ENSICAEN
2ème année Informatique

# TP Mathématique pour l'Informatique
# PageRank

## 1 Problème du PageRank

### 1.1 Travail préparatoire

1. Soit le miniweb de la figure 1, qui est un graphe orienté non-valué. Soit $x_i$ le score d'importance associé à la page $i$. Ces scores sont positifs et $x_j > x_i$ indique que la page $j \neq i$ est plus importante que la page $i$ (avec $x_j = 0$ indiquant que la page $j$ possède le score d'importance le plus bas.

   (a) Une approche simple consiste à choisir comme $x_i$ le nombre de liens qui y sont incidents. Pour l'exemple de la figure 1, donner ces scores.

   Cette approche toutefois ignore un aspect important : un lien vers une page $i$ d'une page importante doit engendrer un bien meilleur score de la page $i$ qu'un lien d'une page bien moins importante. Par exemple, un lien de Yahoo vers votre page personnelle lui donne un bien meilleur score d'importance qu'un lien pointant d'une page conventionnelle. Une première approche pour incorporer cette idée est de prendre en compte dans le calcul de $x_i$ tous les scores des pages des liens incidents $j$ et le nombre de liens sortant de chaque page. Soit $n_j$ le nombre de liens sortant de $j$, le score de la page $i$ est ainsi

$$x_i = \sum_{j \sim i} \frac{x_j}{n_j} \ ,$$

   où la notation $j \sim i$ veut dire que les noeud $j$ et $i$ sont adjacents.

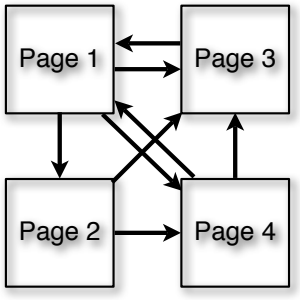


FIGURE 1 –

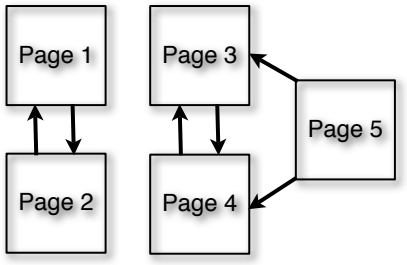


FIGURE 2 –

(a) Soit $x \in \mathbb{R}^4$ le vecteur contenant les scores des pages. Ecrire la relation algébrique auto-référentielle qui le définit à partir de celle des $x_i$.

(b) Soit $A$ la matrice qui intervient dans cette relation, appelée matrice des liens. Quelle propriété satisfait-elle ? Ce type de matrice est dénommé stochastique (en colonne).

(c) Comment alors obtenir $x$ de cette relation ?

(d) Montrer qu'il existe une solution $x$ à ce problème. Est-elle unique en général ?

(e) Calculer cette solution analytiquement dans le cas de l'exemple donné. Quelle est la page de score maximal ?

2. Non-unicité des classements : l'approche que nous venons de traiter contient toutefois une limitation majeure. En effet, la solution du problème de vecteur(s) propre(s) (i.e. classement) n'est pas nécessairement unique. Soit $V_1(A)$ le sous-espace propre (engendré par les vecteurs propres) associé à la valeur propre 1. On peut montrer que $\dim(V_1(A)) = 1$ (solution de classement unique donc) si le graphe est fortement connexe, mais ce n'est pas toujours le cas.

   (a) Soit le miniweb de la figure 2. Ecrire sa matrice de lien.

   (b) Montrer que sa solution de classement n'est pas unique, i.e. $\dim(V_1(A)) > 1$.

   (c) Généraliser ce raisonnement en considérant un web contenant $r$ sous-webs. A quel type de web (graphe) cela correspond-il ?

3. Montrer que le score d'importance de toute page n'ayant pas de lien incident est nul.

4. Dans notre analyse jusqu'à lors est que la façon dont les pages web sont indexées n'a aucun effet sur les scores. On se propose de formaliser cette assertion. Soit un web de $n$ pages indexées de 1 à $n$. Soit $P$ la matrice de permutation $n \times n$ qui échange les indices $i$ et $j$.

   (a) Montrer que $\tilde{A} = PAP$ est la matrice de liens associée au web re-indexé.

   (b) Montrer que si $v \neq 0$ est le vecteur propre associé à une valeur propre de $A$, $Pv$ est le vecteur propre de $\tilde{A}$ de même valeur propre. Conclure.

