par Jacques Bair

Comment Google classe les pages

Les moteurs de recherche comme Google donnent les sites les plus pertinents par rapport à une recherche. Comment font-ils ? De façon assez surprenante, les puissances de matrices sont au centre de la réponse.

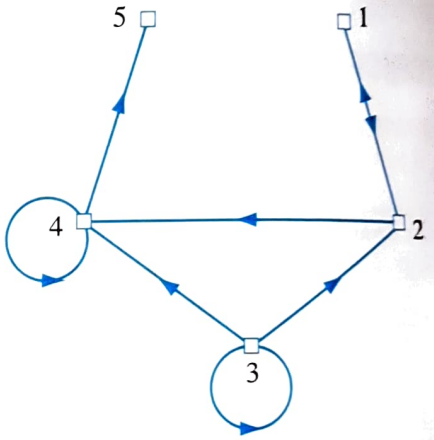
Les deux cofondateurs et le p. d. g de Google (Brin, Schmidt et Page) à la une du Times.

Lorsque vous introduisez des mots-clés dans Google, ce moteur de recherche sélectionne, parmi les pages web sur internet (c'est-à-dire plusieurs centaines de millions), celles qui se rapportent au sujet étudié. C'est ainsi que le mot « tangente » livre environ 1 700 000 résultats (consultation de Google France le 3 juin 2008). De plus, Google classe les pages sélectionnées par ordre décroissant d' « importance ». Les premières pages proposées sont en principe celles qui devraient vous intéresser le plus ; par exemple, la première page sélectionnée par Google pour le mot « tangente » concerne votre revue préférée (adresse électronique : http://tangente.poleditions.com).

Le graphe du Web

Nous nous proposons de vous expliquer simplement comment Google opère son classement. Nous allons exposer la méthode, mise au point il y a une dizaine d'années par deux jeunes informaticiens américains, Larry Page et Sergey Brin, en traitant en détail un exemple fictif d'un système comprenant cinq pages ; le principe, qui exploite la théorie des graphes et l'algèbre linéaire, est le même que dans la réalité… à ceci près que la présentation est alors plus lourde puisque des millions de pages (et non seulement cinq) sont en jeu, ce qui ne peut être traité qu'informatiquement.

La situation étudiée est décrite par le graphe ci-dessous : les nœuds sont les cinq pages 1 à 5, tandis que chaque arc (i, j) traduit le fait que la page i possède un lien (« pointe ») vers j.



Graphe des cinq pages Web. Une flèche entre une page et une autre signifie que la première pointe vers la seconde.

Notons la présence d'une boucle en les nœuds 3 et 4 : il existe un lien de chacune de ces deux pages vers elle-même, par exemple, un lien reliant la fin au début de la page. Par ailleurs, de la page 5 ne part aucun lien, cette page étant, par exemple, composée d'une photo.

Notation de chaque page

L'objectif recherché consiste à classer ces cinq pages de la plus « intéressante » à la moins « intéressante », ce qui se fera en attribuant à chaque page i un score, noté x_i et encore appelé le *PageRank* de i (ce qui se traduit en français par le *rang de la page* … ou le *rang de Page* en l'honneur d'un des concepteurs de la méthode), tel que la page i est plus « intéressante » que j si et seulement si $x_i \geqslant x_j$. Il semble naturel d'apprécier l' « intérêt » d'une page en fonction des arcs qui en partent ou qui y arrivent. Le principe de base pour construire le PageRank des pages repose sur cette règle triple :

Théorie des graphes et algèbre linéaire sont au cœur du moteur de recherche Google.

Larry Page et Sergueï Brin

Le père de Lawrence (dit Larry) Page était professeur d'informatique, celui de Sergueï Brin, mathématicien et spécialiste des systèmes dynamiques. La conjonction des deux se trouve dans la méthode de classement des pages Web.
Le théorème de Perron-Frobenius utilisé par Brin et Page a été découvert bien avant la création de Google, preuve supplémentaire en faveur de ce qui est quelquefois appelé « la déraisonnable efficacité » des mathématiques.

