UNIVERSITÉ CATHOLIQUE DE LOUVAIN

Projet LINFO1114 - Mathématiques discrètes
Groupe 13

Ranking de réseaux sociaux et pages web : PageRank

Auteurs :
Quentin COLLA
Gilles MAES
Cédric KHEIRALLAH

Professeur :
Marco SAERENS
Assistants :
Sylvain COURTAIN
Pierre LELEUX

15 mai 2022

1 Introduction

Au cours de ce projet, il nous a été demandé d'implémenter en **Python** l'algorithme **PageRank**. Celui-ci est utilisé par le moteur de recherche Google dans le but de mesurer la qualité et la quantité des liens reçus par une page sur un site Internet.

Notre implémentation doit fournir deux méthodes différentes : la résolution du système linéaire (`pageRankLinear`) ainsi que la *power method* (`pageRankPower`).

Détails de l'implémentation

- La téléportation est considérée comme *non-uniforme*. Cela implique que nous devons utiliser une matrice de transitions P (ainsi qu'un vecteur de personnalisation v qui est propre à notre groupe). Le calcul de la matrice de transitions suit la formule suivante :

$$w_{j\cdot} = \sum_{i=1}^{n} w_{ji}$$

$$P(page(k+1) = i \mid page(k) = j) = \frac{w_{ji}}{w_{j\cdot}}$$

  Notons que $w_{j\cdot}$ est le nombre d'arêtes *sortantes* du noeud j.

- Notre fichier contenant l'implémentation des différentes méthodes doit également contenir une méthode *"main"* qui :
  1. Lit un fichier `.csv` contenant une matrice d'adjacence **A**.
  2. Exécute le calcul de **PageRank** (via les deux méthodes).
  3. Affiche les résultats.

2 Rappel théorique

PageRank est un algorithme ayant pour but de classer des noeuds dans un graphe. Nous présentons ici deux méthodes de calculs différentes dont les résultats concordent.

2.1 PageRank linéaire

La méthode linéaire se base sur la résolution d'un système d'équations linéaires. En partant de la définition de la matrice Google, on peut former un système contenant :

- La matrice identité $\mathbf{I}$
- Le paramètre de téléportation α
- Le vecteur de personnalisation $\mathbf{v}$
- La matrice de probabilités de transition $\mathbf{P}$

$$\begin{aligned}
&\mathbf{x}^T\mathbf{G} = \mathbf{x}^T \\
&\Rightarrow \mathbf{x}^T(\alpha\mathbf{P} + (1-\alpha)\mathbf{e}\mathbf{v}^T) = \mathbf{x}^T \\
&\Rightarrow \alpha\mathbf{x}^T\mathbf{P} + (1-\alpha)\mathbf{v}^T = \mathbf{x}^T \\
&\Rightarrow \mathbf{x}^T(\mathbf{I} - \alpha\mathbf{P} = (1-\alpha)\mathbf{v}^T \\
&\Rightarrow (\mathbf{I} - \alpha\mathbf{P})^T\mathbf{x} = (1-\alpha)\mathbf{v}
\end{aligned}$$

Le paramètre de téléportation ainsi que le vecteur de personnalisation sont donnés, il ne reste qu'à calculer la matrice de probabilités de transition. Pour cela, il suffit de diviser chaque ligne de la matrice d'adjacence par sa somme, et de transposer la matrice résultante.

2.2 PageRank power method

La *power method* n'inclut pas l'utilisation d'un système d'équations, mais le calcul du vecteur propre de gauche de la matrice Google. Celle-ci peut être obtenue en utilisant la formule suivante [3] :

$$\mathbf{G} = \alpha\mathbf{P} + (1-\alpha)\mathbf{e}\mathbf{v}^T$$

Pour obtenir le vecteur résultant de l'algorithme PageRank, il suffit de multiplier la matrice Google avec notre vecteur de personnalisation v jusqu'à ce que ce qu'on atteigne le niveau de précision demandé.

$$\begin{cases} \mathbf{x} \leftarrow \mathbf{G}\mathbf{x} \\ \mathbf{x} \leftarrow \dfrac{\mathbf{x}}{||\mathbf{x}||} \end{cases}$$

2.3 Impact du paramètre de téléportation α

Le paramètre de téléportation α définit la probabilité qu'un utilisateur (qui clique aléatoirement sur les différentes pages disponibles) se déplace sur une page voisine à celle sur laquelle il se trouve, ou se "téléporte" sur une page aléatoire.

