L'algorithme PageRank de Google

Paul Jolissaint

Un moteur de recherche tel que Google doit faire trois choses :

1. Naviguer sur le web et localiser (si possible) toutes les pages ayant un accès public.
2. Indexer les données ci-dessus pour qu'elles puissent être retrouvées efficacement par mots-clés ou groupes de mots significatifs.
3. Classer l'importance de chaque page dans la base de données de sorte que lorsqu'un internaute fait une recherche et que le sous-ensemble des pages correspondantes dans la base de données a été trouvé, **les pages les plus importantes soient présentées en premier.**

Le but de l'algorithme PageRank est de résoudre le troisième problème ci-dessus, et nous allons expliquer comment on procède pour élaborer le classement des pages web avec cet algorithme.

L'ensemble des pages web peut être modélisé par un **graphe orienté** : chaque page est représentée par un point (appelé **sommet** du graphe) et on dessine une flèche d'un sommet A vers un sommet B si et seulement si la page A contient un lien hypertexte *vers* la page B. Une telle flèche s'appelle une **arête** (orientée).

Notons que le nombre de pages dans le web est estimé à plus de 60 milliards, et Google en répertorie environ 35 milliards. Aussi, si la partie du web indexée par Google contient n pages, on convient de représenter chaque page par un numéro i compris entre 1 et n, et on désignera par x_i le **score** de la page i. Il s'agit d'un nombre compris entre 0 et 1 qui doit indiquer l'importance de la page i au sens du point 3 ci-dessus : plus x_i est élevé, plus l'URL de la page i apparaît tôt dans la liste que propose le moteur de recherche lorsque la page i contient les mots-clés demandés. Comme première condition, on doit avoir $x_i > x_j$ si la page i est plus importante que la page j. Une approche simple consisterait à définir x_i de la façon suivante : notons m_i le nombre de liens vers la page i, alors x_i pourrait être défini par le rapport m_i/n.

Cette formule ignore toutefois un point crucial : *une page est importante si d'autres pages importantes pointent vers elle.* Ainsi, si la page j est importante (c'est-à-dire si x_j est grand) et s'il existe un lien de la page j vers la page i, cela donne d'autant plus d'importance à la page i, donc l'établissement d'un tel lien devrait augmenter x_i davantage que si on établit un lien d'une page sans importance vers la page i. Toutefois, on veut éviter que des pages sans grande importance obtiennent un grand score par des moyens artificiels comme par exemple la création d'un grand nombre de pages fictives contenant chacune un lien vers la page en question. Ainsi, si la page j contient n_j liens vers d'autres pages, dont la page i, on veut que la contribution de la page j au score x_i de la page i soit égal à x_j/n_j et non pas égal à x_j.

Les concepteurs de Google, Larry Page et Sergey Brin, alors étudiants en informatique à Stanford, ont adopté en 1998 la définition suivante de la suite des scores $x_1, x_2, \ldots, x_n$: rappelons que n_j désigne le nombre de liens contenus dans la page j pour tout j (donc issus de la page j vers d'autres pages, et on convient qu'un lien d'une page vers elle-même est ignoré), et soit $L_i \subset \{1,\ 2,\ \ldots,\ n\}$ l'ensemble des pages qui ont un lien vers la page i. On demande alors que pour tout i,

$$x_i = \sum_{j \in L_i} \frac{x_j}{n_j}$$

et que les composantes de $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} =: (x_1, \ldots, x_n)^T$ soient normalisées de sorte que $\sum_i x_i = 1$.

(Si M est une matrice quelconque, on note M^T sa transposée.)