1. Le PageRank d'une page dépend des références qui lui sont faites : plus il y a des liens vers elle, plus grand sera son score. Ceci se traduit par le fait que le PageRang d'une page est la somme de « points » accordés par les pages qui possèdent un lien vers elle : en formule, $x_j = \sum_{i \in E_j} v_{ij}$, où E_j désigne l'ensemble des pages pointant vers j, tandis que v_{ij} est la valeur des points accordés par i à j.
2. Le PageRank d'une page dépend de la notoriété des pages qui la citent : il est d'autant plus élevé que le score des pages qui y font référence est grand. Cette condition s'exprime en supposant chaque valeur v_{ij} proportionnelle à x_i.
3. Le PageRank d'une page varie avec le nombre de liens que possède la page qui la référence : plus le nombre de références venant d'une page est grand, moins importante sera chacune des références en question. Cette exigence se manifeste en supposant chaque valeur v_{ij} inversement proportionnelle au nombre n_i de pages pointées par i.

En résumé, on peut définir le PageRank de p_j par : $x_j = \sum_{i \in E_j} \frac{x_i}{n_i}$.

En appliquant cette formule à notre exemple, nous obtenons un système linéaire, que nous pouvons écrire matriciellement :

$$\begin{cases} x_1 = \frac{x_2}{2} \\ x_2 = x_1 + \frac{x_3}{3} \\ x_3 = \frac{x_3}{3} \\ x_4 = \frac{x_2}{2} + \frac{x_3}{3} + \frac{x_4}{2} \\ x_5 = \frac{x_4}{2} \end{cases} \Leftrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & 0 & 0 & 0 \\ 1 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{3} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_2 \end{pmatrix} \Leftrightarrow X = RX$$

Problème mathématique

Le problème mathématique consiste à trouver un vecteur X solution du système linéaire homogène écrit sous la forme matricielle X = R X. La seule solution du système proposé est le vecteur nul, qui n'est évidemment pas acceptable. Il faut modifier ce système, en remplaçant la matrice R par une nouvelle matrice P obtenue en modifiant R aussi peu que possible pour tenir compte au maximum de la règle donnée pour construire le PageRank de chaque page, de manière à obtenir un vecteur X non nul solution du nouveau système X = P X . Mais, ce système possède forcément une infinité de solutions, puisque tout multiple non nul d'une solution en est encore une. C'est pourquoi, nous conviendrons de choisir une solution X « normalisée », c'est-à-dire telle que la somme $\sum_{1 \leq j \leq 5} x_j$ des composantes est égale à 1. Nous procédons à cette modification en deux étapes.

Matrice stochastique

Nous remarquons que toutes les colonnes de R, sauf la dernière, sont des vecteurs de probabilité, c'est-à-dire que ses composantes sont des nombres non négatifs dont la somme est égale à 1. Nous remplaçons la dernière colonne (nulle) de R par une colonne de nombres tous égaux à 1/5. Nous obtenons ainsi une matrice S. De la sorte, le système X = S X possède une et une seule solution normalisée :

$$x_1 = \frac{7}{46}, x_2 = \frac{5}{23}, x_3 = \frac{3}{46}, x_4 = \frac{8}{23}, x_5 = \frac{5}{23}.$$

Le remplacement d'une colonne de 0 par une colonne de nombres égaux consiste à ajouter fictivement des liens de la page 5 vers chacune des cinq pages. Dans le cas général de n pages, on change une colonne de 0 par une colonne de nombres tous égaux à l'inverse de n, ce qui est négligeable puisque n est grand. Une telle modification a pour effet de construire une matrice qualifiée de « stochastique suivant les colonnes » en ce sens que toutes ses colonnes sont des vecteurs de probabilité.
Une telle matrice possède toujours 1 comme valeur propre, de sorte que l'existence d'une solution normalisée du système X = S X est garantie. Toutefois, l'unicité d'une solution normalisée n'est pas assurée pour toute matrice stochastique, ainsi qu'en atteste le cas défini par le nouveau système

$$x_1 = x_2, x_2 = x_1, x_3 = x_4, x_4 = x_3, x_5 = x_5.$$

qui admet les solutions normalisées

$$x_1 = x_2 = \frac{1}{2}, x_3 = x_4 = x_5 = 0 \text{ et } x_1 = x_2 = x_5 = 0, x_3 = x_4 = \frac{1}{2}.$$

Une façon de garantir à la fois l'existence et l'unicité d'une solution normalisée pour un tel système consiste à modifier légèrement la matrice stochastique de manière à ce qu'elle reste stochastique, mais avec tous ses éléments strictement positifs. Effectuons cette seconde transformation sur notre exemple-type du début.

