

Un texte , une modélisation

Laurent Dumas

Texte 1: la matrice de Google

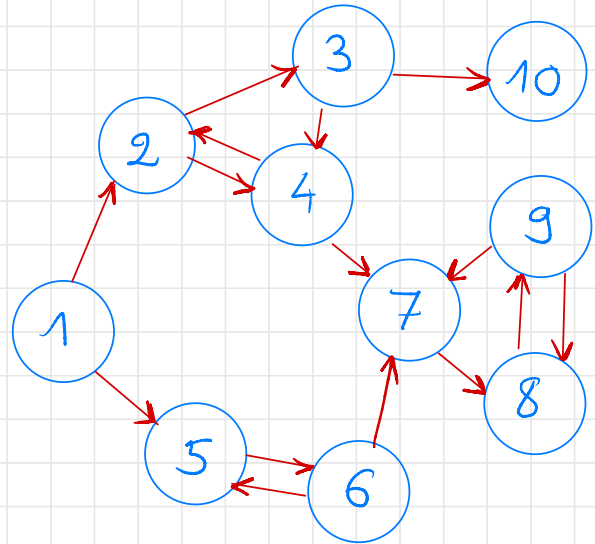
* Objectif: expliquer et illustrer

l'algorithme PageRank de Google

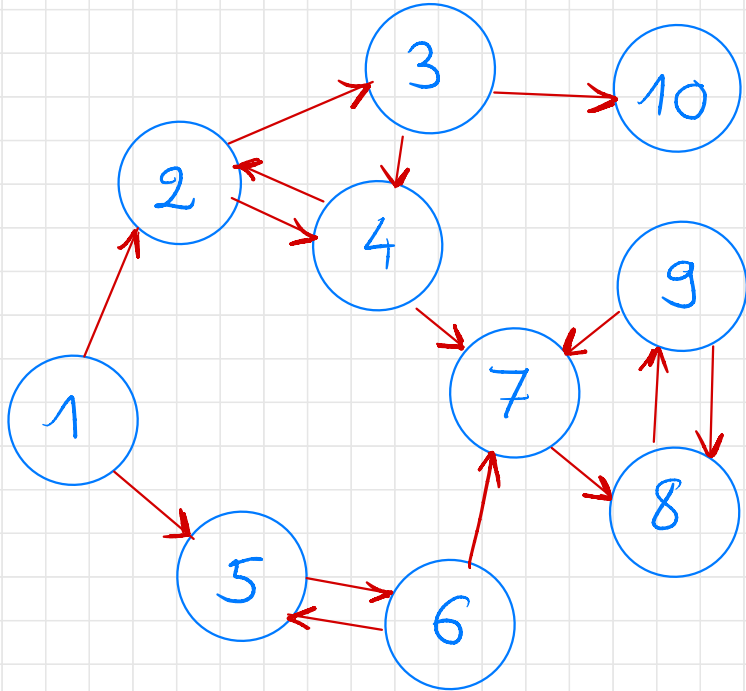
* Exemple étudié: un mini-réseau
de 10 pages web reliées entre elles

* Outils mathématiques:

* Algèbre linéaire (recherche
de valeurs propres)



Etape 1 : traduction du réseau sous forme matricielle



→ Matrice C de connectivité :
 $C \in \mathcal{M}_N(\mathbb{R})$ ($N=10$) avec :

$$C_{ij} = \begin{cases} 1 & \text{si le site } j \text{ pointe vers le site } i \\ 0 & \text{sinon} \end{cases}$$

G_n note :

$$N_j = \sum_{i=1}^N C_{ij} : \text{nombre de liens du site } j$$

→ Matrice Q avec liens pondérés :

$$Q_{ij} = \begin{cases} c_{ij}/N_j & \text{si } N_j \neq 0 \\ 0 & \text{sinon} \end{cases}$$

→ Matrice P avec coefficients > 0

$$P_{ij} = \begin{cases} Q_{ij} & \text{si } N_j \neq 0 \\ \frac{1}{N} & \text{sinon} \end{cases}$$

On a alors

$$\begin{aligned} * P_{ij} &> 0 \quad \forall (i,j) \in \{1, \dots, N\}^2 \\ * \sum_{i=1}^N P_{ij} &= 1, \quad \forall j \in \{1, \dots, N\} \end{aligned}$$

* On dit que P est une matrice stochastique.

