Algorithmique des réseaux sociaux **2013/2014**

Cours 1 — 7/8 Octobre

Enseignant: Marc Lelarge *Scribe: Marc Lelarge*

Pour information

– Page web du cours
`http://www.di.ens.fr/~lelarge/soc.html`

1.1 Recherche sur Internet en utilisant les hyper-liens

Dans ce cours, nous présentons deux algorithmes utilisant la structure de réseau des pages web pour améliorer les résultats de recherche par mot-clé. Les pages web constituent les noeuds du réseau tandis que les hyperliens constituent les arêtes dirigées du réseau.

1.1.1 L'algorithme HITS

Cet algorithme a été proposé par Jon Kleinberg en 1999 et est constitué de deux étapes :

1. la première étape consiste à identifier un sous-graphe de pages pertinentes sur lequel l'algorithme va travailler dans la deuxième étape. L'algorithme identifie tout d'abord un coeur d'environs 200 pages en utilisant des techniques standards d'extraction d'informations basées sur le texte. Comme certaines pages pertinentes peuvent ne pas contenir les mots de la requête, l'algorithme élargit le coeur en rajoutant toutes les pages pointées par une page du coeur ainsi que quelques pages pointant vers le coeur. Le but de cette première étape est de générer un sous-graphe induit G d'environs 3000 noeuds.
2. la deuxième étape va consister à trouver dans G les pages les plus pertinentes. L'idée est qu'une page peut être pertinente pour deux raisons : (a) elle contient de l'information sur le sujet de la requête : c'est une **autorité** ; (b) elle contient de nombreux liens vers de bonnes autorités : c'est un **hub**.

Une bonne autorité doit être connectée à de nombreux bons hubs et un bon hub doit pointer vers de nombreuses bonnes autorités. Pour mettre en pratique cette idée, nous définissons pour chaque sommet v un poids a_v correspondant à son autorité et un poids h_v correspondant à sa qualité de hub. Initialement, toutes les pages v ont poids a_v^0 et h_v^0 égaux à 1. Ensuite ces poids sont mis à jour selon la récursion :

$$\begin{aligned} a_v(t+1) &= \sum_{u \to v} h_u(t), \text{ pour tout } v, \\ h_v(t+1) &= \sum_{v \to u} a_u(t+1), \text{ pour tout } v. \end{aligned}$$

On note B la matrice d'adjacence du graphe i.e. $B_{uv} = 1$ si il y a une arête dirigée dans G de u vers v et $B_{uv} = 0$ sinon. On peut donc écrire les équations précédentes comme suit :

$$\mathbf{a}(t+1) = B^t\mathbf{h}(t) \text{ et, } \mathbf{h}(t+1) = B\mathbf{a}(t+1).$$

On a donc

$$\mathbf{a}(t+1) = B^tB\mathbf{a}(t) \text{ et, } \mathbf{h}(t+1) = BB^t\mathbf{h}(t). \tag{1.1}$$

La matrice symétrique B^tB est la matrice de co-citation. Pour (espérer) avoir convergence, à chaque étape, nous renormalisons les vecteurs $\mathbf{a}$ et $\mathbf{h}$ de telle sorte que $\sum_v a_v^2 = \sum_v h_v^2 = 1$. Pour $\mathbf{a} \in \mathbb{R}^n$, on note $s(\mathbf{a}) = \frac{1}{\|\mathbf{a}\|}\mathbf{a}$, avec $\|\mathbf{a}\|^2 = \sum_v a_v^2$. Si l'algorithme converge (avec cette renormalisation), d'après (**??**), la limite doit satisfaire $\|\mathbf{a}^\infty\| = 1$ et :

$$\mathbf{a}^\infty = s(B^tB\mathbf{a}^\infty).$$

Donc $\mathbf{a}^\infty$ doit être une valeur propre de la matrice symétrique B^tB. Soit $\omega_1, \omega_2, \ldots, \omega_n$ une base orthonormée de vecteurs propres de B^tB associés aux valeurs propres $\lambda_1, \ldots, \lambda_n$ numérotées de telle sorte que $|\lambda_1| \geq |\lambda_2| \geq \ldots$

On a alors en décomposant le vecteur $\mathbf{1}$ sur la base orthonormée :

$$\mathbf{a}^0 = \mathbf{1} = \sum_{i=1}^{n} \alpha_i\omega_i \text{ avec, } \alpha_i = \langle\omega_i, \mathbf{1}\rangle.$$

