AAIA TP 3IF - PageRank

Florian Rascoussier Christine Solnon

28 décembre 2023

Table des Matières

1 Exercice 1 : Marche aléatoire et Matrice de Transition dans PageRank

Montrez que pour tout $k \geq 1$ et tout couple de sommets $(i, j) \in S \times S, M^k[i][j]$ est égal à la probabilité d'arriver sur la page j en k clics à partir de la page i.

Pour comprendre comment la matrice de transition M se rapporte à la probabilité d'arriver sur une page donnée en k clics, considérons d'abord ce que signifie une marche aléatoire sur le graphe du web $G = (S, A)$.

1.1 Rappels et explications

Les propriétés étudiées dans le cours[1] sur la théorie des graphes aborde les notions de marche aléatoire et de matrice de transition [1]. Ces notions sont essentielles pour comprendre l'algorithme PageRank.

1.1.1 Marche aléatoire sur le graphe

Dans le contexte de l'algorithme PageRank, la marche aléatoire représente un utilisateur naviguant sur le Web en cliquant sur des liens de manière aléatoire. À chaque étape :

- L'utilisateur est sur une page i (un sommet dans le graphe).
- Il clique sur un lien sortant au hasard, le menant à une page j (un autre sommet).

1.1.2 Matrice de transition M

La matrice de transition M reflète la probabilité de passer d'une page à une autre en un seul clic. Ainsi, $M[i][j]$ représente la probabilité de passer de la page i à la page j. Elle est définie comme :

- $M[i][j] = \frac{1}{d^+(i)}$ si un lien (arc) existe de i à j (c'est-à-dire $i \to j$).
- $M[i][j] = 0$ sinon.

Où $d^+(i)$ est le nombre de liens sortants de la page i (son demi-degré extérieur).

1.1.3 Puissances de la matrice de transition M^k

Lorsque l'on considère M^k, avec $k > 1$, on regarde le processus de navigation après k clics. Voici comment cela fonctionne :

- $M^1 = M$ représente la probabilité de passer d'une page à une autre en un seul clic.
- $M^2 = M \times M$ représente la probabilité de passer d'une page à une autre en deux clics.
- Et ainsi de suite, jusqu'à M^k, qui représente la probabilité de passer d'une page à une autre en k clics.

1. Voir le cours d'AAIA, section 3.1 p11. `http://perso.citi-lab.fr/csolnon/supportAlgoGraphes.pdf`

Pour montrer que $M^k[i][j]$ est égal à la probabilité d'arriver sur la page j en k clics à partir de la page i, on utilise la propriété des marches aléatoires et des matrices de transition. La démonstration informelle peut être faite par induction sur k :

- **Initialisation (k=1)** : Par définition, $M[i][j]$ est la probabilité d'aller de i à j en un clic.
- **Récurrence** : Supposons que $M^{k-1}[i][j]$ représente la probabilité d'aller de i à j en $k-1$ clics. M^k est alors le produit de M et M^{k-1}, ce qui signifie que chaque élément $M^k[i][j]$ est la somme des probabilités de passer de i à un certain intermédiaire m en un clic (selon $M[i][m]$) et puis d'aller de m à j en $k-1$ clics (selon $M^{k-1}[m][j]$). Cela reflète la nature cumulative des probabilités dans les marches aléatoires.

1.2 Démonstration formelle par récurrence

Nous cherchons à démontrer que pour une matrice de transition M d'une marche aléatoire sur le graphe du web, $M^k[i][j]$ représente la probabilité d'arriver sur la page j en k clics à partir de la page i.

Rappel En Mathématique, démonter qu'une proposition $P(n)$ est vraie pour tout $n \in \mathbb{N}$ signifie démontrer que $P(1)$ est vrai quelle que soit la valeur de n. Dans ce cas, une preuve par récurrence sur k est appropriée.

1.2.1 Notations et Déclaration des Variables

- $M \in \mathbb{R}^{|S|\times|S|}$: Matrice de transition, où chaque élément $M[i][j]$ représente la probabilité de passer de la page i à la page j.
- $i, j \in \{1, 2, \ldots, |S|\}$: Indices représentant des pages spécifiques dans le graphe, où $|S|$ est le nombre total de pages ou sommets dans le graphe.
- $k \in \mathbb{N}$: Le nombre de clics (ou étapes) dans la marche.
- $m \in \{1, 2, \ldots, |S|\}$: Un sommet intermédiaire dans le graphe lors de la marche.
- S : L'ensemble des pages Web ou sommets du graphe. $S \in \mathbb{N}$ et représente le nombre total de sommets.
- $d^+(i) \in \mathbb{N}$: Le demi-degré extérieur du sommet i, c'est-à-dire le nombre de liens sortants de i dans un graphe orienté.
- $P(n)$: La proposition que nous cherchons à démontrer. Spécifiquement, $P(n)$ stipule que pour tout $n \in \mathbb{N}$ et pour tout couple de sommets $(i, j) \in S \times S$, $M^n[i][j]$ donne la probabilité d'une marche de longueur n de i à j.

