Algorithme PageRank

Objectif : Déterminer le degré d'importance d'une page Web.

Soit l'ensemble des pages contenant les mots clés d'une requète. On cherche à établir un classement de ces pages (basé non pas sur la consultation des pages par les internautes mais sur le nombre de liens qui pointent vers ces pages).

On considère une numérotation de 1 à n de cet ensemble et la **matrice de connectivité** $C \in \mathcal{M}_n(\mathbb{R}^+)$ définie par :

$$c_{ij} = \begin{cases} 1 & \text{si la page j pointe vers la page i, avec } i \neq j \\ 0 & \text{sinon} \end{cases}.$$

Algorithme PageRank

On définit :

- Le nombre total de liens présents sur chaque page :

$$n_j = \sum_{k=1}^{n} c_{kj},$$

- L'indice de pertinence d'une page :

$$x_i = \sum_{j=1}^{n} c_{ij} \frac{x_j}{n_j}$$

(dépendant des pages pointant vers celle-ci, pondéré par l'indice de pertinence et le nombre total de liens des pages ciblantes).

Le problème revient donc à trouver $x \in (\mathbb{R}^{+,*})^n$ tel que

$$x = \widetilde{C}x, \quad \text{avec} \quad \widetilde{c}_{ij} = \frac{c_{ij}}{n_j} \quad (\text{ et } \widetilde{c}_{ij} = 0 \text{ si } n_j = 0).$$

$\Rightarrow$ Trouver un vecteur propre de $\widetilde{C}$ associé à la valeur propre 1.

Algorithme PageRank

Problème : $\widetilde{C}$ n'admet pas forcément 1 comme valeur propre.

On définit une perturbation E de $\widetilde{C}$ en rajoutant des liens artificiels :

$$e_{ij} = \widetilde{c}_{ij} + \frac{1}{n} d_j \quad \text{où} \quad d_j = \begin{cases} 1 & \text{si } n_j = 0 \\ 0 & \text{sinon} \end{cases}.$$

E^T est une **matrice stochastique** ($e_{ij} \in \mathbb{R}^+$ et $\sum_i e_{ij} = 1, \ \forall j$) et donc

$$E^T e = e, \qquad \text{avec } e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Puisque 1 est valeur propre de E^T, 1 est aussi valeur propre de E (par contre, on ne connait pas un vecteur propre qui lui est associé).

Algorithme PageRank

Problème : 1 n'est pas forcément valeur propre simple de E alors qu'on souhaiterait un classement unique.

On considère, pour $0 < \alpha < 1$ (en pratique $\alpha = 0.85$),

$$A = \alpha E + (1 - \alpha)\frac{1}{n} e e^T.$$

Comme E^T est stochastique, A^T est aussi stochastique. De plus, A est une **matrice primitive** :

$$\exists k \in \mathbb{N}, \quad (A^k)_{ij} > 0$$

(pour cette matrice, $k = 1$).

D'après le **théorème de Perron-Frobenius**, 1 est valeur propre simple de A, $\rho(A) = 1$ et il existe un unique $x \in (\mathbb{R}^{+,*})^n$ tel que $\|x\|_1 = 1$ et $Ax = x$.

Algorithme PageRank

Théorème (Perron-Frobenius) : Soit A une matrice de $\mathcal{M}_n(\mathbb{R}^+)$ (pas nécessairement de $\mathcal{M}_n(\mathbb{R}^{+,*})$) et primitive. Alors elle admet une valeur propre réelle strictement positive $r > 0$ telle que

- pour toute autre valeur propre s de A, on a $|s| < r$,
- r est une valeur propre simple,
- il existe un unique vecteur x^+ de norme 1 à coordonnées strictement positives telles que $Ax^+ = rx^+$.

Démonstration : admise.