Projet 1 - Chaînes de Markov

Eléments de processus stochastiques

Floriane Magera
Romain Mormont
Fabrice Servais
Troisième bachelier en sciences de l'ingénieur

Chapitre 1

Question 1

1.1 Etude du modèle de base

2) Avant de calculer cette matrice, il est nécessaire de définir les poids/probabilités que nous appliquerons aux arcs du graphe modélisant la partie du Web qui nous intéresse. Nous avons décidé que la répartition des probabilités sur les arcs quittant un noeud doivent être uniforme (toutes les transitions sont donc équiprobables).

Suite à ce choix, la formation de la matrice de transition est très simple. Si l'on note A la matrice d'adjacence, alors il suffit d'appliquer la formule suivante pour calculer l'élément Q(i,j) :

$$Q(i,j) = A(i,j) \times \frac{1}{n} \sum_{j=1}^{n} A(i,j)$$

Cette formule permet de placer à 0 les éléments de Q représentant une transition impossible et de placer à une certaine probabilité les autres éléments de Q.

3) Nous avons choisi un nombre de pas $t = 20$. Le cas où le surfeur démarre aléatoirement sur le graphe est représenté par une distribution initiale π_0 uniforme et le cas où le surfeur démarre d'une page donnée est représenté par une distribution initiale π_0 où toutes les probabilités sont nulles sauf à l'index correspondant au noeud de départ. L'évolution des probabilités dans les deux cas est donnée sur le Figure 1.1.

La matrice $Q^{(20)}$ obtenue est la suivante :

$$Q^{(20)} = \begin{pmatrix} 0.3751 & 0.1874 & 0.1875 & 0.2500 \\ 0.3751 & 0.1877 & 0.1874 & 0.2499 \\ 0.3749 & 0.1875 & 0.1875 & 0.2500 \\ 0.3749 & 0.1875 & 0.1875 & 0.2500 \end{pmatrix}$$

On constate une convergence des distributions de probabilités vers un distribution π_s commune pour les deux distributions initiales ainsi qu'une convergence de la matrice Q.

4) La distribution stationnaire de la chaîne de Markov calculée par la méthode des puissances est la suivante :

$$\pi_\infty = \begin{pmatrix} 0.3750 & 0.1875 & 0.1875 & 0.2500 \end{pmatrix}$$

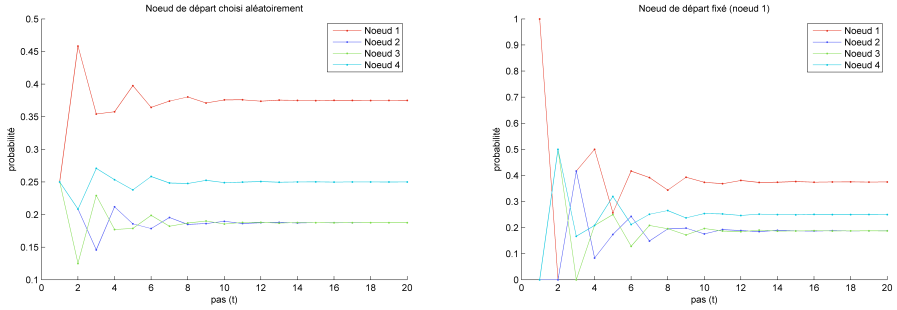


FIGURE 1.1 – Évolution de la distribution de probabilité

5) On constate que le noeud 1 possède le meilleur PageRank, suivi des noeuds 2 et 3 à égalité et du noeud 4. On peut expliquer ce classement intuitivement :

- le **noeud 1** possède le plus d'arêtes entrantes donc ayant le plus de chance d'être visité
- le **noeud 3** possède le moins d'arêtes entrantes donc ayant le moins de chance d'être visité
- les **noeuds 2** et **4** possèdent le même nombre intermédiaire (par rapport aux deux autres) d'arêtes entrantes. Le PageRank du noeud 4 est néanmoins plus élevé que celui du noeud 2 puisque le noeud 4 possèdent une arête entrante venant du noeud 1 qui est le plus visité.
- malgré un nombre d'arête entrante plus élevé que pour le noeud 3, le **noeud 2** possède une PageRank égal. Cela est dû au fait que, d'une part, le noeud 3 peut être visité depuis le noeud le plus visité (noeud 1) ce qui améliore son PageRank et, d'autre part, que le noeud 2 ne peut être accéder depuis des noeuds moins visités (noeud 3 et 4) ce qui abaisse son PageRank.

