Graphe du Web – Classement de pages webs

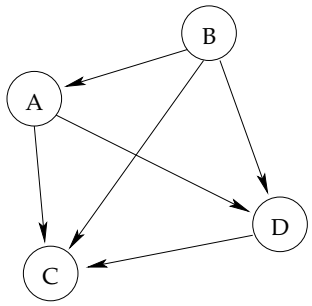
1 HITS et ses limites

On rappelle que l'algorithme HITS proposé par Kleinberg définit pour chaque page un "poids de pivot" PP (*hubness* en anglais) et un poids d'autorité PA. On nomme M la matrice d'adjacence du graphe du web : $M_{u,v} = 1$ si la page u possède un lien vers v, $M_{u,v} = 0$ sinon.

Question 1.1. *Calculez PA' (le vecteur PA à l'itération suivante) en fonction de PA et de M, PP' en fonction de PP et de M.*

Question 1.2. *Que représentent $[M^T \times M]_{u,v}$ et $[M \times M^T]_{u,v}$?*

Question 1.3. *Faire tourner à la main l'algorithme HITS sur le petit exemple ci-dessous.*



Question 1.4. *Vous avez écrit une page web, et vous voulez qu'elle soit référencée comme une autorité (pour apparaître en tête du classement de Google...), que faites-vous ? Comment un moteur de recherche peut-il contrer cette attitude ?*

2 Topic distillation, par K. Bharat et M. Henziger

Lire les parties 2,3 et 4 de l'article "Improved Algorithms for Topic Distillation in a Hyperlinked Environment" [1], et répondre au questions suivantes.

Question 2.1. *Quels sont les problèmes de HITS mis en évidence par cette étude ? (partie 2)*

Question 2.2. *Quelle pondération est introduite pour résoudre le premier problème ? Sur quoi se base-t'elle ?*

Question 2.3. *Expliquez rapidement ce que les auteurs proposent pour résoudre les 2 autres problèmes. Que pensez-vous de leur solution ? (partie 4)*

3 SALSA, par R. Lempel et S. Moran

Lire l'extrait de l'article "SALSA : The stochastic Approach for Link-Structure Analysis" [2], et répondre au questions correspondantes.

Question 3.1. *Lire l'extrait d'article joint. Étant donné le graphe G de la partie 1, que vaut $\tilde{G}$?*

Question 3.2. *Comment comprenez-vous la référence à la marche aléatoire (random walk, p. 140) ? Faites le lien avec PageRank.*

Question 3.3. *Que représentent $\tilde{h}_{i,j}$ et $\tilde{a}_{i,j}$ (p. 141) ?*

Question 3.4. *Montrer que $\tilde{H} = W_r {W_c}^T$ et que $\tilde{A} = {W_c}^T W_r$.*

Question 3.5. *À quelle approche les auteurs font-ils référence en temps que "Mutual Reinforcement" ? (p. 141)*

Question 3.6. *Quelle approche les auteurs choisissent-ils pour montrer que leur méthode est moins sensible à l'effet des communautés étroitement liées que HITS ? Que pensez-vous de cette approche ?*

Question 3.7. *Ces deux papiers cherchent à résoudre des problèmes très similaires. Comparer ces deux approches, en termes de :*
- *méthode et données utilisées,*
- *efficacité,*
- *complexité, possibilité de mise en oeuvre pratique.*

4 Coefficient de clusterisation

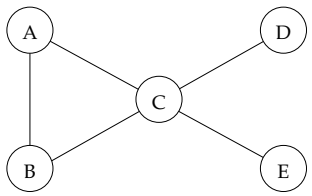
Il existe deux définitions pour le coefficient de clusterisation
- une définition globale :

$$C_1 = \frac{\sum_{u \in V} \text{nombre de triangles dont } u \text{ est un sommet}}{\sum_{u \in V} \text{nombre de triplets dont } u \text{ est un sommet}}$$

- une définition basée sur la moyenne de coefficients locaux :

$$C_2 = \frac{1}{n} \sum_{u \in V} \frac{\text{nombre de triangles dont } u \text{ est un sommet}}{\text{nombre de triplets dont } u \text{ est un sommet}}$$

Question 4.1. *Calculez la valeurs des coefficients de clusterisation du graphe ci-dessous.*



Question 4.2. *Construire un graphe tel que $C_1 = \Theta(1)$ et $C_2 = \Theta(1/n)$.*

Sources et références

[1] Krishna Bharat and Monika Rauch Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *SIGIR*, pages 104–111, 1998.

[2] R. Lempel and S. Moran. Salsa : the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2) :131–160, 2001.