Matrice stochastique strictement positive

La matrice S est remplacée par la matrice P, notée ainsi en souvenir de L. Page, définie par

$$P = 0{,}85S + 0{,}15M$$

où M est la matrice de 5 lignes et de 5 colonnes dont tous les éléments sont égaux à 1/5. En d'autres termes, P est une « moyenne » de S, la matrice stochastique issue de la règle de départ, et de M, la matrice correspondant au cas où toute page est reliée par un lien à toute page, cette moyenne étant pondérée de manière à ce que le poids le plus important (ici 0,85) soit accordé au premier terme.
La matrice P est toujours stochastique, et de plus strictement positive par construction. Le théorème de Perron-Frobenius (voir l'article sur le sujet dans ce numéro) peut lui être appliqué : il assure que le système X = P X possède une et une seule solution normalisée X.

Calcul d'un vecteur propre

Celle-ci peut être obtenue en constatant que $X = PX = P(PX) = P^2X = P^2(PX) = P^3X = \ldots = P^kX = \ldots$ quel que soit l'exposant entier positif k. Or, d'après le même théorème, la suite des puissances P^k possède, lorsque k tend vers l'infini, une limite P, qui est une matrice dont toutes les colonnes sont égales au vecteur normalisé X trouvé ci-dessus.
La solution de notre problème peut donc être obtenue en élevant la matrice P à des puissances toujours plus élevées : lorsque les colonnes se stabilisent, les scores des différentes pages sont obtenus et le classement attendu peut être établi.
Ainsi, pour notre exemple de départ, on trouve successivement (grâce à un ordinateur) :

$$P = \begin{pmatrix} 0{,}03 & 0{,}455 & 0{,}03 & 0{,}03 & 0{,}2 \\ 0{,}88 & 0{,}03 & 0{,}313333 & 0{,}03 & 0{,}2 \\ 0{,}03 & 0{,}03 & 0{,}313333 & 0{,}03 & 0{,}2 \\ 0{,}03 & 0{,}455 & 0{,}313333 & 0{,}455 & 0{,}2 \\ 0{,}03 & 0{,}03 & 0{,}03 & 0{,}455 & 0{,}2 \end{pmatrix}$$

$$P^{20} = \begin{pmatrix} 0,160258 & 0,160207 & 0,160235 & 0,16023 & 0,160225 \\ 0,225802 & 0,22587 & 0,225833 & 0,22584 & 0,225846 \\ 0,0896481 & 0,0896451 & 0,0896468 & 0,0896465 & 0,0896462 \\ 0,32285 & 0,32282 & 0,322836 & 0,322834 & 0,322831 \\ 0,201441 & 0,201457 & 0,201449 & 0,22145 & 0,2201452 \end{pmatrix}$$

$$P^{50} = \begin{pmatrix} 0,160229 & 0,160229 & 0,160229 & 0,160229 & 0,160229 \\ 0,225841 & 0,225841 & 0,225841 & 0,225841 & 0,225841 \\ 0,0896464 & 0,0896464 & 0,0896464 & 0,0896464 & 0,0896464 \\ 0,322833 & 0,322833 & 0,322833 & 0,322833 & 0,322833 \\ 0,201451 & 0,201451 & 0,201451 & 0,201451 & 0,201451 \end{pmatrix}$$

On constate que l'élévation de P à une puissance supérieure à 50 ne change pas le dernier résultat. En conséquence, les scores recherchés sont donnés par

$$x_1 = 0,160229 \; ; \; x_2 = 0,225841 \; ; \; x_3 = 0,0896464 \; ; \; x_4 = 0,322833 \; ; \; x_5 = 0,201451.$$

Le classement de nos cinq pages est donc le suivant : la page 4 occupe la première position, suivie de la page 2, puis successivement 5, 1 et enfin 3.

