Pertinence d'une page web

PAUL MILAN

Lycée d'Adultes de la ville de Paris :
http://www.lyceedadultes.fr/index.html

1 Que fait un moteur de recherche

Contrairement à une base de données structurée dont on peut facilement extraire des informations, le Web est une immense collection de textes de toutes natures qui évolue en permanence.

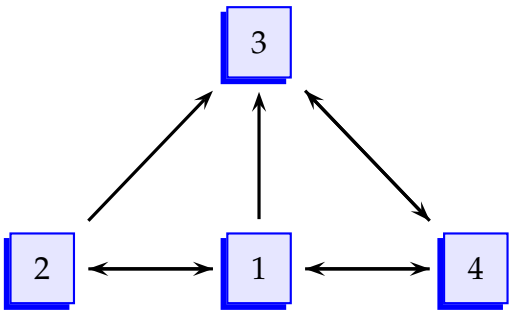
Le moteur de recherche copie préalablement les pages sur des milliers d'ordinateurs (60 000 pour Google !) et les trie par ordre alphabétique (selon les mots clé). Lors d'une requête relative à un mot clé, le moteur répond par la liste des pages contenant ce mot clé. Mais il y en a des dizaines de milliers en général.

C'est ici qu'intervient l'innovation de Larry Page (fondateur avec Serguei Brin de Google) connue sous le nom de *Pagerank* (*to rank : classer*). L'idée est de répondre à la requête en citant les pages par ordre de pertinence.

Le Web a une structure de graphe due au fait que les pages se citent mutuellement. C'est le principe de l'hypertexte : les pages se citent mutuellement par les fameux liens. Essayons de déterminer la pertinence d'une page web.

2 Un exemple

Soit le graphe suivant qui relie 4 pages Web :



3 Mesurer la pertinence

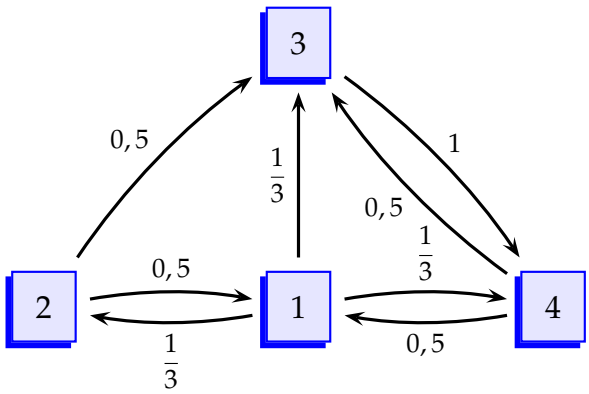
3.1 Comptage pondéré

Il est clair qu'une page importante reçoit de nombreux liens. Cependant certaines pages émettent beaucoup de liens ce qui d'une certaine façon diminue leur poids. On pondère alors les liens qui relient les pages.

La page 1 pointe vers 3 pages (2,3,4). Chacun de ces liens sera alors pondéré du coefficient $\frac{1}{3}$

On peut associer à ce graphe la matrice $\mathbf{M} = (a_{ij})$ où a_{ij} représente le coefficient de la page i qui pointe sur j.

$$\mathbf{M} = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{pmatrix}$$



On peut alors définir la mesure de pertinence μ_i de la page i en comptant le nombre de liens pondérés qui pointent vers elle :

$$\mu_i = \sum_j a_{ji}$$

On obtient les pertinences des pages 1, 2, 3, 4 : $\mu_1 = 1$, $\mu_2 = \frac{1}{3}$, $\mu_3 = \frac{4}{3}$, $\mu_4 = \frac{4}{3}$
Les pages 3 et 4 sont donc les plus pertinentes.

On pose la matrice ligne des pertinences **P** et la matrice ligne **J** avec que des 1, on a alors :

$$\mathbf{P} = \mathbf{J} \times \mathbf{M}$$

Remarque : Mais ce comptage est très facile à manipuler, puisqu'il suffit de créer des "fausses" pages pointant vers la page i pour en augmenter l'importance.

3.2 Comptage récursif

La pertinence d'une page est renforcée par la pertinence des pages qui pointent vers elle et elle est diminuée par la dispersion éventuelle des liens issus de ces dernières.

En reprenant la pondération précédente, on peut définir la pertinence d'une page i de la façon suivante :

$$\mu_i = \sum_j a_{ji} \times \mu_j$$

Le risque de manipulation consistant en l'ajout de pages vides de sens est alors ici annulé puisqu'une telle page recevrait une mesure de pertinence nulle.