Le problème avec la définition ci-dessus est que le calcul de x_i fait intervenir les autres x_j qui sont a priori eux aussi inconnus. On peut par exemple calculer des valeurs approximatives des x_i en utilisant une méthode itérative : notons $x_i^{(k)}$ la valeur de x_i après la k-ième itération. On initialise le processus en posant par exemple $x_i^{(0)} = 1/n$ pour tout i (ce qui s'interprète en stipulant qu'au départ toutes les pages possèdent le même score), puis en définissant $x_i^{(k+1)} = \sum_{j \in L_i} \frac{x_j^{(k)}}{n_j}$ pour tout $1 \leq i \leq n$ et pour tout $k \geq 0$. On simplifie les notations en introduisant la matrice A qui est définie comme suit : pour $1 \leq j \leq n$, on remplit la colonne j en posant $\frac{1}{n_j}$ dans la ligne i si $j \in L_i$ (c'est-à-dire s'il y a un lien de j vers i) et 0 sinon, et on note comme ci-dessus $\vec{x}^{(k)}$ le vecteur dont les composantes sont les $x_i^{(k)}$.

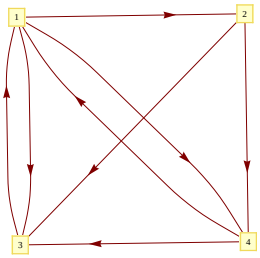
De la sorte, on a

$$A\vec{x} = \vec{x} \quad \text{et} \quad A\vec{x}^{(k)} = \vec{x}^{(k+1)}\ \forall k \geq 0.$$

Il semble que Google recalcule les scores une fois par semaine, en itérant un processus proche de celui décrit ci-dessus car ce dernier pose quelques problèmes que nous allons décrire plus loin.

En attendant, voici un exemple simple qui permet de se faire une idée de la situation :

Exemple. Considérons le web donné par le graphe suivant :



Cela conduit au système linéaire :

$$\begin{cases} x_1 = x_3/1 + x_4/2 \\ x_2 = x_1/3 \\ x_3 = x_1/3 + x_2/2 + x_4/2 \\ x_4 = x_1/3 + x_2/2. \end{cases}$$

On pose $\vec{x} = (x_1, x_2, x_3, x_4)^T$ et la matrice du système est $A = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$ de sorte que $A\vec{x} = \vec{x}$.

En résumé, on cherche un vecteur $\vec{x}$ comme ci-dessus dont les composantes x_j sont positives, de somme 1, et qui satisfait l'équation $A\vec{x} = \vec{x}$. Elle admet une solution : c'est un multiple du vecteur $\vec{v} = (12, 4, 9, 6)^T$, donc $\vec{x} = (\frac{12}{31}, \frac{4}{31}, \frac{9}{31}, \frac{6}{31})^T \approx (0.387, 0.129, 0.290, 0.194)^T$.

Remarque importante. Dans la description ci-dessus, on n'a pas tenu compte des "pages cul-de-sac" (en anglais : "dangling nodes") : ce sont les pages qui n'ont aucun lien vers une autre page, donc ce sont les pages j pour lesquelles $n_j = 0$. Brin et Page ont résolu le problème en remplaçant chaque colonne correspondant à une page cul-de-sac par la colonne $(1/n, \ldots, 1/n)^T$. On interprète cette modification ainsi : lorsqu'un internaute visite une page cul-de-sac, la probabilité de visiter les autres pages du web en quittant celle-ci est uniforme.

Les équations ci-dessus posent un certain nombre de questions :
- l'équation $A\vec{x} = \vec{x}$ avec $\sum_i x_i = 1$ admet-elle une solution avec $x_i > 0$ pour tout i ?
- si une solution à l'équation ci-dessus existe, est-elle unique ?
- dans le calcul de la suite $(\vec{x}^{(k)})$ qui est censée approcher $\vec{x}$, le processus converge-t-il ? et sous quelles conditions ?

Afin de garantir l'existence et l'unicité de $\vec{x}$, et la convergence du processus, Brin et Page ont modifié l'équation ci-dessus : le vecteur de score $\vec{x}$ est la solution de l'équation

$$\boxed{\vec{x} = ((1-m)A + mS)\,\vec{x}}$$

où $0 < m < 1$ est un paramètre et S est la matrice $n \times n$ dont tous les coefficients sont égaux à $1/n$. On démontrera que cette équation admet toujours une unique solution $\vec{x}$ telle que $\sum_i x_i = 1$ et que tous les x_i sont strictement positifs. En 2005, Google utilisait la valeur $m = 0.15$. Nous ignorons si c'est encore la valeur utilisée aujourd'hui ; de plus, les diverses valeurs des x_i ne sont pas publiques.