De manière plus précise :
- ➢ α représente la probabilité que l'utilisateur se déplace sur une page voisine
- ➢ $(1-\alpha)$ représente représente la probabilité que l'utilisateur soit téléporté sur une page aléatoire.

Un paramètre de téléportation trop élevé pourrait empêcher l'utilisateur d'atteindre certaines pages isolées, tandis qu'un paramètre de téléportation réduirait l'intérêt des liens entre les pages, car l'utilisateur serait téléporté aléatoirement trop fréquemment.

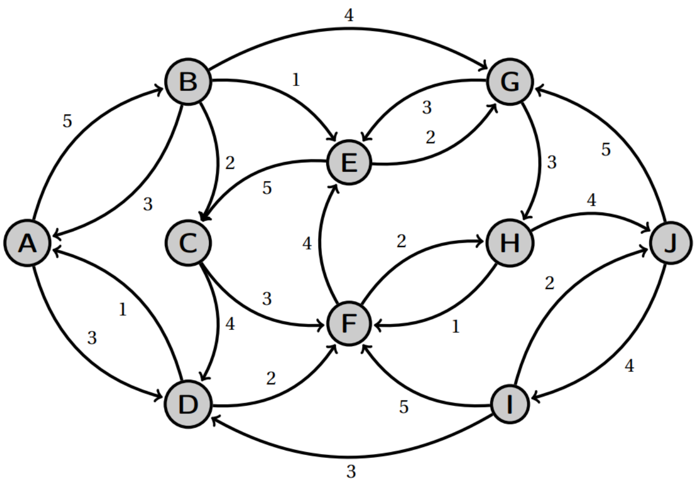
2.4 Impact du vecteur de personnalisation v

Le vecteur de personnalisation v gère la téléportation de l'utilisateur et définit les probabilités que celui-ci soit téléporté sur chacune des différentes pages. Etant donné qu'il s'agit d'un vecteur de probabilité, la somme de ses éléments sera égale à 1.

Les différentes valeurs présentes dans ce vecteur ont intérêt à être réparties équitablement. Dans le cas contraire, si, par exemple, l'utilisateur sera toujours téléporté sur la même page, il se peut que certaines pages ne soient pas atteignables.

3 PageRank power method

La première étape des deux méthodes est de traduire le graphe donné en consignes ci-dessous et d'écrire sa matrice d'adjacence A :



$$A = \begin{bmatrix} 0 & 5 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 2 & 0 & 1 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 3 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 4 \\ 0 & 0 & 0 & 3 & 0 & 5 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 5 & 0 & 4 & 0 \end{bmatrix}$$

FIGURE 1 – Graphe dirigé et pondéré représentant un réseau de 10 noeuds, G

À partir de la matrice d'adjacence A, nous pouvons simplement calculer la matrice de transition P en divisant chaque ligne de la matrice d'adjacence par sa somme, et finalement en transposant la matrice finale :

$$\text{Matrice de probabilités de transition P} = \begin{bmatrix} 0 & 3/10 & 0 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5/8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2/10 & 0 & 0 & 5/7 & 0 & 0 & 0 & 0 & 0 \\ 3/8 & 0 & 4/7 & 0 & 0 & 0 & 0 & 0 & 3/10 & 0 \\ 0 & 1/10 & 0 & 0 & 0 & 4/6 & 3/6 & 0 & 0 & 0 \\ 0 & 0 & 3/7 & 2/3 & 0 & 0 & 0 & 1/5 & 5/10 & 0 \\ 0 & 4/10 & 0 & 0 & 2/7 & 0 & 0 & 0 & 0 & 5/9 \\ 0 & 0 & 0 & 0 & 0 & 2/6 & 3/6 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4/9 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4/5 & 2/10 & 0 \end{bmatrix}$$

Nous avons à présent tous les éléments pour commencer la *power method*. Calculons d'abord la matrice Google G grâce à la formule vue au point précédent :

$$\text{Matrice Google G} = \begin{bmatrix} 0.01 & 0.28 & 0.01 & 0.31 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.5725 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.19 & 0.01 & 0.01 & 0.65285714 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 \\ 0.3475 & 0.01 & 0.52428571 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.28 & 0.01 \\ 0.01 & 0.1 & 0.01 & 0.01 & 0.01 & 0.61 & 0.46 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.39571429 & 0.61 & 0.01 & 0.01 & 0.01 & 0.19 & 0.46 & 0.01 \\ 0.01 & 0.37 & 0.01 & 0.01 & 0.26714286 & 0.01 & 0.01 & 0.01 & 0.01 & 0.51 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.31 & 0.46 & 0.01 & 0.01 & 0.01 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.41 \\ 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.01 & 0.73 & 0.19 & 0.01 \end{bmatrix}$$

