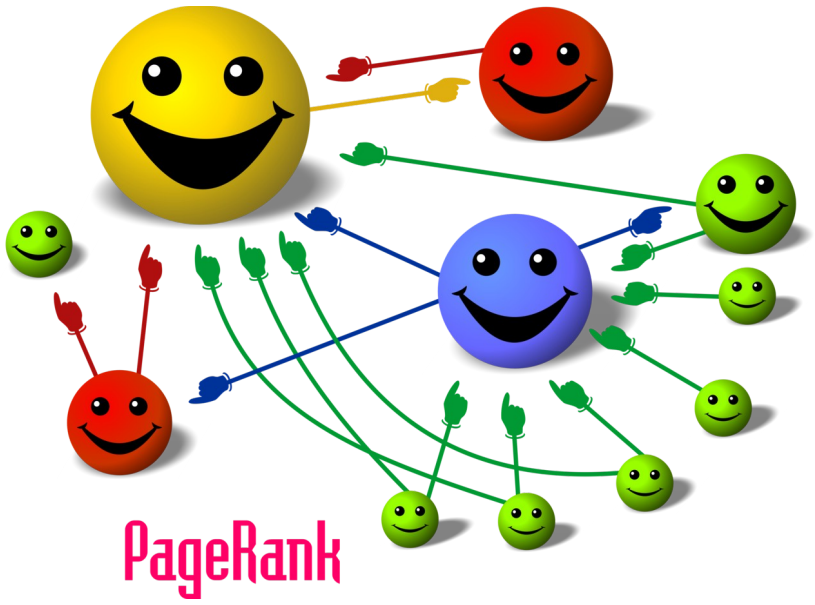
PERTINENCE D'UNE PAGE WEB

Notions réinvesties : inverse d'une matrice carrée



Le *PageRank* (ou *PR*) est l'algorithme d'analyse des liens utilisé par le moteur de recherche **Google**. Il mesure quantitativement la popularité d'une page web. Le *PageRank* n'est qu'un indicateur parmi d'autres dans l'algorithme qui permet de classer les pages du Web dans les résultats de recherche de Google.
Ce système a été inventé par Larry Page, cofondateur de Google.

D'autres critères que les liens entrants sont pris en compte dans le calcul du *PageRank*, la recette exacte étant gardée secrète par Google : les liens sortants, les ancres, le trafic de la page, le comportement des visiteurs, le nom de domaine ou encore l'hébergement influencent le score.
Depuis le 15 avril 2016, Google a donc officiellement arrêté d'afficher le PageRank sur sa Toolbar. Cela ne signifie pas pour autant que la qualité des liens entrants et sortants n'est plus un critère majeur pour le positionnement d'une page, bien au contraire. Si Google met fin à la communication publique de ce score, elle utilise toujours ce système en interne et l'obtention de liens de qualité reste à ce jour un des meilleurs moyens pour progresser dans les classements sur les pages de résultats.

Partie 1 : un problème difficile de classement

Un moteur de recherche sur Internet a pour but, à partir de mots-clés, d'afficher une liste de pages contenant ces mots. Ces pages sont alors triées par ordre de « pertinence ».

Mais comment peut-on estimer qu'une page est plus pertinente (ou plus importante) qu'une autre ?

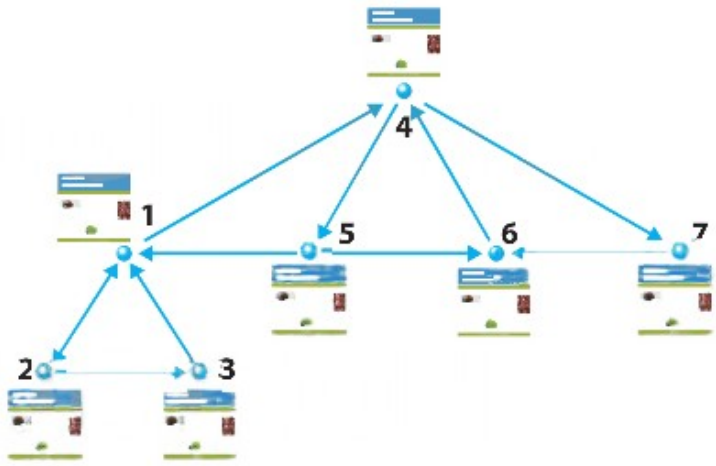
Étudions trois modèles de calcul de la pertinence d'une page web, le troisième modèle faisant appel au calcul matriciel.

Prenons comme exemple (voir ci-contre) un réseau de sept pages, numérotées de 1 à 7.