1.2.2 Preuve par Induction sur k

Proposition $P(n)$: Pour tout $n \geq 1$ et tout couple de sommets $(i, j) \in S \times S$, $M^n[i][j]$ est égal à la probabilité d'arriver sur la page j en n clics à partir de la page i.

$P(n) : \forall n \in \mathbb{N}^*, \forall i, j \in \{1, 2, \ldots, |S|\}, M^n[i][j] =$ "probabilité d'une marche de i à j de longueur n".

1.2.3 Initialisation : $P(1)$

Par définition :

- Cas 1 : $M[i][j] = \frac{1}{d^+(i)}$ si $(i, j) \in A$.
- Cas 2 : $M[i][j] = 0$ si $(i, j) \notin A$.

Dans les deux cas, $M^1[i][j]$ correspond aux probabilités définies pour un clic unique, donc $P(1)$ est vrai.

1.2.4 Récurrence : $P(k) \to P(k+1)$

Supposons que $P(k)$ soit vrai, c'est-à-dire que $M^k[i][j]$ représente la probabilité d'aller de i à j en k clics.

Pour tout chemin de longueur $k+1$ de i à j, il doit passer par un intermédiaire m. Ainsi, la probabilité d'une marche de i à j en $k+1$ clics est la somme sur tous les sommets intermédiaires possibles m des produits des probabilités d'aller de i à m en k clics et de m à j en un clic.

$$M^{k+1}[i][j] = \sum_{m=1}^{|S|} M^k[i][m] \cdot M[m][j]$$

Cela correspond à la définition du produit de matrices (dot-product). Par hypothèse de récurrence, $M^k[i][m]$ est la probabilité d'une marche de longueur k de i à m, et $M[m][j]$ est la probabilité d'un clic de m à j. Ainsi, l'induction est complète et $P(n)$ est démontrée [2].

En Conclusion

Pour démontrer formellement que $M^k[i][j]$ est la probabilité d'arriver sur la page j en k clics à partir de la page i, vous devrez utiliser des arguments de probabilité et la définition des marches aléatoires, en plus d'une preuve par récurrence sur la puissance de la matrice. C'est un concept fondamental en théorie des graphes et dans l'analyse des algorithmes de marche aléatoire, comme celui de PageRank.

2 Exercice 2 : PageRank sur liste d'adjacence

Implémentez l'algorithme PageRank sur un graphe représenté par une liste d'adjacence. Donnez le résultat de l'algorithme sur le graphe G_1 fourni dans le fichier `res/example_1.txt` pour $k = 4$.

Dans cet exercice, nous allons implémenter l'algorithme PageRank sur un graphe représenté par une liste d'adjacence. L'implémentation C de cette liste d'adjacence est fournie.

L'étape la plus cruciale de l'algorithme PageRank est bien sûr le calcul d'une rangée de score (ou vecteur de probabilités de transition à l'étape k). Ce calcul se base sur la cardinalité des liens sortants du sommet considéré, ainsi que sur la valeur du score de ce sommet à l'étape $k-1$.

L'implémentation de cet algorithme donne ainsi le résultat suivant pour le graphe G_1 fourni :

