Classement de pages web

On propose d'étudier l'algorithme d'analyse de liens "PageRank", employé par Google comme mesure de l'importance relative d'un document dans un réseau d'hyperliens (notamment, l'importance d'une page sur le web).

On peut considérer que le web est une collection de N pages, la plupart des pages incluent des liens hypertextes vers d'autres pages : on dit qu'elles pointent vers ces autres pages. L'idée de base utilisée par les moteurs de recherche pour classer les pages par ordre de pertinence décroissante consiste à considérer que plus une page est la cible de liens venant d'autres pages, plus elle a de chances d'être fiable et intéressante pour l'utilisateur. Il s'agit de quantifier cette idée, c'est-à-dire d'attribuer un rang numérique ou un score de pertinence à chaque page.

On se donne un ordre sur l'ensemble des pages que l'on numérote de 1 à N. La structure de connectivité du web est alors représentée par une matrice C telle que $c_{i,j} = 1$ si la page i pointe vers la page j, sinon $c_{i,j} = 0$. Les liens d'une page vers elle même ne sont pas significatifs, donc $c_{i,i} = 0$.

On souhaite attribuer à chaque page i un score r_i pour pouvoir classer l'ensemble des pages par ordre décroissant et présenter à l'utilisateur une liste de pages ainsi classées. L'algorithme Pagerank part du principe qu'un lien de la page j pointant sur la page j contribue positivement au score de cette dernière, avec une pondération par le score r_j de la page dont est issu le lien (une page ayant un score élevé a ainsi plus de poids qu'une n'ayant qu'un score médiocre) et par le nombre total de liens présents sur ladite page $N_j = \sum_{k=1}^{N} N c_{k,j}$. On introduit donc la matrice Q définie par $q_{i,j} = c_{i,j}/N_j$ si $N_j \neq 0$, $q_{i,j} = 0$ sinon. La somme des coefficients des colonnes non nulles de Q vaut toujours 1. L'application des principes ci-dessus conduit donc une équation pour le vecteur $r \in \mathbb{R}^N$ des scores des pages de la forme $r_i = \sum_{j=1}^{N} qi, jr_j$ c'est-à-dire $r = Qr$.

Le problème du classement des pages du Web se trouve ainsi ramené à la recherche dun vecteur propre d'une énorme matrice, associé à la valeur propre 1 ! Il peut arriver que la matrice Q n'admette pas la valeur propre 1 ce qui invalide quelque peu la philosophie originale de l'algorithme. Pour remédier ce défaut, on considère alors $e =^t (1, 1, \ldots, 1) \in \mathbb{R}^N$ et $d \in \mathbb{R}^N$ tel que $d_j = 1$ si $N_j = 0$, $d_j = 0$ sinon. La matrice

$$P = Q + \frac{1}{N} e^t d$$

est alors la transposée d'une matrice stochastique : ses coefficients sont tous positifs et la somme des coefficients de chaque colonne vaut 1 (remarquons quil sagit dune " petite " correction Q car N est très grand). Du fait que ${}^t eP =^t e$, on voit que P admet bien la valeur propre 1. Comme cette valeur propre est en général multiple, on effectue une dernière modification en choisissant un nombre $0 < \alpha < 1$ et en posant

$$A = \alpha P + (1 - \alpha)\frac{1}{N} e^t e$$

On pourra admettre quune telle matrice admet 1 comme plus grande valeur propre, que cette valeur propre est simple et que l'on peut choisir un vecteur propre correspondant composantes toutes positives (il sagit du théorème de Perron-Frobenius). Finalement, PageRank calcule un tel vecteur propre $r \in \mathbb{R}^N$, normalisé d'une façon ou d'une autre,

$$r = Ar$$

dont les N composantes fournissent le classement recherché des pages du Web. On sait combien cette stratégie s'est révélée efficace, puisque Google a totalement laminé les moteurs de recherche de première génération, comme Altavista, lesquels ont essentiellement disparu du paysage.

Tâche 1 *Calculer le vecteur r à l'aide de la méthode de la puissance.*

La difficulté de cete méthode vient de la taille de la matrice A qui est une matrice pleine alors que la matrice initiale Q est creuse.

Tâche 2 *Améliorer cette méthode en remarquant que*

- *le vecteur initial dans le calcul de la puissance peut être choisi de norme 1 (ce qui évite de renormaliser à chaque étape)*
- *$y = Az$ eut s'écrire sous la forme $y = \alpha Qz + \frac{\beta}{N}e$ avec $\beta = 1 - ||\alpha Qz||_1$.*