* Elle possède $\lambda = 1$ comme plus grande valeur propre (en module).

Etape 2 : classement des pages:

A chaque page i , on associe un score r_i . Celui-ci vérifie :

$$r_i = \sum_{j=1}^N p_{ij} r_j$$

c'est à dire

$$\boxed{r = Pr} \text{ avec } r = \begin{pmatrix} r_1 \\ \vdots \\ r_N \end{pmatrix}$$

$\Rightarrow$ le vecteur r est un vecteur propre de P par la valeur propre $\lambda = 1$

Etape 3 : utilisation du

théorème de Perron - Frobenius

Afin d'obtenir une unique solution r telle que :

$$\boxed{\|r\|_1 = \sum_{i=1}^N r_i = 1 \text{ et } \forall i \in \{1, \dots, N\}, r_i > 0}$$

on transforme la matrice P en une nouvelle matrice A :

$$A = \alpha I + (1-\alpha) \frac{1}{N} e^t e$$

où $e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$, et $\alpha \in]0, 1[$.

On peut alors montrer que la valeur propre $\lambda = 1$ de A est simple, associée à un vecteur propre π à coefficients strictement positifs (et que les autres valeurs propres de A sont de module inférieur à α).

Etape 4 : estimation numérique du vecteur de score

On utilise la méthode de la puissance pour approcher π :

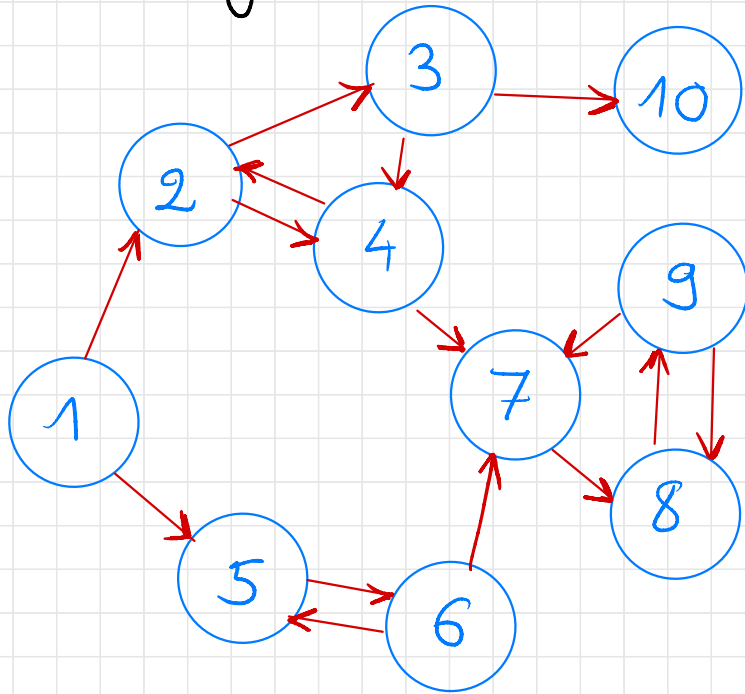
$\pi_0 \in \mathbb{R}^N$, aléatoire tq $\|\pi_0\|_1 = 1$ et à coefficients > 0 et

$$\pi_{k+1} = A \pi_k$$

On peut montrer que

$$\lim_{k \rightarrow +\infty} \pi_k = \pi$$

Etape 3: simulation avec le logiciel Scilab



$$N = 10, \alpha = 0,85$$

Pour aller plus loin :

- Utilisation d'une matrice creuse dans la méthode de la puissance.
- Etude du conditionnement numérique de l'algorithme.

Référence :

www.agreg.org

(texte de modélisation, 2008)