Donc en itérant (**??**), on obtient

$$\mathbf{a}^t = \sum_{i=1}^{n} \alpha_i\lambda_i^t\omega_i$$

Nous allons voir dans la section suivante que pourvu que B^tB soit primitive (définition donnée ci-dessous), le théorème de Perron-Frobenius **??** assure que $\lambda_1 > |\lambda_2|$ et que toutes les composantes de ω_1 sont strictement positives. Dans notre cas précis, cela implique que $\alpha_1 > 0$ et que lorsque $t \to \infty$,

$$s(\mathbf{a}^t) = \omega_1 + O\left(\frac{|\lambda_2|^t}{\lambda_1^t}\right).$$

Donc le vecteur normalisé des autorités converge vers le vecteur propre associé à la plus grande valeur propre de la matrice de co-citation B^tB et la vitesse de convergence dépend du ratio $\frac{|\lambda_2|}{\lambda_1}$.

1.1.2 Le théorème de Perron-Frobenius

Un vecteur est (resp. strictement) positif si chacune de ses composantes est (resp. strictement) positive. Pour une matrice $T = (T_{ij})$, on note $T \geq 0$ (resp. $T > 0$) si pour tout i, j, $T_{ij} \geq 0$ (resp. $T_{ij} > 0$). L'élément ij de la matrice carrée T élevée à la puissance k est noté $T^{(k)}_{ij}$.

Définition 1.1.1 *Une matrice carrée positive T est dite primitive si il existe $k > 0$ tel que $T^k > 0$, c'est à dire $T^{(k)}_{ij} > 0$ pour tout i, j.*

Théorème 1.1.1 *Soit T une matrice positive et primitive. Alors il existe une valeur propre r telle que*

1. *$r \in \mathbb{R}^+_*$, i.e. $r > 0$.*
2. *à r sont associés des vecteurs propres à gauche et à droite strictement positifs.*
3. *$r > |\lambda|$ pour toute valeur propre $\lambda \neq r$.*
4. *les vecteurs propres associés à r sont uniques à une constante mulptiplicative près.*

Démonstration. Pour $\mathbf{x} \geq 0$ avec $\mathbf{x} \neq \mathbf{0}$, on définit

$$r(\mathbf{x}) = \min_j \frac{\sum_i x_i T_{ij}}{x_j},$$

où le ratio est infini si $x_j = 0$. Il est clair que $0 \leq r(\mathbf{x}) < \infty$.

Nous montrons maintenant que $r(\mathbf{x})$ est borné uniformément en $\mathbf{x}$:

$$\begin{aligned} x_j r(\mathbf{x}) &\leq \sum_i x_i T_{ij} \text{ pour tout } j, \\ \mathbf{x}^t r(\mathbf{x}) &\leq \mathbf{x}^t T \\ r(\mathbf{x}) \mathbf{x}^t \mathbf{1} &\leq \mathbf{x}^t T \mathbf{1}, \end{aligned}$$

mais $T\mathbf{1} \leq K\mathbf{1}$ avec $K = \max_j \sum_i T_{ij}$, donc on a :

$$r(\mathbf{x}) \leq \frac{\mathbf{x}^t K \mathbf{1}}{\mathbf{x}^t \mathbf{1}} = K.$$

Comme T est primitive, elle ne contient pas de colonne nulle et donc $r(\mathbf{1}) > 0$. On a donc

$$r = \sup_{\mathbf{x} \geq \mathbf{0}, \mathbf{x} \neq \mathbf{0}} \min_j \frac{\sum_i x_i T_{ij}}{x_j} \geq r(\mathbf{1}) > 0,$$

et $0 < r \leq K$. On peut normalisé $\mathbf{x}$ sans affecter r donc

$$r = \sup_{\mathbf{x} \geq \mathbf{0}, \mathbf{x}^t \mathbf{x} = 1} \min_j \frac{\sum_i x_i T_{ij}}{x_j} \geq r(\mathbf{1}) > 0.$$

L'ensemble $\{\mathbf{x} \geq \mathbf{0}, \mathbf{x}^t\mathbf{x} = 1\}$ étant un compact de $\mathbb{R}^n$ et la fonction $\mathbf{x} \mapsto r(\mathbf{x})$ étant semi-continue supérieurement, elle atteint sa borne supérieure. Il existe donc $\hat{\mathbf{x}} \geq \mathbf{0}, \neq \mathbf{0}$ tel que

$$\sum_i \hat{x}_i T_{ij} \geq r\hat{x}_j, \text{ pour tout } j, \tag{1.2}$$

et avec égalité pour certains j.

