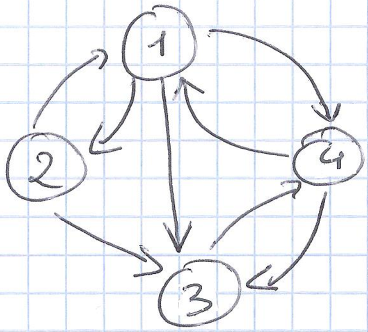
Calcul de la pertinence d'une page Web : algorithme de PageRanking

<u>Objectif</u> : classer les pages Web (qui sont triés à partir de mots clés)

<u>Exemple de réseau simplifié</u> :

Les flèches indiquent l'existence d'un lien entre les pages : $i \to j$



* <u>1ère méthode</u> : comptage des liens

$$M = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$
$$\begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ 2 & 1 & 3 & 2 \end{array}$$

$m_{ij} = \begin{cases} 1 \text{ s'il existe un lien } i \to j \\ 0 \text{ sinon} \end{cases}$

On obtient donc le classement suivant :

Rang	1	2	3	4
Page	3	1 ou 4		2

<u>Problème</u> : très facile à manipuler.
Par exemple, créer une page avec beaucoup de liens vers le site que l'on souhaite voir bien classé

* <u>2ème méthode</u> : comptage pondéré
(chaque lien est divisé par le nombre de liens de la page)

$$M = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$
$$\begin{array}{cccc} \downarrow & \downarrow & \downarrow & \downarrow \\ 1 & \frac{1}{3} & \frac{4}{3} & \frac{4}{3} \end{array}$$

$m_{ij} = \begin{cases} \frac{1}{\lambda_i} \text{ s'il existe } i \to j \text{ et } \lambda_i = \text{nb de liens de la page } i \\ 0 \text{ sinon} \end{cases}$

On obtient alors le classement suivant

Rang	1	2	3	4
Page	3 ou 4		1	2

<u>Problème</u> : également facile à manipuler.
Par exemple créer beaucoup de pages pointant chacune vers le site que l'on souhaite bien classé

* $3^{\text{ème}}$ méthode : Comptage récursif.

On continue avec le comptage pondéré et la pertinence de la page ciblant le site sera prise en compte

La pertinence va donc évoluer selon les changements de pages

Soit L_n la matrice ligne mesurant la pertinence des pages à chaque étape n (les clics).

On a donc $L_{n+1} = L_n \times M$ avec $M = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$

soit $L_n = L_0 \times M^n$

* L_0 donne la position du "surfeur" au départ.
* On s'interesse donc à $\lim\limits_{n \to +\infty} L_n$

Remarque : On assimile le déplacement sur le réseau à une marche aléatoire : le score d'une page correspond à la probabilité que le surfeur s'y arrête

L'étude matricielle nous montre que L_n converge vers L avec $L = \begin{pmatrix} \frac{3}{13} & \frac{1}{13} & \frac{4}{13} & \frac{5}{13} \end{pmatrix}$. On obtient donc

Rang	1	2	3	4
Page	4	3	1	2

Problème : certaines pages ne possède pas de liens extérieurs ou peut également entrer dans une partie du réseau sans pouvoir en sortir.

On introduit alors un coefficient d'échappement qui offre la possibilité, à chaque étape de se diriger vers n'importe quelle autre page

$c \in]0;1[$ et $c = \frac{1}{\lambda}$ où λ est le nombre moyen de pages visitées sur le réseau

Pour $c = 0{,}2 = \frac{1}{5}$ on obtient la nouvelle "matrice de transition" N

$$N = \begin{pmatrix} \frac{1}{20} & \frac{19}{60} & \frac{19}{60} & \frac{19}{60} \\ \frac{9}{20} & \frac{1}{20} & \frac{9}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{17}{20} \\ \frac{9}{20} & \frac{1}{20} & \frac{9}{20} & \frac{1}{20} \end{pmatrix} \quad \text{soit } N = \frac{c}{4}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} + (1-c)\begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix} = \frac{c}{4}J + (1-c)M$$

Ainsi $L_{n+1} = L_n \times N$ soit $L_n = L_0 \times N^n$

L'étude matricielle nous montre que L_n converge vers L_s avec

$L_s = \begin{pmatrix} \frac{135}{572} & \frac{323}{2860} & \frac{171}{572} & \frac{1007}{2860} \end{pmatrix}$. Ainsi

Rang	1	2	3	4
Page	4	3	1	2