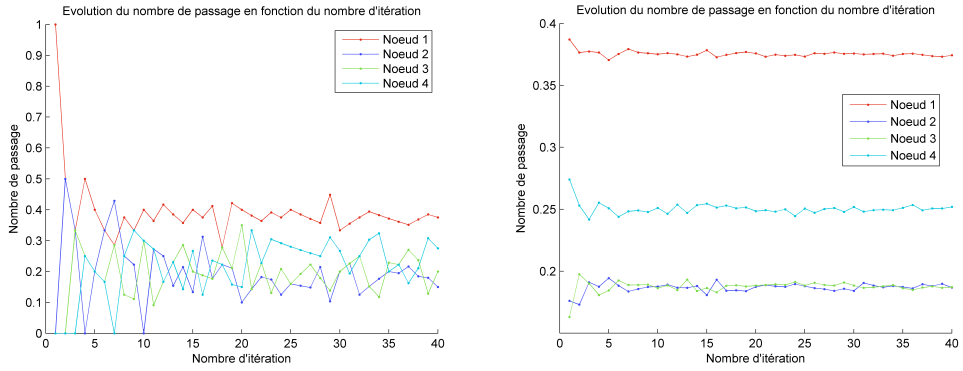
6) Dans un premier temps, nous avons généré une chaîne pour chaque longueur. Le résultat obtenu est donné sur la Figure 1.2(a). On peut déjà observer que les différentes courbes obtenues oscillent autour de leur probabilité stationnaire associée. Néanmoins, le résultat nous semblant trop incertain étant donné les oscillations, nous avons décidé de refaire l'expérience en générant cette fois-ci 1000 chaînes pour chaque longueur (voir Figure 1.2(b)). Nous avons ensuite moyenné afin d'obtenir un résultat exploitable. Les courbes obtenues nous permettent de confirmer les premières observations.

7)

1.1.1 Analyse des matrices A_2 et A_3

1) Les évolutions des distributions de probabilité sont données dans la Figure 1.3. On constate d'un part l'apparition d'**oscillations** et d'autre part que la **probabilité d'aller sur un certain noeud** (noeud 3 pour les deux matrices) devient **nulle** à partir d'un moment donné.

2)



(a) 1 génération (b) 1000 générations

FIGURE 1.2 – Évolution du nombre de passage par un noeud

1.2 Téléportation

1) La formule utilisée pour calculer la matrice de transition Q_t du modèle du surfeur avec téléportation est la suivante :

$$Q_t = (1 - \alpha)Q' + \alpha\tilde{Q}$$

où Q' et $\tilde{Q}$ sont des matrices de transition et α la probabilité de téléportation.

La première est la matrice de transition du graphe initial auquel on a rajouté des arêtes partant des *dangling nodes*. Elle a été calculée en remplaçant tous les éléments de la matrice Q dans des lignes ne contenant que des 0 par $\frac{1}{n}$. Cette valeur $\frac{1}{n}$ a été choisie en considérant une densité de probabilité uniforme entre les différentes arêtes partant des *dangling nodes*.

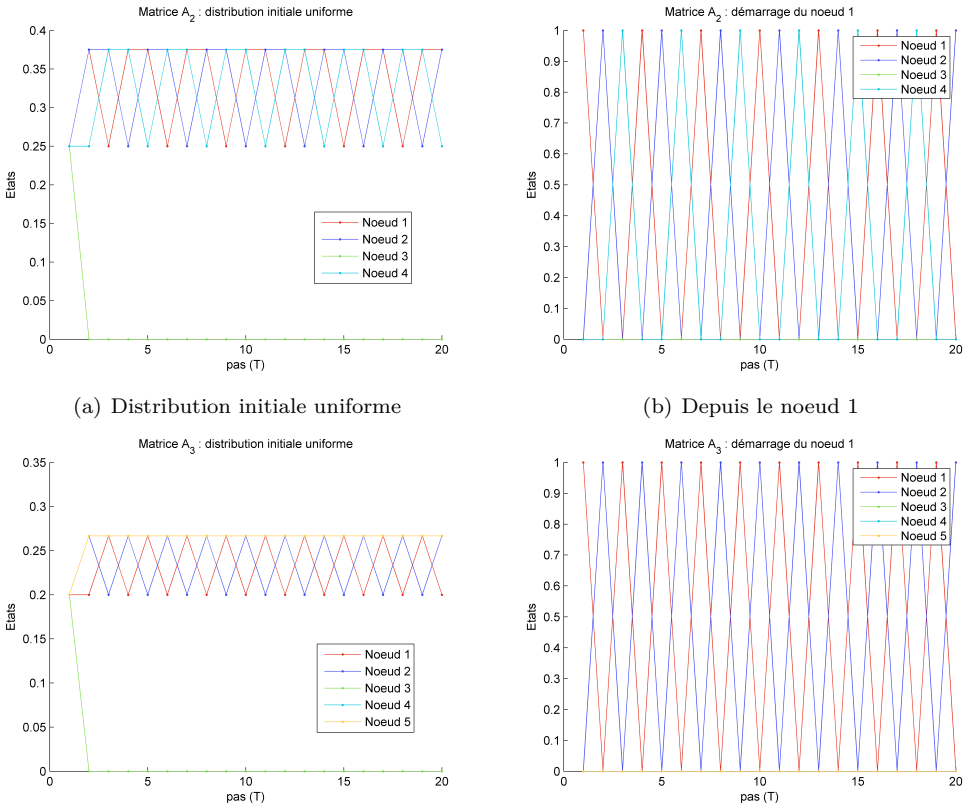
La seconde est la matrice de transition du graphe complet formé des noeuds du graphe initial. Autrement dit, la matrice de transition représentant la téléportation. Une combinaison linéaire de paramètre α est ensuite appliquée aux deux matrices pour trouver la matrice Q_t.