Afin de simplifier les notations, posons $(1-m)A + mS =: M = \begin{pmatrix} m_{1,1} & \ldots & m_{1,n} \\ \vdots & \ddots & \vdots \\ m_{n,1} & \ldots & m_{n,n} \end{pmatrix}$. Elle possède deux propriétés cruciales : elle est stochastique par rapport à ses colonnes, *i.e.* la somme des coefficients de chaque colonne vaut 1, et tous les coefficients $m_{i,j} > 0$. On désigne par $V_1(M) = \{\vec{x} \in \mathbb{R}^n | M\vec{x} = \vec{x}\}$ le sous-espace propre associé à la valeur propre 1 de M. Remarquons que $V_1(M)$ n'est pas réduit à $\vec{0}$: en effet, les valeurs propres de M sont exactement les mêmes que celle de sa transposée M^T, et cette dernière admet 1 comme valeur propre avec le vecteur propre constant $(1, \ldots, 1)^T$, puisque M est stochastique. La positivité stricte de M admet la conséquence importante suivante :

Proposition 1 *Soit M comme ci-dessus. Alors tout vecteur propre dans $V_1(M)$ a ses composantes soit toutes positives, soit toutes négatives.*

Preuve. Observons d'abord que si $\vec{y}$ est un vecteur à n composantes $y_1,\ y_2,\ \ldots\ y_n$, alors $|\sum_i y_i| \leq \sum_i |y_i|$ et que l'inégalité est stricte si et seulement si certains y_i sont positifs et d'autres négatifs. Supposons alors par l'absurde qu'il existe $\vec{x} \in V_1(M)$ dont les composantes ont des signes mélangés (certaines positives et d'autres négatives). De l'équation $\vec{x} = M\vec{x}$, on déduit que $x_i = \sum_j m_{i,j}x_j$ pour tout i, et comme $m_{i,j} > 0$ pour tous i et j, on a :

$$|x_i| = \left|\sum_{j=1}^{n} m_{i,j}x_j\right| < \sum_j m_{i,j}|x_j|.$$

En sommant ces inégalités sur i de 1 à n, puis en intervertissant les sommes sur i et sur j on obtient

$$\sum_{i=1}^{n}|x_i| < \sum_{i=1}^{n}\sum_{j=1}^{n} m_{i,j}|x_j| = \sum_{j=1}^{n}\left(\sum_{i=1}^{n} m_{i,j}\right)|x_j| = \sum_{j=1}^{n}|x_j|$$

qui donne une contradiction. □

Proposition 2 *Soient $\vec{v}$ et $\vec{w}$ deux vecteurs linéairement indépendants dans $\mathbb{R}^n$. Alors il existe une valeur réelle s telle que le vecteur $\vec{x} = \vec{v} + s\vec{w}$ ou le vecteur $\vec{x} = s\vec{v} + \vec{w}$ admet des composantes positives et des composantes négatives.*

Preuve. Comme $\vec{v}$ et $\vec{w}$ sont linéairement indépendants, ils sont tous deux non nuls. Posons $d = \sum_i v_i$. Si $d = 0$, alors $\vec{v}$ possède des composantes de chaque signe, et il suffit de prendre $s = 0$ et donc de poser $\vec{x} = \vec{v}$. Si $d \neq 0$, posons $s = -\frac{\sum_i w_i}{d}$ et $\vec{x} = s\vec{v}+\vec{w}$, de sorte que $\sum_i x_i = 0$ et $\vec{x} \neq \vec{0}$. Par suite, $\vec{x}$ a nécessairement des composantes positives et des composantes négatives. □

Corollaire 3 *Si M est une matrice stochastique et positive, alors $V_1(M)$ est de dimension 1 et formé des multiples d'un vecteur $\vec{x}$ dont les composantes $x_i > 0$ pour tout i et telles que $\sum_i x_i = 1$.*

Preuve. Supposons par l'absurde que $\dim(V_1(M)) \geq 2$, et soient $\vec{v}$ et $\vec{w}$ deux vecteurs linéairement indépendants dans $V_1(M)$. Alors le vecteur $\vec{x}$ construit dans la proposition 2 fournit une contradiction à la proposition 1. □

Enfin, il reste à montrer comment on procède pratiquement pour trouver le vecteur $\vec{x}$.