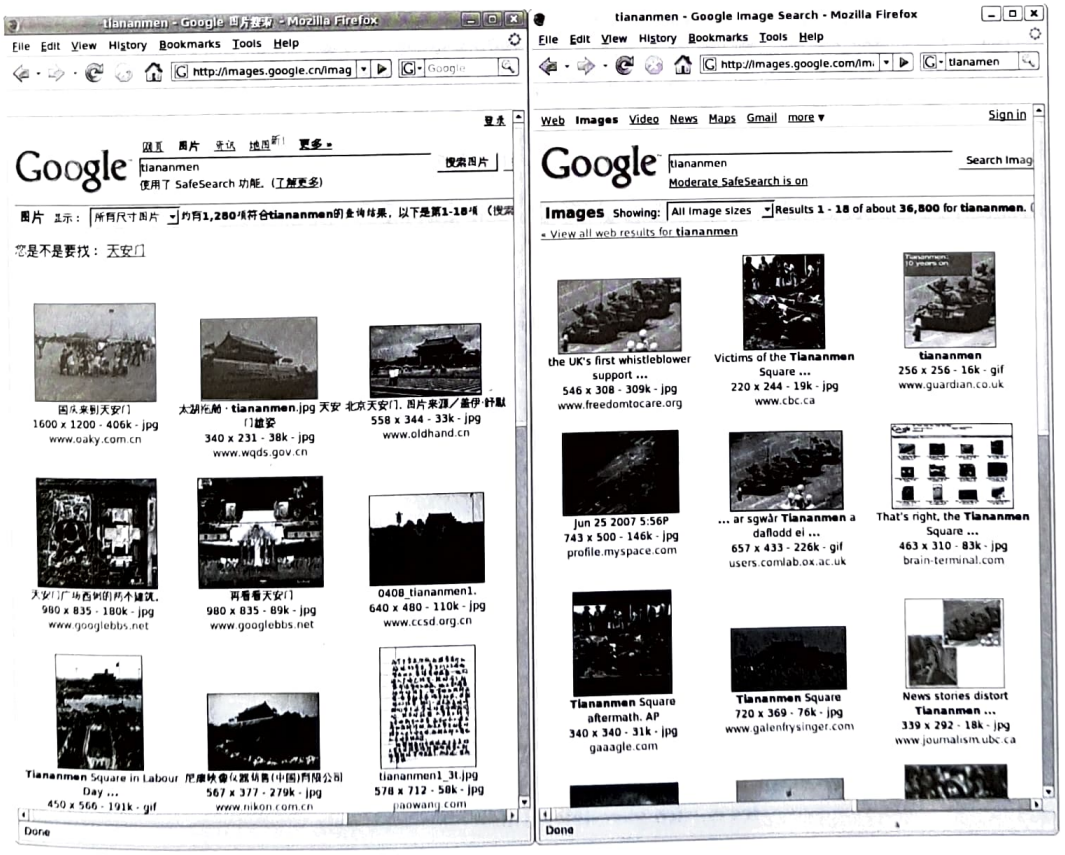
Interprétation probabiliste

Terminons en donnant une interprétation probabiliste de ce qui précède. Chaque élément p_{ij} de la matrice P peut être regardé comme étant la probabilité pour qu'un surfeur passe, en une unité de temps, de la page i à la page j. Ainsi, ce surfeur, se trouvant en un instant déterminé sur la page i, va choisir j une unité de temps plus tard, et ceci de la manière suivante :

- avec une probabilité de 0,85, il choisit au hasard un des liens de la page i
- avec une probabilité de 0,15, il choisit au hasard n'importe quelle page du web.

L'élément se situant sur la i-ème ligne et la j-ème colonne de P^2 représente la probabilité pour que le surfeur, étant initialement sur la page i, se retrouve sur la page j deux unités de temps plus tard. De même, l'élément se situant sur la i-ème ligne et la j-ème colonne de P peut être regardé comme étant la probabilité pour que le surfeur passe de la page i à la page j en un temps « infiniment long » ; cette probabilité peut encore être assimilée à la fréquence théorique vers laquelle tend la fréquence observée de cet événement lorsque le nombre de répétitions de cette expérience aléatoire tend vers l'infini.

J. B.



L'algorithme de classement des pages n'est pas le même en Chine et dans le reste du monde.