2) Pour que la distribution stationnaire π_s soit unique, **il faut que la chaîne de Markov soit irréductible**. Autrement dit, il faut que pour tout couple de nœuds (i_1, i_2), il existe une arête les reliant (un probabilité non-nulle de passer de i_1 à i_2). Cette propriété est vérifiée avec le modèle du surfeur modifié puisque la téléportation permet, depuis tout noeud, de se diriger vers un autre noeud tant que $\alpha > 0$.

A partir du moment où $\alpha = 0$, on est plus assuré que chaque paire de nœuds est reliée par une arête et donc que π_s est bien stationnaire.

3) Les sites les plus visités, obtenus à l'aide de la distribution stationnaire, sont les suivants :

1. http ://purl.org/rss/1.0/modules/content
2. http ://www.ulg.ac.be
3. http ://ogp.me/ns♯
4. http ://www.gre-liege.be



(a) Distribution initiale uniforme

(b) Depuis le noeud 1

(c) Distribution initiale uniforme

(d) Depuis le noeud 1

FIGURE 1.3 – Évolution des distributions de probabilités

5. http ://blog.intelliterwal.net
6. http ://www.jalios.com
7. http ://www.vmfnet.be
8. http ://www.alinoa.be
9. http ://www.ulb.ac.be
10. http ://www.cedia.ulg.ac.be

4)

1.3 Effet de α

1) Pour prouver que le score PageRank de toute page est au moins $\frac{\alpha}{n}$ (n est le nombre de pages), on peut développer une expression "*explicite*" des éléments de la matrice Q_t en utilisant la formule donnée précédemment :

$$Q_t(i,j) = q_{ij}(1-\alpha) + \frac{1}{n}\alpha$$

où q_{ij} est un élément de la matrice Q'. Connaissant la relation qui lie $\pi^{(k)}$ et $\pi^{(k-1)}$, on a :

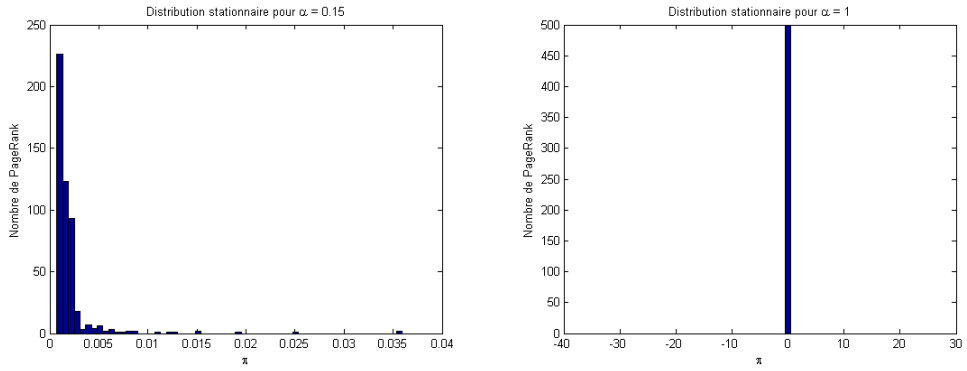
$$\begin{aligned}
\pi_j^{(k)} &= \sum_{i=1}^{n} Q_t(i,j)\pi_i^{(k-1)} \\
&= \sum_{i=1}^{n} \left(q_{ij}(1-\alpha) + \frac{\alpha}{n} \right) \pi_i^{(k-1)} \\
&= \sum_{i=1}^{n} q_{ij}(1-\alpha)\pi_i^{(k-1)} + \sum_{i=1}^{n} \frac{\alpha}{n}\pi_i^{(k-1)} \\
&= \frac{\alpha}{n} + \underbrace{(1-\alpha)\sum_{i=1}^{n} q_{ij}\pi_i^{(k-1)}}_{>0}
\end{aligned}$$

Le deuxième terme est inférieur à 1 (et même inférieur à $(1 - \frac{\alpha}{n})$ afin de respecter le deuxième axiome de Kolmogorov) et surtout, positif. De ce fait, on peut affirmer que :

$$\pi_j^{(k)} \geq \frac{\alpha}{n}$$

Le cas où α tend vers 1 correspond à la situation où le surfeur a majoritairement tendance à se téléporter lorsqu'il change de page. Si on considère qu'en cas de téléportation, la distribution de probabilité est uniforme entre les nœuds de destination, on observera un PageRank uniforme.

Afin de vérifier cette conclusion, nous avons calculé la distribution stationnaire pour un $\alpha = 1$. Le résulat est donné sur la Figure 1.4 où il est mis en parallèle avec la distribution des PageRank pour $\alpha = 0.15$. Le résultat est édifiant, on constate en effet un PageRank uniforme dans le cas où $\alpha = 1$ (écart-type du PageRank : 4.7753×10^{-18}).



(a) $\alpha = 0.15$ (b) $\alpha = 1$

FIGURE 1.4 – Distribution des PageRank lorsque α tend vers 1