Une fois que nous avons obtenu notre matrice Google G, il suffit d'itérer dessus grâce à la formule :

$$\boxed{\mathbf{x}^T\mathbf{G} = \mathbf{x}^T}$$

Vecteur de scores à l'itération 1 :

$$\begin{bmatrix} 0.071722 & 0.090269 & 0.035064 & 0.122633 & 0.213394 & 0.151232 & 0.082210 & 0.139570 & 0.036800 & 0.057106 \end{bmatrix}$$

Vecteur de scores à l'itération 2 :

$$\begin{bmatrix} 0.071163 & 0.050344 & 0.163430 & 0.062175 & 0.145858 & 0.138787 & 0.125923 & 0.092364 & 0.032842 & 0.117114 \end{bmatrix}$$

Vecteur de scores à l'itération 3 :

$$\begin{bmatrix} 0.042245 & 0.050029 & 0.112828 & 0.126935 & 0.154468 & 0.141747 & 0.124187 & 0.108301 & 0.056846 & 0.082414 \end{bmatrix}$$

Vecteur de scores final x :

$$\begin{bmatrix} 0.051099 & 0.038743 & 0.119933 & 0.102126 & 0.160159 & 0.159144 & 0.113746 & 0.108929 & 0.048892 & 0.097229 \end{bmatrix}$$

4 PageRank linéaire

Il s'agissait ici de résoudre le système linéaire :

$$I - (\alpha \;.\; P) \;.\; x = (1 - \alpha) \;.\; v$$

Ce qui, avec la matrice d'adjacence du graphe donné en consignes, le vecteur de personnalisation de notre groupe et un paramètre de téléportation α à 0.9 par défaut, nous donne comme système linéaire :

$$\begin{bmatrix} 1 & -27/100 & 0 & -3/10 & 0 & 0 & 0 & 0 & 0 & 0 \\ -9/16 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -9/50 & 1 & 0 & -9/14 & 0 & 0 & 0 & 0 & 0 \\ -27/80 & 0 & -18/35 & 1 & 0 & 0 & 0 & 0 & -27/100 & 0 \\ 0 & -9/100 & 0 & 0 & 1 & -3/5 & -9/20 & 0 & 0 & 0 \\ 0 & 0 & -27/70 & -3/5 & 0 & 1 & 0 & -9/50 & -9/20 & 0 \\ 0 & -9/25 & 0 & 0 & -9/35 & 0 & 1 & 0 & 0 & -1/2 \\ 0 & 0 & 0 & 0 & 0 & -3/10 & -9/20 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2/5 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -18/25 & -9/50 & 1 \end{bmatrix} . \; x = \begin{bmatrix} 0.01427 \\ 0.00996 \\ 0.00957 \\ 0.01161 \\ 0.00111 \\ 0.02162 \\ 0.01438 \\ 0.00513 \\ 0.00565 \\ 0.0067 \end{bmatrix}$$

Ce qui, une fois résolut donne un vecteur de scores x :

$$x = \begin{bmatrix} 0.056354 & 0.041659 & 0.118115 & 0.102789 & 0.157182 & 0.167189 & 0.115577 & 0.107296 & 0.042275 & 0.091563 \end{bmatrix}$$

5 Conclusion

Pour conclure, nous avons obtenu des réponses très proches avec les deux différentes méthodes ce qui prouve que nous avons bien réussi à implémenter les deux algorithmes de **PageRank** demandés avec pour seules valeurs de départ, un graphe dirigé pondéré et un vecteur de personnalisation, tous deux donnés dans les consignes de ce projet.
Après avoir rappelé la théorie de l'algorithme du PageRank, nous avons détaillé la résolution des deux méthodes étape par étape.
Ce projet nous a permis de mettre en pratique et d'améliorer notre compréhension des différentes notions vues et étudiées au cours durant le quadrimestre.

Références

[1] Dr. Leonardo Ermann. Google matrix. http://www.scholarpedia.org/article/Google_matrix, 2016.

[2] Jonathan Machado. Linear algebra application : Google pagerank algorithm. https://mathstats.uncg.edu/sites/yasaki/publications/machado-google-pagerank-linear-algebra-project.pdf.

[3] Marco Saerens. Slides du chaptitre 10, 2022. p128 (chaînes de Markov) et p137 (matrice Google).