On a alors : $\mathbf{P} = \mathbf{P} \times \mathbf{M}$

On obtient le système suivant :

$$\begin{cases} \mu_1 = \frac{1}{2}\mu_2 + \frac{1}{2}\mu_3 \\ \mu_2 = \frac{1}{3}\mu_1 \\ \mu_3 = \frac{1}{3}\mu_1 + \frac{1}{2}\mu_2 + \frac{1}{2}\mu_4 \\ \mu_4 = \frac{1}{3}\mu_1 + \mu_3 \end{cases}$$

D'où les pertinences en fonction de μ_1 :

$$\mu_2 = \frac{1}{3}\mu_1, \quad \mu_3 = \frac{4}{3}\mu_1, \quad \mu_4 = \frac{5}{3}\mu_1$$

Si on fixe la somme des pertinences à 1, on obtient alors :

$$\mu_1 = \frac{3}{13}, \quad \mu_1 = \frac{1}{13}, \quad \mu_3 = \frac{4}{13}, \quad \mu_4 = \frac{5}{13}$$

Remarque : La page 4 est alors la plus pertinente.

4 Pertinence et probabilité

Dans la matrice [1] $\mathbf{M}$, on peut remarquer que la somme des coefficients a_{ij} sur une ligne est égal à 1.

Ces coefficients a_{ij} peuvent donc s'interpréter comme la probabilité, pour un "surfeur" qui se trouverait à la page i de suivre le lien qui l'amènerait à la page j.

On suppose que le surfeur est sur une page donné à l'instant 0 et qu'il évolue de page en page en cliquant sur les liens au hasard.
En notant $\mathbf{U}_n$ la matrice ligne admettant pour coefficient à la colonne i la probabilité que le surfeur se trouve à la page i au bout de n clics, les relations précédentes peuvent se traduire par la relation matricielle suivante :

$$\mathbf{U}_{n+1} = \mathbf{U}_n \times \mathbf{M} \quad \Rightarrow \quad \mathbf{U}_n = \mathbf{U}_0 \times (\mathbf{M})^n$$

On peut montrer que cette suite $(\mathbf{U}_n)$ converge vers $\left(\frac{3}{13}; \frac{1}{13}; \frac{4}{13}; \frac{5}{13}\right)$ et ce quelque soit l'état d'origine. On a donc bien une probabilité plus grande de se retrouver en page 4 après un grand nombre de clics !

Remarque : Toutefois, il peut arriver que certaines pages ne comportent aucun lien vers d'autres pages ; dans ce cas, lorsque le surfeur aléatoire arrive sur l'une d'entre elles, il lui est impossible de la quitter.

5 Saut aléatoire

C'est pourquoi Google utilise une astuce : à chaque page, avec une probabilité p le surfeur peut renoncer à suivre les liens et abandonner sa page actuelle pour une autre page choisie au hasard parmi les n pages du Web.

On obtient alors une nouvelle matrice $\mathbf{M}'$ caractérisant le nouveau système.

$$\mathbf{M}' = \frac{p}{4}\mathbf{J} + (1-p)\mathbf{M}$$

Si on prend $p = \frac{1}{5}$, on obtient :

$$\mathbf{M}' = \frac{1}{20}\mathbf{J} + \frac{4}{5}\mathbf{M} = \begin{pmatrix} \frac{1}{20} & \frac{19}{60} & \frac{19}{60} & \frac{19}{60} \\ \frac{9}{20} & \frac{1}{20} & \frac{9}{20} & \frac{1}{20} \\ \frac{1}{20} & \frac{1}{20} & \frac{1}{20} & \frac{17}{20} \\ \frac{9}{20} & \frac{1}{20} & \frac{9}{20} & \frac{1}{20} \end{pmatrix}$$

On démontre que les puissances de la matrice $\mathbf{M}'$ conduisent à une matrice limite et à des indices de pertinence qui sont :

$$\frac{135}{572}, \quad \frac{323}{2860}, \quad \frac{171}{572}, \quad \frac{1007}{2860}$$

à comparer aux indices trouvés précédemment

	μ_1	μ_2	μ_3	μ_4
Sans saut aléatoire	0,23	0,08	0,31	0,38
Avec saut aléatoire	0,24	0,11	0,30	0,35

Dans les deux cas la page 4 demeure la plus pertinente

Pour résumer, plus vous aurez de liens de qualité vers votre site, plus votre *PageRank* sera élevé, plus vos chances d'apparaître en bonne position dans le moteur de recherche Google seront accrues.

1. Cette matrice est une matrice de transition d'une chaîne de Markov. Une chaîne de Markov est un processus aléatoire portant sur un nombre fini d'états, avec des probabilités de transition sans mémoire