2. Adapté d'une preuve consultable sur *math.stackexchange.com* [2]

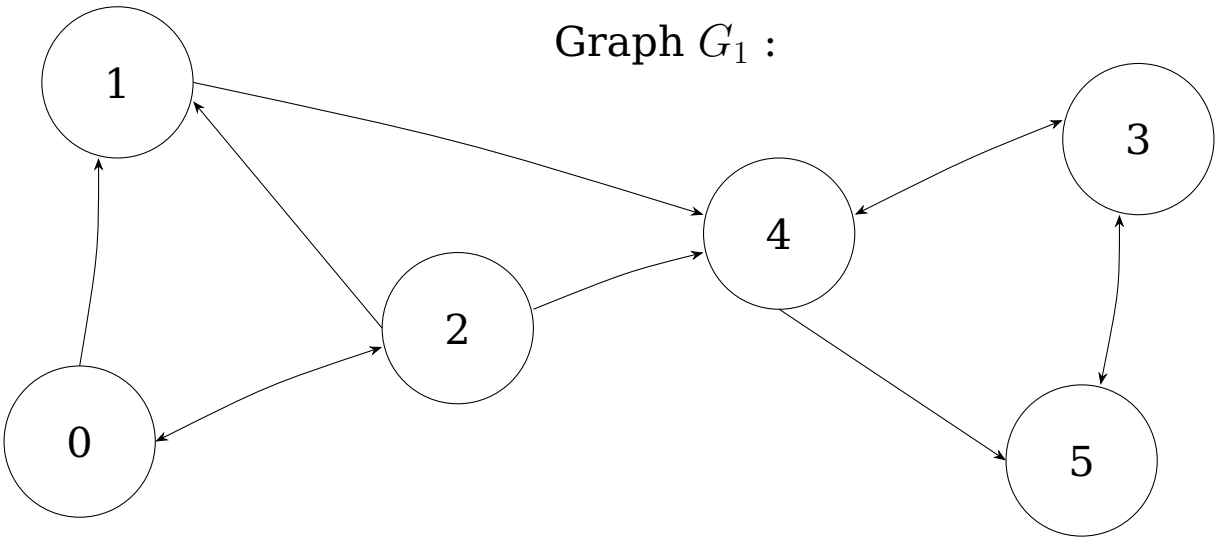

Figure 1 – Représentation du graphe G_1 fourni dans le fichier `res/example_1.txt`
.

```
    Scores at step 0: [0.166667, 0.166667, 0.166667, 0.166667, 0.166667, 0.166667]
    Scores at step 1: [0.055556, 0.138889, 0.083333, 0.250000, 0.305556, 0.166667]
    Scores at step 2: [0.027778, 0.055556, 0.027778, 0.319444, 0.291667, 0.277778]
    Scores at step 3: [0.009259, 0.023148, 0.013889, 0.423611, 0.224537, 0.305556]
    Scores at step 4: [0.004630, 0.009259, 0.004630, 0.417824, 0.239583, 0.324074]
```

Code 1 – Résultat de l'algorithme PageRank sur le graphe G_1 fourni, pour $k = 4$.

Le résultat est donc, pour $k = 4$: $[0.004630, 0.009259, 0.004630, 0.417824, 0.239583, 0.324074]$. Le sommet 3 est donc le plus important, suivi des sommets 5 et 4 à l'issue de l'étape 4.

3 Exercice 3 : Problème de PageRank sur matrice d'adjacence

Exécutez l'algorithme PageRank sur le graphe G_2 fourni dans le fichier `res/example_2.txt` pour $k = 4$. Calculez la somme des valeurs du vecteur s_k à chaque itération k. Qu'observe-t-on ? Comment évolue la somme des valeurs du vecteur s_k lorsque k augmente ?

Le graphe G_2 fourni dans le fichier `res/example_2.txt` est représenté dans la figure 2. Il s'agit du graphe G_1 dans lequel on a retiré le lien $1 \to 4$.

On relance l'algorithme PageRank sur le graphe G_2. On obtient le résultat suivant :

```
    S0: [0.166667, 0.166667, 0.166667, 0.166667, 0.166667, 0.166667] (sum: 1.000000)
    S1: [0.055556, 0.138889, 0.083333, 0.250000, 0.138889, 0.166667] (sum: 0.833333)
    S2: [0.027778, 0.055556, 0.027778, 0.236111, 0.152778, 0.194444] (sum: 0.694444)
    S3: [0.009259, 0.023148, 0.013889, 0.270833, 0.127315, 0.194444] (sum: 0.638889)
    S4: [0.004630, 0.009259, 0.004630, 0.258102, 0.140046, 0.199074] (sum: 0.615741)
```

Code 2 – Résultat de l'algorithme PageRank sur le graphe G_2 fourni, pour $k = 4$, avec la somme des valeurs du vecteur s_k à chaque itération k.

On remarque que la somme du vecteur s_k diminue à chaque itération. Hors, ce vecteur est censé donné la probabilité de se trouver sur chaque sommet du graphe à l'étape k. La somme de ces probabilités devrait donc être égale à 1.