Dans ce problème, les nombres i et j représentent des entiers entre 1 et 7.

Lorsqu'il existe un lien de la page *j* vers la page *i*, on symbolise ce lien par une flèche allant de *j* à *i*.

On notera p_i la pertinence de la page *i*.



A. Premier modèle : comptage du nombre de liens

Ce modèle repose sur l'affirmation : « Plus une page reçoit de liens, plus elle est importante ».

La pertinence p_i de la page *i* est alors le nombre de liens reçus par la page *i*, ce qui s'écrit :

$$p_i = \sum_{j \to i} 1.$$

1. Déterminez la ou les pages les plus pertinentes avec ce modèle.

2. Supposons que l'on ajoute dans ce réseau deux pages (pages 8 et 9), ne recevant aucun lien et émettant chacune un lien unique vers la page 5. Quelles seraient alors les pages les plus pertinentes dans ce nouveau réseau ?

3. À partir de la question précédente, expliquez l'inconvénient majeur de ce modèle.

B. Deuxième modèle : comptage pondéré du nombre de liens

- Prenons par exemple la page 1 : elle reçoit un lien des pages 2, 3 et 5.

De plus, la page 2 émet deux liens. On va alors considérer dans ce modèle que le lien de la page 2 vers la page 1 apporte un « poids » de $\frac{1}{\text{nombre de liens émis}}$ c'est-à-dire $\frac{1}{2}$ dans le calcul de la pertinence de la page 1.

De même, la page 3 apporte un poids de 1 et la page 5 un poids de $\frac{1}{2}$.

Ainsi $p_1 = \frac{1}{2} + 1 + \frac{1}{2} = 2$.

- Plus généralement, si on note ℓ_j le nombre de liens émis par la page *j*, chacun de ces ℓ_j liens émis apportera le même « poids » $\frac{1}{\ell_j}$ dans le calcul de la pertinence de chacune des pages recevant ces liens.

Ainsi la pertinence p_i de la page *i* est donnée par la formule : $p_i = \sum_{j \to i} \frac{1}{\ell_j}$.

Reprenez les trois questions du paragraphe précédent avec ce nouveau modèle.

C. Troisième modèle : prise en compte de la pertinence des pages émettrices

Ce modèle repose sur l'affirmation : « Une page est importante si beaucoup de pages importantes la citent ».

La pertinence p_i de la page i vérifie la formule : $p_i = \sum_{j \to i} \frac{1}{\ell_j} p_j$.

1. a) Vérifiez que $\frac{1}{2}p_2 + p_3 + \frac{1}{2}p_5 = p_1$, ce qui revient à : $-p_1 + \frac{1}{2}p_2 + p_3 + \frac{1}{2}p_5 = 0$.

Écrivez ainsi les six autres équations induites par ce modèle.

b) Vérifiez que si on note L_i la i-ième ligne de ce système, alors $L_7 = -\sum_{i=1}^{6} L_i$.

Déduisez-en que la dernière équation de ce système est inutile.

Aide

Notez que si le 7-uplet $(p_1, p_2, p_3, p_4, p_5, p_6, p_7)$ vérifie les six premières équations du système, alors il vérifie la septième.

2. Pour que les pertinences soient définies de manière unique, on leur impose la condition :

$$p_1 + p_2 + p_3 + p_4 + p_5 + p_6 + p_7 = 1.$$

Cela revient à considérer les pertinences comme des pourcentages, dont la somme est 100 %.

La ligne L_7 du système est alors remplacée par cette équation.

a) Vérifiez que ce nouveau système peut se traduire par l'égalité matricielle : $AX = B$ (*) où :

$$A = \begin{pmatrix} -1 & 0{,}5 & 1 & 0 & 0{,}5 & 0 & 0 \\ 0{,}5 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0{,}5 & -1 & 0 & 0 & 0 & 0 \\ 0{,}5 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0{,}5 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0{,}5 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \\ p_5 \\ p_6 \\ p_7 \end{pmatrix} \quad \text{et} \quad B = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

Cette écriture est appelée l'**écriture matricielle** du système.

b) À l'aide de votre calculatrice ou d'un tableur, calculez la matrice $A^{-1}B$ puis justifiez que cette matrice est égale à la matrice X dans l'égalité (*).

Déduisez-en la pertinence, à 10^{-2} près, de chacune de ces sept pages.

c) Supposons que l'on ajoute dans ce réseau deux pages (pages 8 et 9) ne recevant aucun lien et émettant chacune un lien unique vers la page 5. Quelles seraient les pertinences de ces deux pages dans le nouveau réseau ainsi formé ?

