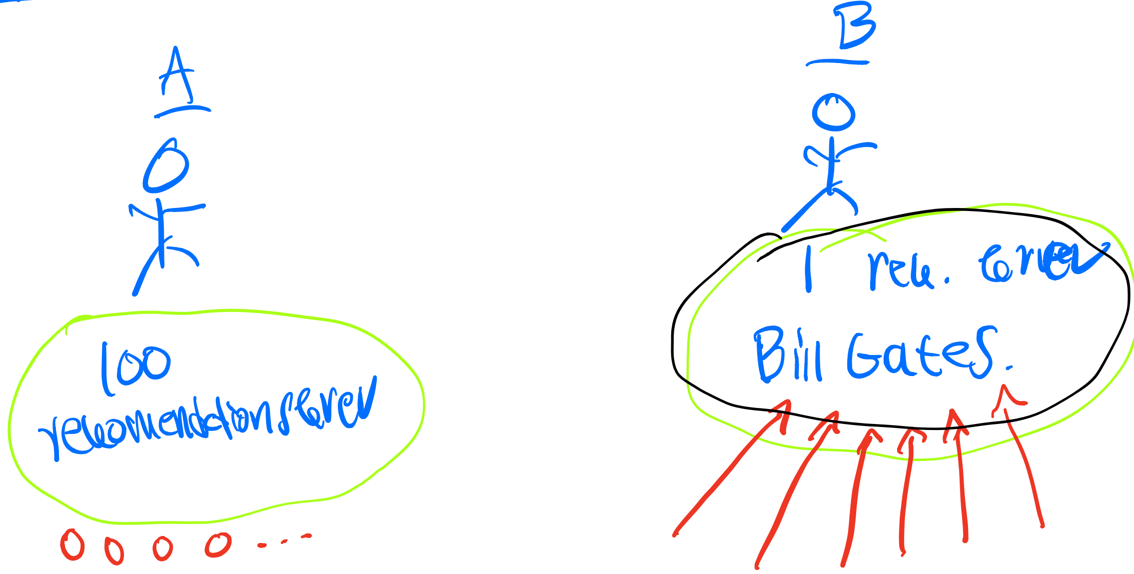
matris: 1, 83, 1004, ...

Olika sätt att rangordna på:

- Antalet klick på en sida.
- Rangordna efter sidor som innehåller flest ord från sökningen.
- Hur många länkar som en sida har till sig.

Problem: Aktörer kan skapa hemsidor för att få fler länkar till sig.

Ex: Två personer söker jobb inom IT.



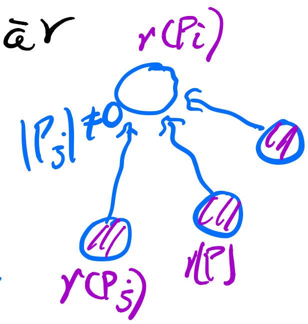
PageRank:

Matematisk modell: Låt $P_1, \dots, P_n$ beteckna hemsidor och låt $|P_j|$ vara antalet länkar från hemsida P_j.

Definition. Rangen till en hemsida P_i skriver vi som $r(P_i)$ och där

$$r(P_i) = \sum_{j \to i} \frac{r(P_j)}{|P_j|}.$$

(rekursiv def).

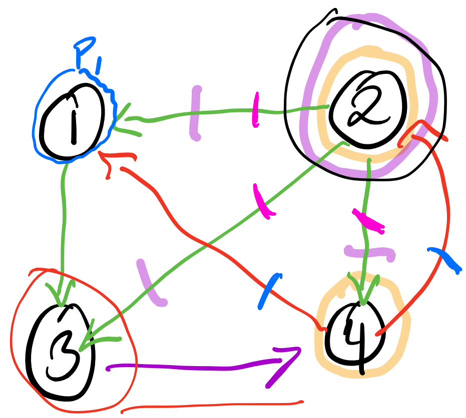


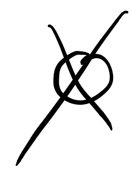
Definition. Länkmatrisen. Låt $P_1, \dots, P_n$ vara hemsidor. Länkmatrisen H är en $n \times n$-matris där

(i,j) i H

$$H_{ij} = \begin{cases} 1/|P_j| & \text{om } j \mapsto i \\ 0 & \text{annars.} \end{cases}$$

