L'algorithme PageRank

Lionel Fourquaux

1er février 2017

L'algorithme « PageRank » (nommé d'après son inventeur Larry Page) est utilisé pour classer par importance des documents d'après les liens de ces documents entre eux. Il peut s'agir aussi bien de liens hypertextes entre pages web, de citations d'articles scientifiques, ou de toute autre forme de graphe. C'est l'un des principaux moyens de classement utilisé par le moteur de recherche Google, fondé par Larry Page et Sergey Brin.

Le mot « PageRank » lui-même est une marque déposée par Google. Le procédé est breveté par l'université Stanford, qui en a accordé une licence exclusive à Google.

1 Déplacements sur un graphe, probabilités et matrices

Considérons un graphe fini, orienté ou non, et un point qui se déplace aléatoirement sur ce graphe, la probabilité d'emprunter chacune des arêtes partant du nœud étant fixée. On obtient ainsi une chaîne de Markov homogène (particulièrement simple puisque l'espace de valeurs est fini), dont le comportement se lit sur les puissances de la matrice des probabilités de transition d'un nœud à l'autre.

Numérotons les nœuds de 1 à n, et notons $a_{i,j}$ (pour $1 \leqslant i, j \leqslant n$) la probabilité de passer du nœud j au nœud i. Cette probabilité est en particulier nulle s'il n'y a pas d'arête allant de j vers i. Si $x = (x_i)_{1 \leqslant i \leqslant n}$ donne la probabilité d'être au nœud i à l'instant 0, alors la même probabilité après m déplacements est donnée par le vecteur $A^m x$.

Notons que pour tout j, on a

$$\sum_{i=1}^{n} a_{i,j} = 1,$$

donc la transposée de la matrice $A = (a_{i,j})$ a 1 pour valeur propre, et la matrice A a aussi 1 pour valeur propre.

Considérons la norme définie par $\|x\|_1 = \sum_{i=1}^{n} |x_i|$. On a $\|Ax\|_1 \leqslant \|x\|_1$.

On en déduit donc l'existence d'une distribution de probabilité *stationnaire* sur le graphe, c'est-à-dire des réels $0 \leqslant p_i \leqslant 1$ tels que

$$\sum_{j=1}^{n} a_{i,j} p_j = p_i \qquad \text{et} \qquad p_1 + \cdots + p_n = 1.$$

Concernant les autres valeurs propres, notons qu'elles sont toutes de module au plus 1, et plus précisément qu'elles se trouvent dans le disque fermé de centre $\min a_{i,i}$ ayant 1 sur son bord.

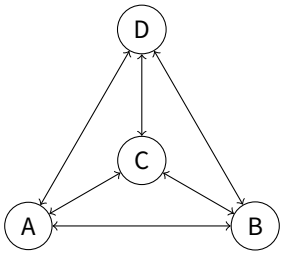
Une telle matrice est appelée matrice stochastique à gauche, et ces propriétés sont précisées par le théorème de Perron-Frobenius :

Théorème 1.1. *Soit $A = (a_{i,j})$ une matrice $n \times n$ à coefficients réels positifs, irréductible, c'est-à-dire telle que le graphe dont les nœuds numérotés par les entiers $1, ..., n$ et dont les arêtes sont données par les couples (i, j) tels que $a_{i,j} \neq 0$ est connexe. Alors on a les propriétés suivantes.*

(i) Le rayon spectral r de la matrice A, c'est-à-dire le maximum des modules de ses valeurs propres complexes, est une valeur propre de A.
(ii) Si A est à coefficients strictement positifs, c'est une valeur propre simple.
(iii) Il y a un vecteur propre de A associé à la valeur propre r dont les composantes sont des réels positifs.
(iv) Si A a exactement p valeurs propres complexes de module r, alors l'ensemble des valeurs propres complexes de A est stable par rotation d'angle $2\pi/p$ autour de l'origine.