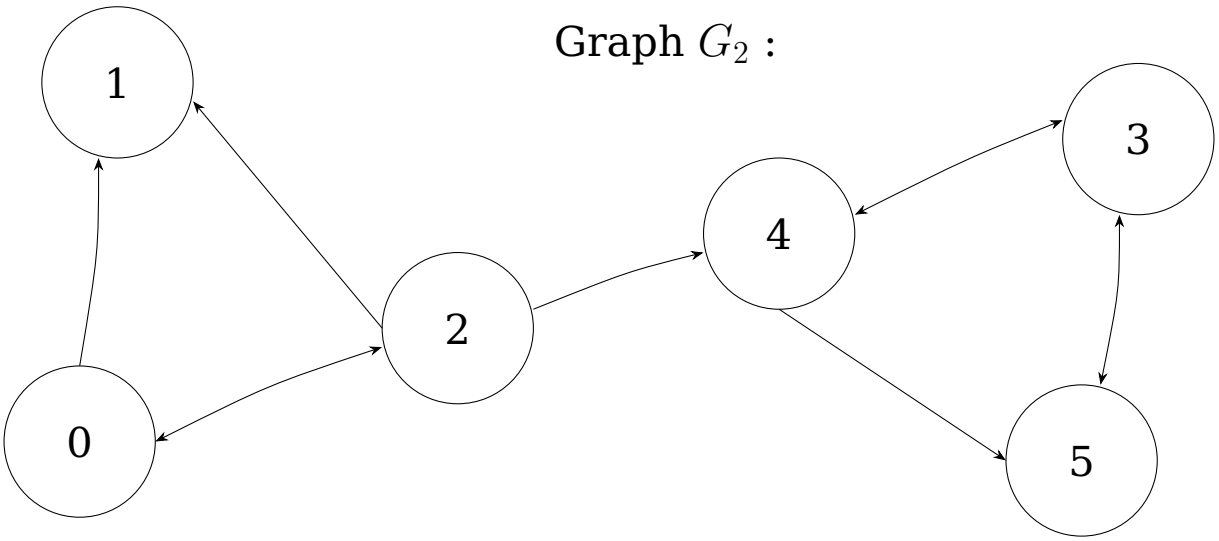

Figure 2 – Représentation du graphe G_2 fourni dans le fichier `res/example_2.txt`

.

4 Exercice 4 : Correction de l'algorithme PageRank sur matrice d'adjacence

On définit un sommet comme étant *absorbant* si il ne possède aucun lien sortant, c'est-à-dire si son demi-degré $d^+(i)$ extérieur est égal à 0. Dans le graphe G_2 fourni, le sommet 1 est absorbant.

Dans ce cas, l'algorithme PageRank actuel ne fonctionne pas correctement. En effet, la somme des valeurs du vecteur s_k à chaque itération k n'est pas forcément égale à 1. Cela est dû au fait que la matrice de transition M n'est pas stochastique, c'est-à-dire que la somme des valeurs de chaque ligne n'est pas égale à 1. Ainsi, lors du calcul des valeurs de s_k, la somme des valeurs du vecteur s_{k-1} n'est pas forcément égale à 1 non-plus.

Pour corriger ça, on va modifier la matrice de transition M pour la rendre stochastique. Pour cela, on va ajouter une probabilité de transition de chaque sommet vers tous les autres sommets du graphe. Ainsi, la somme des valeurs de chaque ligne de la matrice sera égale à 1. Le fait d'ajouter ces connexions virtuelles vers tous les autres sommets conserve le fait que le sommet absorbant ne discrimine aucun autre sommet. En effet, dans le cas de PageRank, un sommet qui n'est connecté à aucun autre sommet ou qui est connecté à tous les autres sommets du graphe est équivalent puisqu'il ne discrimine aucun autre sommet par rapport à un autre dans la marche aléatoire.

Après modification de l'algorithme, on obtient les résultats suivants :

```
    S0: [0.166667, 0.166667, 0.166667, 0.166667, 0.166667, 0.166667] (sum: 1.000000)
    S1: [0.083333, 0.166667, 0.111111, 0.277778, 0.166667, 0.194444] (sum: 1.000000)
    S2: [0.064815, 0.106481, 0.069444, 0.305556, 0.203704, 0.250000] (sum: 1.000000)
    S3: [0.040895, 0.073302, 0.050154, 0.369599, 0.193673, 0.272377] (sum: 1.000000)
    S4: [0.028935, 0.049383, 0.032665, 0.381430, 0.213735, 0.293853] (sum: 1.000000)
```

Code 3 – Résultat de l'algorithme PageRank sur le graphe G_2 fourni, pour $k = 4$, avec la somme des valeurs du vecteur s_k à chaque itération k.

Pour s_4, on obtient ainsi : $[0.028935, 0.049383, 0.032665, 0.381430, 0.213735, 0.293853]$. On obtient donc bien une somme de 1 à chaque itération k. On remarque également que les valeurs de s_k sont

différentes de celles obtenues précédemment. En effet, le sommet 1 est maintenant connecté à tous les autres sommets du graphe, ce qui fait qu'il est moins discriminant que les autres sommets. Ainsi, les valeurs de s_k sont plus équilibrées entre les différents sommets du graphe.

Acronymes

AAIA Algorithmique Avancée pour l'Intelligence Artificielle et les graphes. 2

Références

[1] Christine Solnon. *Première partie : Algorithmique avancée pour les graphes*. AAIA. CITI Lab, INSA Lyon dépt. Informatique. 2016. url : http://perso.citi-lab.fr/csolnon/supportAlgoGraphes.pdf (visité le 28/12/2023).

[2] Simón Ramírez Amaya et BMBM. *Proof - raising adjacency matrix to n-th power gives n-length walks between two vertices*. 2017. url : https://math.stackexchange.com/questions/2434064/proof-raising-adjacency-matrix-to-n-th-power-gives-n-length-walks-between (visité le 28/12/2023).