Soit M la matrice ci-dessus. L'idée pour calculer le vecteur $\vec{x}$ tel que $\sum_i x_i = 1$ et $M\vec{x} = \vec{x}$ consiste à choisir un vecteur initial arbitraire $\vec{x}^{(0)}$ à composantes positives et de somme 1, puis à définir la suite de vecteurs $\vec{x}^{(k)} = M\vec{x}^{(k-1)}$ et de démontrer que cette suite converge vers le vecteur cherché.

Notons V le sous-espace vectoriel de $\mathbb{R}^n$ formé des vecteurs $\vec{w}$ dont la somme des composantes $\sum_i w_i = 0$. C'est, pour le produit scalaire standard, l'orthogonal du vecteur $\vec{e} = (1, \ldots, 1)^T$. On munit également $\mathbb{R}^n$ de sa **1-norme** : $\|\vec{v}\|_1 := \sum_{i=1}^{n} |v_i|$.

Proposition 4 *Posons $c = \max_{1\leq j\leq n} |1-2\cdot\min_{1\leq i\leq n} m_{i,j}|$. Alors $0 < c < 1$, et on a $M\vec{v} \in V$ et $\|M\vec{v}\|_1 \leq c\|\vec{v}\|_1$ pour tout $\vec{v} \in V$.*

Preuve. Pour $\vec{v} \in V$ fixé, posons $\vec{w} = M\vec{v}$, de sorte que $w_i = \sum_j m_{i,j}v_j$. Alors

$$\sum_i w_i = \sum_i\sum_j m_{i,j}v_j = \sum_j v_j\left(\sum_i m_{i,j}\right) = \sum_j v_j = 0$$

puisque la matrice M est stochastique et que $\sum_j v_j = 0$. Ainsi, $\vec{w} = M\vec{v} \in V$. Pour prouver l'inégalité, notons $e_i = \mathrm{sgn}(w_i) = \pm 1$ le signe de w_i, de sorte que

$$\|\vec{w}\|_1 = \sum_i e_i w_i = \sum_i e_i\left(\sum_j m_{i,j}v_j\right).$$

Observons que les e_i n'ont pas tous le même signe car $\vec{w} \in V$ (sauf bien sûr au cas où $\vec{w} = \vec{0}$, mais alors l'inégalité est trivialement vraie). Posons encore $a_j = \sum_i e_i m_{i,j}$ et intervertissons la double somme ci-dessus pour obtenir

$$\|\vec{w}\|_1 = \sum_j v_j\left(\sum_i e_i m_{i,j}\right) = \sum_j a_j v_j.$$

Puisque les signes des $e_i = \pm 1$ varient, que $e_i + 1 = 0$ ou 2 et que $\sum_i m_{i,j} = 1$, avec $0 < m_{i,j} < 1$, on obtient d'une part

$$a_j + 1 = \sum_i (e_i + 1) m_{i,j} \geq 2 \cdot \min_i m_{i,j}$$

et d'autre part, puisque $e_k - 1 = -2$ ou 0 et que $m_{k,j} \geq \min_i m_{ij}$, on a

$$a_j - 1 = \sum_k (e_k - 1) m_{k,j} \leq -2 \cdot \min_i m_{i,j}.$$

Cela implique immédiatement que

$$-1 < -1 + 2 \cdot \min_i m_{i,j} \leq a_j \leq 1 - 2 \cdot \min_i m_{i,j} < 1.$$