5. Le calcul du vecteur propre d'une matrice demande beaucoup de ressources calculatoires qu'il est convenable d'utiliser à bon escient lors d'une requête sur un web avec des milliards de pages. Il est donc fortement souhaitable d'assurer l'unicité du classement qui en

découle. On propose ci-après une modification de la méthode ci-dessus qui assure l'unicité des scores. Elle se fonde pour cela sur le théorème de Perron-Frobenius (que nous demandons d'admettre) :

**Théorème 1** *Soit $M \in \mathcal{A}_{n\times n}(\mathbb{R}^+ \setminus \{0\})$. Alors*

*(i) Il existe un seul réel strictement positif $r$, appelé la racine de Perron ou la valeur propre de Perron-Frobenius, qui est valeur propre maximale de $M$, i.e. $r = \rho(M)$.*

*(ii) La valeur propre $r$ est simple $\iff \dim(V_r) = 1$.*

*(iii) Il existe un vecteur propre $v$ associé à la valeur propre $r$ tel que $v > 0$. Ce vecteur est de plus unique avec $\|v\|_1 = 1$ (appelé vecteur de Perron).*

*(iv) La valeur propre $r$ a l'encadrement*

$$\min_i \sum_j M_{i,j} \leq r \leq \max_i \sum_j M_{i,j} = \|M\|_\infty \ .$$

Revenons à notre problème de calcul unique de scores d'importance. Soit $S$ la matrice $n \times n$ d'entrées toutes égales à $1/n$. C'est une matrice stochastique et $\dim(V_1(S)) = 1$ (une seule valeur propre égale à 1 et $n-1$ toutes nulles). On remplace la matrice $A$ par la combinaison convexe

$$M = (1-\alpha)A + \alpha S, \quad \alpha \in ]0,1] \ .$$

La valeur de $\alpha$ originellement utilisée par Google est 0.15.

(a) Montrer que $M$ est stochastique en colonne.

(b) Ecrire le nouveau problème à résoudre pour obtenir les scores.

(c) Montrer que $\rho(M) = 1$, que $\dim(V_1(M)) = 1$ et que le seul vecteur propre associé à $V_1(M)$ est strictement positif (vecteur de Perron).

(d) Que se passe-t-il pour les pages sans lien incident ?

# 2 Optimisation pour PageRank

On dénote $e = (1, \ldots, 1)^T$. Rappelons que le problème de PageRank est de résoudre :

$$\text{Trouver } x \geq 0 : \quad Mx = x \quad \text{et} \quad e^T x = 1 \ .$$

Clairement, ceci peut se formuler sous la forme du problème d'optimisation suivant

$$\min_{x\in\mathbb{R}^n} \{ f(x) \triangleq \frac{1}{2} \|Mx - x\|_2^2 + \frac{\gamma}{2} \left(e^T x - 1\right)^2 \} \ ,$$

où $\gamma > 0$ est un paramètre de pénalité pour la contrainte égalité $\sum_i x_i = 1$, et $\|\cdot\|_2$ est la norme Euclidienne.

Comme la fonction objective $f$ est quadratique et strictement convexe, et donc que le minimiseur est unique [1], on peut appliquer une simple descente de gradient pour la minimiser. On rappelle que l'itération de la descente de gradient s'écrit

$$x^{(k+1)} = x^{(k)} - \mu \nabla f(x^{(k)}) \ , \quad k \in \mathbb{N}$$

où $0 < \mu < 2/\beta$, et $\beta > 0$ est la constante de Lipschitz de $\nabla f$, i.e. la constante $\beta$ telle que

$$\|\nabla f(y) - \nabla f(z)\|_2 \leq \beta \|y - z\|_2 \quad \forall (y, z) \in \mathbb{R}^n \times \mathbb{R}^n \ .$$

## 2.1 Travail préparatoire

1. Calculer $\nabla f(x)$ pour $f$ du PageRank.
2. Dans ce cas, calculer la constante $\beta$ de $\nabla f$.

## 2.2 Travail d'implémentation

### Descente de gradient

1. Implémenter la descente de gradient avec $\mu = 1.99/\beta$, $\gamma = 1/n$ et $\gamma = 1/n^2$, et tracer $\left\|x^{(k)} - x^\star\right\|_2$ et $f(x^{(k)})$ en fonction de $k$, où $x^\star$ est le vecteur de score trouvé précédemment, par exemple en utilisant `eig`. Commenter les résultats obtenus.
2. A quelle vitesse décroit $\left\|x^{(k)} - x^\star\right\|$, comparer à la vitesse théorique (voir TD 2).
3. Quelle est la complexité de la descente de gradient à chaque itération ?
4. Peut-on l'accélérer et comment ? Quelle est sa nouvelle complexité ?
5. En utilisant la fonction `sparse`, implémenter la version accélérée. Commenter.

### Gradient conjugué

6. Implémenter la méthode du gradient conjugué pour minimiser $f$ du PageRank.
7. Quel est le nombre d'itérations au maximum pour avoir la solution exacte ?
8. Refaire les questions 3-5 pour le gradient conjugué.

1. En effet, cette assertion est une conséquence immédiate du point (iii) du Théorème de Perron-Frobenius (question 5(c)), et de ce fait $f(x) = 0$ si et seulement si $x$ est vecteur de Perron.