Ces deux pages augmenteraient-elles la pertinence de la page 5 dans ce nouveau réseau ? Expliquez.

Des questions restent sans réponse…

Ce dernier modèle est celui retenu pour mesurer la pertinence d'une page Web : on ne peut pas truquer le comptage en ajoutant des pages artificielles comme dans les deux premiers modèles.

On pouvait avoir, dès le début, l'intuition que la page 4 serait la plus « pertinente » de toutes : parmi les pages 1, 2, 3, la page 1 sert de référence commune et semble un bon point de départ pour chercher des informations, comme la page 6 parmi les pages 5, 6, 7. Ces deux pages « importantes » ont un lien vers la page 4, qui contient alors de l'information essentielle pour l'ensemble des pages de ce réseau et qui semble donc être la page la plus pertinente…

Il reste cependant un problème difficile : le calcul effectif de A^{-1} pour une matrice A qui comporte plusieurs milliards de lignes et de colonnes.
Voyons comment ce problème est résolu (ou plutôt contourné) par les moteurs de recherche.

Partie 2 : une méthode probabiliste pour contourner la difficulté

• Sur Internet, les moteurs de recherche attribuent à chaque page disponible un **indice de pertinence** qui permet de les classer par ordre d'importance. Plusieurs méthodes ont été exposées au chapitre 4 (problème 7, p. 119), à partir du modèle ci-contre.
Nous allons voir ici une méthode probabiliste.

Comment calculer l'indice de pertinence d'une page à l'aide des probabilités ?

A Le principe

Imaginons un internaute qui surfe au hasard de page en page. Quand il est sur une page, il peut soit choisir au hasard l'un des liens disponibles, soit, si aucun lien ne lui plaît, la quitter. Dans ce cas, il choisit au hasard $\left(\text{avec la même probabilité } \frac{1}{7}\right)$ l'une des sept pages disponibles (y compris celle où il est déjà).

On estime ici à 0,14 la probabilité qu'il quitte une page et à 0,86 la probabilité qu'il y choisisse un lien.
Ainsi, par exemple, la probabilité de passer :

- de la page 1 à la page 2 est $0{,}14 \times \frac{1}{7} + 0{,}86 \times \frac{1}{2} = 0{,}45$;
- de la page 4 à la page 3 est $0{,}14 \times \frac{1}{7} = 0{,}02$;
- de la page 7 à la page 6 est $0{,}14 \times \frac{1}{7} + 0{,}86 = 0{,}88$.

Pour tout naturel n, on note P_n la répartition de probabilité entre les sept pages à l'étape n ; ainsi, $P_{n+1} = P_n T$, où T est la matrice de transition.

Matrice de transition T à compléter

$$T = \begin{pmatrix} \dots & 0{,}45 & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & 0{,}02 & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 0{,}88 & \dots \end{pmatrix}$$

Or, la matrice T n'a aucun coefficient nul. Donc d'après le théorème 6, la suite (P_n) converge vers la répartition stable de probabilité P, vérifiant P = PT.
Cela suggère de définir la pertinence d'une page web comme la probabilité qu'elle soit visitée par un internaute surfant au hasard pendant une longue période ; autrement dit, par son coefficient dans P.
Mais au lieu de déterminer P à l'aide de l'égalité P = PT, ce qui imposerait d'inverser la matrice I – T (ici d'ordre 7, mais gigantesque dans la réalité), on peut obtenir des valeurs approchées de ses coefficients en calculant les premiers termes de la suite (P_n).

B Le calcul

1. Pour alléger les calculs, on introduit la matrice carrée K d'ordre 7 dont tous les coefficients sont égaux à 1, et la matrice ligne V de format (1, 7) dont tous les coefficients sont égaux à 1.

a) Démontrez que, pour tout naturel n, $P_n K = V$.

b) Démontrez que la relation de récurrence $P_{n+1} = P_n T$ peut s'écrire $P_{n+1} = P_n A + B$, avec $A = T - 0{,}02K$ et $B = 0{,}02V$.

La matrice A comportant beaucoup de 0, cette relation est plus simple à mettre en œuvre au tableur.

2. On suppose que l'internaute commence par la page 1, autrement dit $P_0 = (1\ 0\ 0\ 0\ 0\ 0\ 0)$. Au tableur, calculez P_n pour les 35 premières pages visitées, en gardant 3 décimales pour les valeurs affichées. Que constatez-vous ?

3. On admet que les valeurs obtenues sont des valeurs approchées à 10^{-2} près des indices de pertinence. Quelles sont ces valeurs ?

Source : Transmaths, T°S spécialité, édition Nathan (2012)