Nous allons montrer que $\hat{\mathbf{x}}$ est en fait un vecteur propre à gauche pour T associé à la valeur propre r. On définit donc $\mathbf{z}^t = \hat{\mathbf{x}}^t T - r\hat{\mathbf{x}}^t \geq \mathbf{0}^t$ et on suppose par l'absurde que $\mathbf{x} \neq \mathbf{0}$. Par hypothèse, il existe k tel que $T^k > 0$ et donc

$$\mathbf{z}^t T^k = \hat{\mathbf{x}}^t T^k T - r\hat{\mathbf{x}}^t T^k > \mathbf{0}^t.$$

En écrivant $\mathbf{y}^t = \hat{\mathbf{x}}^t T^k$, on a donc

$$\sum_j y_i T_{ij} > r y_j, \text{ pour tout } j,$$

ce qui est une contradiction de la définition de r. On donc bien $\mathbf{z} = \mathbf{0}$ et le premier point est démontré :

$$\hat{\mathbf{x}}^t T = r\hat{\mathbf{x}}^t \tag{1.3}$$

En itérant (**??**), on a :

$$\hat{\mathbf{x}}^t T^k = r^k \hat{\mathbf{x}}^t.$$

En choisissant $T^k > 0$ et comme $\hat{\mathbf{x}} \geq \mathbf{0}, \neq \mathbf{0}$, on a $\hat{\mathbf{x}}^t T^k > \mathbf{0}^t$ et donc $\hat{\mathbf{x}} > \mathbf{0}$, ce qui prouve le second point (le cas du vecteur à droite se fait de manière similaire).

Nous montrons maintenant le troisième point. Soit λ une valeur propre de T, c'est à dire pour un $\mathbf{x} \neq \mathbf{0}$ ayant des valeurs possiblement complexes,

$$\sum_i x_i T_{ij} = \lambda x_j. \tag{1.4}$$

En prenant le module, on a donc

$$|\lambda||x_j| \leq \sum_i |x_i| T_{ij},$$

c'est à dire, $|\lambda| \leq \frac{\sum_i |x_i| T_{ij}}{|x_j|}$, donc $|\lambda| \leq r$.

Supposons par l'absurde que $|\lambda| = r$, on a alors

$$\sum_i |x_i| T_{ij} \geq r|x_j|.$$

Cette équation est exactement la même que (**??**), donc par le même raisonnement, on en déduit que

$$\sum_i |x_i| T_{ij} = r|x_j| > 0, \text{ pour tout } j,$$

c'est-à-dire que le vecteur $|\mathbf{x}|$ est un vecteur propre de T (à gauche) associé à la valeur propre r. On a donc

$$\sum_i |x_i| T_{ij}^{(k)} = |\lambda^k||x_j| > 0.$$

Or par définition, on a $\sum_i x_i T_{ij}^{(k)} = \lambda^k x_j$ donc au final, on a pour tout j, $|\sum_i x_i T_{ij}^{(k)}| = \sum_i |x_i| T_{ij}^{(k)}$ et $|x_j| > 0$.

On note que pour deux complexes $z, z' \neq 0$, si $|z + z'| = |z| + |z'|$ alors z et z' sonc colinéaires. On en déduit que tous les x_j sont colinéaires si bien que (**??**) se simplifie en

$$\sum_i |x_i| T_{ij} = \lambda |x_j|.$$

Comme $|x_j| > 0$ pour tout j, λ est réel, positif et comme nous avons supposé $|\lambda| = r$, on a $\lambda = r$.

Nous montrons maintenant le dernier point. Soit $\mathbf{y} \neq \mathbf{0}$ un vecteur propre à gauche associé à r. Pour tout c, le vecteur $\eta = \hat{\mathbf{x}} - c\mathbf{y}$ (pourvu qu'il soit non nul) est un vecteur propre à gauche associé à r. Donc si $\mathbf{y}$ n'est pas un multiple de $\hat{\mathbf{x}}$, on peut choisir c tel que $\eta \neq \mathbf{0}$ mais que certaines composantes de η soient nulles. Or par l'argument précédent, $|\eta|$ est également un vecteur propre associé à r et est strictement positif. Ce dernier point contredit le fait que $\mathbf{y}$ et $\hat{\mathbf{x}}$ ne soient pas multiples.

Nous avons donné toutes les preuves pour les vecteurs propres à gauche et tous les arguments peuvent être répétés pour les vecteurs propres à droite. Le point 3 montre que le nombre r produit par l'analyse des vecteurs propres à droite est le même que ci-dessus. □

1.1.3 PageRank