2 Quelques exemples de graphes

Considérons le graphe des sommets et arêtes d'un tétraèdre :

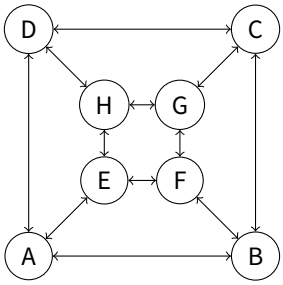


En donnant à chaque arête la même probabilité d'être empruntée, on trouve la matrice de transition suivante :

$$A_{\text{tétraèdre}} = \begin{pmatrix} 0 & 1/3 & 1/3 & 1/3 \\ 1/3 & 0 & 1/3 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}$$

et un calcul des puissances de cette matrice donne que si l'on est au nœud A au temps 0, alors la probabilité d'être en A au temps n est $\frac{1-(-3)^{1-n}}{4}$ et celle d'être en B (ou C, ou D) est $\frac{1-(-3)^{-n}}{4}$. Quand n tend vers l'infini, elle converge vers la distribution de probabilité stationnaire, i.e. $1/4$ pour chacun des quatre sommets, la vitesse de convergence étant contrôlée par la valeur propre de plus grand module après 1.

Il n'y a cependant pas toujours convergence. Par exemple, dans le cas du graphe d'un cube, avec une même probabilité pour chaque arête :



il est assez facile de montrer qu'il n'y a en général pas convergence vers la distribution stationnaire ($1/8$ partout). En particulier, si l'on est au nœud A au temps 0, un argument de parité montre que l'on ne peut pas être sur B, D, E ou G au temps $2n$. La matrice de transition a alors des valeurs propres de module 1 autres que 1 (en l'occurence, -1). (Le cube est le seul polyèdre régulier pour lequel ce phénomène se produit).

En revanche, si les coefficients diagonaux de la matrice de transition sont non nuls, alors 1 est la seule valeur propre de module 1, et l'on est assuré de la convergence vers une distribution de probabilité stationnaire.

3 L'algorithme « PageRank »

L'idée clé du PageRank est de mesurer l'« importance » d'un nœud d'un graphe à partir des arêtes qui mènent à ce nœud et de l'importance des nœuds sources de ces arêtes. Pour cela, on peut considérer une marche aléatoire sur le graphe, les arêtes partant d'un nœud donné ayant la même probabilité d'être empruntées. (Dans le cas d'un nœud n'ayant aucune arête sortante, on peut faire comme s'il avait des arêtes pointant vers tous les autres nœuds, ou supprimer ces nœuds du graphe). La probabilité attribuée à chaque nœud dans la distribution stationnaire mesure alors l'importance du nœud.

Prenons par exemple les cinq villes ayant la plus grande population en France, et le nombre de trains directs relevés sur une journée :

vers :	Paris	Marseille	Lyon	Toulouse	Nice
Paris	0	15	23	3	5
Marseille	16	0	18	5	16
Lyon	20	18	0	2	4
Toulouse	6	5	3	0	3
Nice	4	12	4	2	0

On trouve la matrice de transition suivante :

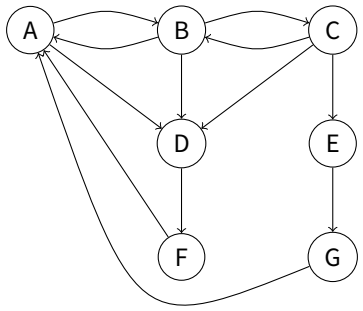
$$\begin{pmatrix} 0 & 16/55 & 5/11 & 6/17 & 2/11 \\ 15/46 & 0 & 9/22 & 5/17 & 6/11 \\ 1/2 & 18/55 & 0 & 3/17 & 2/11 \\ 3/46 & 1/11 & 1/22 & 0 & 1/11 \\ 5/46 & 16/55 & 1/11 & 3/17 & 0 \end{pmatrix}$$

La distribution stationnaire (« PageRank ») obtenue est :

Paris	Marseille	Lyon	Toulouse	Nice
0.249	0.284	0.255	0.067	0.145

ce qui correspond (assez grossièrement : Paris et Toulouse se retrouvent plus bas qu'on aurait pu l'attendre) au classement par population.

Notons qu'il ne s'agit pas simplement d'une manière de comptabiliser le nombre d'arêtes pointant vers un nœud donné : le PageRank des nœuds sources est aussi pris en compte, comme on le voit sur l'exemple suivant.



PageRank :

A	B	C	D	E	F	G
$16/54$	$9/54$	$3/54$	$12/54$	$1/54$	$12/54$	$1/54$

On peut constater que le nœud F a un plus grand PageRank que le nœud G, alors que tous les deux ont une arête entrante et une sortante.