Ainsi, $|a_j| \leq |1 - 2 \cdot \min_i m_{i,j}| \leq c < 1$. On a finalement

$$\|\vec{w}\|_1 = \sum_j a_j v_j \leq \left| \sum_j a_j v_j \right| \leq \sum_j |a_j||v_j| \leq c \sum_j |v_j| = c\|\vec{v}\|_1,$$

ce qui prouve la proposition. □

Voici enfin le résultat qui garantit la convergence de la suite des $\vec{x}^{(k)}$:

Théorème 5 *Soit M une matrice comme ci-dessus. Alors elle admet un unique vecteur $\vec{q} \in V_1(M)$ à composantes toutes positives et tel que $\|\vec{q}\|_1 = 1$. Il peut être calculé par*

$$\vec{q} = \lim_{k \to \infty} M^k \vec{x}^{(0)}$$

à partir de n'importe quel vecteur initial $\vec{x}^{(0)}$ à composantes positives et tel que $\|\vec{x}^{(0)}\|_1 = 1$. La convergence de $M^k \vec{x}^{(0)}$ vers $\vec{q}$ a lieu par rapport à la 1-norme.

Preuve. On sait déjà que $\vec{q}$ existe et est unique. Une fois le vecteur intial $\vec{x}^{(0)}$ choisi tel que $x_j^{(0)} > 0$ pour tout j et $\|\vec{x}^{(0)}\|_1 = 1$, posons $\vec{v} = \vec{x}^{(0)} - \vec{q}$ de sorte que $\vec{x}^{(0)} = \vec{q} + \vec{v}$ et $\vec{v}$ appartient à V, c'est-à-dire que la somme des composantes de $\vec{v}$ vaut 0. On a alors pour tout k :

$$M^k \vec{x}^{(0)} = M^k \vec{q} + M^k \vec{v} = \vec{q} + M^k \vec{v}$$

car $M\vec{q} = \vec{q}$ donc $M^k \vec{q} = \vec{q}$ pour tout $k > 0$ et

$$M^k \vec{x}^{(0)} - \vec{q} = M^k \vec{v}.$$

Par une récurrence évidente, on obtient, par la proposition 4 : $\|M^k \vec{v}\|_1 \leq c^k \|\vec{v}\|_1$ pour tout k, et comme $0 < c < 1$, on a

$$\lim_{k \to \infty} \|M^k \vec{v}\|_1 = 0$$

ce qui permet de conclure que

$$\vec{q} = \lim_{k \to \infty} M^k \vec{x}_0.$$

□

Remarque finale. Dans le cas de la matrice de Google, le calcul direct de la suite $(\vec{x}^{(k)})_{k \geq 1}$ par récurrence comme présenté ci-dessus serait malgré tout très fastidieux puisque chaque coefficient de $\vec{x}^{(k)}$ nécessiterait des milliards de multiplications et d'additions. En fait, le calcul est raisonnable et efficace grâce à la forme spéciale de la matrice M ; comme on l'a vu, $M = (1-m)A + mS$ et A est une matrice creuse, c'est-à-dire qu'elle contient beaucoup de 0. On observe alors la particularité suivante de S : Soit $\vec{v} \in \mathbb{R}^n$ un vecteur dont la somme des composantes vaut 1. Alors $S\vec{v} = (\frac{v_1 + \cdots + v_n}{n}, \ldots, \frac{v_1 + \cdots + v_n}{n})^T = (1/n, \ldots, 1/n)^T =: \vec{u}$.

Par suite, quel que soit k, on a $mS\vec{v}^{(k)} = m\vec{u}$ qui est *indépendant* de k et ainsi la récurrence est en réalité :

$$\vec{x}^{(k+1)} = (1 - m)A\vec{x}^{(k)} + m\vec{u}$$

qui peut se calculer plus facilement puisque le calcul de chaque composante de $\vec{x}^{(k+1)}$ ne fait intervenir que quelques dizaines de composantes de $\vec{x}^{(k)}$. On peut démontrer qu'une cinquantaine d'itérations suffisent pour obtenir un vecteur de score satisfaisant.

Lycée cantonal de Porrentruy
pajolissaint@sunrise.ch