Exempel: Betrakta följande "internet"





$$\begin{cases} r(P_1) = \dfrac{r(P_2)}{3} + \dfrac{r(P_4)}{2} \\ r(P_2) = \dfrac{r(P_4)}{2} \\ r(P_3) = r(P_1) + \dfrac{r(P_2)}{3} \\ r(P_4) = \dfrac{r(P_3)}{3} + r(P_3) \end{cases} \qquad (*)$$

och

$$H = \begin{array}{cccc} H_{21} & \downarrow & P_3 & P_4 \\ \end{array}$$
$$H = \begin{pmatrix} 0 & 1/3 & 0 & 1/2 \\ 0 & 0 & 0 & 1/2 \\ 1 & 1/3 & 0 & 0 \\ 0 & 1/3 & 1 & 0 \end{pmatrix}$$
$$\begin{array}{cccc} 1 & 1 & 1 & 1 \end{array}$$

$H_{31} =$

$\underline{R = HR.}$

Observera att $(*) \Longleftrightarrow R = HR$ där $R = \begin{pmatrix} r(P_1) \\ r(P_2) \\ r(P_3) \\ r(P_4) \end{pmatrix}$

Definiera $R = \begin{pmatrix} r(P_1) \\ r(P_2) \\ r(P_3) \\ r(P_4) \end{pmatrix} \geq 0$ som rangeringsvektorn.

och den bästa hemsidan är sidan kopplad till $\max_i \{r(P_i)\}$.

Finns det lösningar? För att garantera existensen av lösningar sätts

$$HX = \lambda \cdot X$$

egenvektor

egenvärde

där $X = (x_1, \ldots, x_n)$.

Så, detta betyder att värdet (rangen) av en hemsida P_j antas vara

Proportionell mot värdena som länkad till P_j.

Proposition. Låt $\bar{x}$ vara en vektor så att $H\bar{x} = \lambda\bar{x}$. Antag att elementen i $\bar{x} = (x_1, \dots, x_n)$ inte summerar till noll dvs. $x_1 + x_2 + x_3 + \dots + x_n \neq 0$. Då är $\lambda = 1$ (ett egenvärde).

Bevis. Låt H vara länkmatrisen

$$H = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ \vdots & & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{pmatrix}.$$

Vi vet att H är stokastisk, ty $h_{ij} \geq 0 \ \forall i,j$ men även kolumnsumma

$$\sum_{j=1}^{n} h_{ij} = \left| h_{ij} = \begin{cases} \frac{1}{|P_j|} \\ 0 \end{cases} \right|$$

$$= \sum_{k=1}^{|P_j|} \frac{1}{|P_j|} = \frac{|P_j|}{|P_j|} = 1.$$

Vår ekvation: $H\bar{X} = z\bar{X}$

$$\iff \underbrace{\begin{pmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{n1} & \cdots & h_{nn} \end{pmatrix}}_{H} \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}}_{\bar{X}} = z \underbrace{\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}}_{\bar{X}}$$

$$\iff \begin{pmatrix} \sum_{k=1}^{n} h_{1k} x_k \\ + \\ \sum_{k=1}^{n} h_{2k} x_k \\ \vdots \\ + \\ \sum_{k=1}^{n} h_{nk} x_k \end{pmatrix} = z \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

Summera båda kolumnerna:

$$\sum_{j=1}^{n} \sum_{k=1}^{n} h_{jk} x_k = z \sum_{k=1}^{n} x_k$$

$$\sum_{k=1}^{n} \underbrace{\left[\sum_{j=1}^{n} h_{jk} \right]}_{=1} x_k = z \sum_{k=1}^{n} x_k$$

$$\cancel{\sum_{k=1}^{n} x_k} = z \cancel{\sum_{k=1}^{n} x_k}$$

$$2 = 1.$$

Övning: Fixa felet i beviset !

Hur ser rangordningen ut för "vårt" internet? Lös

$$R = HR.$$

$$\iff (H-I)R = 0 \qquad R = \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix}$$

$$\iff \left(\begin{array}{cccc|c} 1 & -1/3 & 0 & -1/2 & 0 \\ 0 & 1 & 0 & -1/2 & 0 \\ -1 & -1/3 & 1 & 0 & 0 \\ 0 & -1/3 & -1 & 1 & 0 \end{array}\right)$$

Gauss

$$\sim \dots \sim \implies \begin{cases} x = \frac{2}{3}t \\ y = \frac{1}{2}t \\ z = \frac{5}{6}t, & \text{där } t \in \mathbb{R} \\ w = t \end{cases}$$

Så, summan av elementen i $R = 1$

$s=1$

$$\frac{2}{3}t + \frac{1}{2}t + \frac{5}{6}t + t = 1$$

$$\iff \quad t = 1/3$$

$$\Rightarrow \quad \underline{\underline{R = \begin{pmatrix} 2/9 \\ 1/6 \\ 5/18 \\ 1/3 \end{pmatrix}}} > 0.$$

$\therefore$ Hemsida P_4 rangordnas först.
Följs av P_3...