Sous cette forme, l'algorithme présente cependant un certain nombre d'inconvénients :

- si le graphe n'est pas connexe, la distribution de probabilité initiale influe sur le résultat, en particulier en déterminant quelles composantes sont explorées ;
- dans le cas d'arêtes orientés, si l'on peut s'échapper d'une partie du graphe mais pas y revenir, alors tous les nœuds de cette partie auront une probabilité nulle dans la distribution stationnaire ;
- d'autre part, se pose le problème du calcul du PageRank dans des situations réelles, sur des graphes ayant des milliards de nœuds : déterminer les vecteurs propres d'une matrice (creuse) aussi grande serait une lourde tâche pour les algorithmes usuels de résolution de systèmes linéaires.

Ces problèmes se trouvent simultanément résolus par l'introduction d'un *facteur d'amortissement*. Au lieu de considérer la matrice de transition A précédente, on utilise la matrice

$$dA + \frac{1-d}{n}U,$$

où U est la matrice carrée de taille n dont tous les coefficients valent 1, avec un *facteur d'amortissement* $0 < d < 1$. Cela correspond à modifier le processus de déplacement sur un graphe en introduisant une probabilité $1-d$ de « sauter » vers un nœud pris au hasard (équiprobablement) plutôt que d'emprunter les arêtes. En particulier, tous les nœuds peuvent alors être atteints.

D'autre part, la seule valeur propre de module 1 de la matrice est 1, et c'est une valeur propre simple, ce qui montre la convergence vers une unique distribution de probabilité stationnaire.

Enfin, on peut exploiter cette convergence (maintenant garantie) pour calculer la distribution de probabilité stationnaire : on part d'une distribution de probabilité arbitraire, $x^{(0)}$, et l'on considère la suite $x^{(m)} = A^m x^{(0)}$, $m \geqslant 0$, de ses images par les puissances successives de la matrice de transition (modifiée à l'aide du facteur d'amortissement, comme ci-dessus). Pour m assez grand, $x^{(m)}$ fournit une bonne approximation de la distribution de probabilité stationnaire. (Cette méthode est appelée méthode des puissances).

Le facteur d'amortissement permet de contrôler la vitesse de convergence du processus (les autres valeurs propres tendent vers 0 quand d tend vers 0), au prix d'une dégradation de la pertinence des résultats. Typiquement, $d = 0.85$ est la valeur utilisée.

Reprenons par exemple le graphe des trains. Voici les valeurs de PageRank obtenues par calcul exact sans facteur d'amortissement, par calcul exact avec $d = 0.85$, et en partant d'une distribution de probabilité concentrée sur Paris et en appliquant cinq fois la matrice de transition avec $d = 0.85$.

	Paris	Marseille	Lyon	Toulouse	Nice
Exact, sans amortissement :	0.249	0.284	0.255	0.067	0.145
Exact, $d = 0.85$:	0.242	0.274	0.246	0.086	0.152
5 itérations, $d = 0.85$:	0.237	0.272	0.252	0.086	0.153

À l'aide de quelques multiplications d'un vecteur par une matrice (au lieu d'une résolution de système linéaire), on obtient le bon classement.

4 Suggestions

1. Les propriétés des matrices stochastiques énoncées dans le texte mériteraient une démonstration plus soignée (sans aller jusqu'à démontrer le théorème de Perron-Frobenius).
2. Le théorème de Perron-Frobenius pourra être illustré à l'aide d'exemples et de contre-exemples.
3. On pourra essayer d'appliquer des méthodes générales de localisation des valeurs propres au cas des matrices stochastiques.
4. On pourra donner des exemples sur d'autres graphes intéressants que ceux donnés par l'énoncé.
5. Comparer les nombre d'opérations requises entre la méthode des puissances et le calcul exact de la distribution stationnaire. On pourra aussi s'intéresser à la tailles des données conservées en mémoire.
6. Montrer que le facteur d'amortissement permet bien d'éliminer les inconvénients signalés dans le texte.
7. On pourra étudier graphiquement l'effet des changements du facteur d'amortissement sur le PageRank, ainsi que le sensibilité au choix de la distribution initiale dans la méthode des puissance.
8. On pourra montrer que les valeurs propres de $dA + \frac{1-d}{n} U$ autres que 1 sont de module au plus d.