PageRank

Anaïs Vergne et Céline Comte

12 juin 2018

1 Introduction

Objectif L'algorithme PageRank permet de classer les pages web les unes par rapport aux autres en fonction de leur *importance relative* dans le graphe du web. Pour cela, il affecte un *rang* à chaque page, de sorte que le rang d'une page est d'autant plus élevé que cette page est importante. Ce rang est recalculé régulièrement, au fur et à mesure que des pages apparaissent et disparaissent sur le web. Lorsqu'un utilisateur saisit un ou plusieurs mots-clés dans un moteur de recherche, d'autres algorithmes sont utilisés (en temps réel) pour sélectionner les pages qui sont à la fois pertinentes et importantes.

Références PageRank a été introduit dans le papier [1] co-écrit par Sergey Brin et Lawrence Page en 1998 et dans le brevet [4] déposé par Lawrence Page en 1998. Ce travail s'inscrit dans la continuité d'autres travaux, comme l'algorithme HITS proposé par Jon Kleinberg en 1999 [2].

2 Définition du rang en exploitant la structure de graphe du web

Idée Le web est un ensemble de pages reliées les unes aux autres par des hyperliens (ou liens hypertextes). L'algorithme PageRank utilise ces hyperliens pour gagner de l'information sur l'importance des pages les unes par rapport aux autres : *une page est importante si elle est pointée par d'autres pages importantes.*

Le graphe du web Il s'agit d'un graphe orienté dont les sommets sont les (indices des) pages web et les arcs représentent les hyperliens entre les pages. Autrement dit, les sommets du graphe sont les entiers de 1 à N, où N est le nombre total de pages web ; pour chaque $i, j = 1, \dots, N$, le graphe contient un arc de i vers j si la page i contient (au moins) un hyperlien vers la page j, auquel cas on écrira $i \to j$. Si une page i ne contient aucun hyperlien sortant, on ajoute une boucle, c'est-à-dire un arc de i vers i. Cet ajout simplifiera certains raisonnement dans la suite du cours ; à noter cependant qu'il existe d'autres moyens de gérer le cas des pages n'ayant pas d'hyperlien sortant.
La *matrice d'adjacence* $A = (a(i,j))_{i,j=1,\dots,N}$ du graphe du web est définie par

$$a(i,j) = \mathbb{1}_{i \to j} = \begin{cases} 1 & \text{si } i \to j, \\ 0 & \text{sinon,} \end{cases} \qquad \forall i, j = 1, \dots, N.$$

Pour chaque $i = 1, \dots, N$, on note $d(i)$ le *degré sortant* du sommet i dans le graphe, égal au nombre de pages vers lesquelles la page i contient des hyperliens. C'est aussi la somme des coefficients de la ligne i dans la matrice d'adjacence A. Par construction, chaque page a un degré sortant supérieur ou égal à 1.

Remarque. Dans notre modèle simplifié, on ne compte qu'une fois les hyperliens multiples, ce qui revient à considérer que chaque page contient au plus un hyperlien vers chaque autre page. On peut facilement étendre les résultats présentés dans ce cours pour les prendre en compte, par exemple en utilisant un graphe pondéré.

Mesure d'importance d'une page L'algorithme PageRank affecte un rang $\pi(i) \in [0,1]$ à chaque page i du graphe. On supposera que les rangs sont normalisés, c'est-à-dire que $\sum_{i=1}^{N} \pi(i) = 1$. Comme expliqué plus tôt, l'algorithme repose sur l'idée qu'une page est importante si elle est pointée par des pages qui sont elles-mêmes importantes. Cela revient à interpréter un hyperlien d'une page i vers une page j comme un *vote* de la page i pour la page j. Ce vote aura d'autant plus de poids que la page i est elle-même importante.

Définition formelle : 1er essai Si l'on tente de mettre cette idée en équations, on obtient par exemple :

$$\pi(i) = \sum_{\substack{j=1 \\ j \to i}}^{N} \pi(j), \quad \forall i = 1, \dots, N. \tag{1}$$

Littéralement, pour chaque $i = 1, \dots, N$, le rang de la page i est égal à la somme des rangs des pages j qui pointent vers la page i. En utilisant la définition de la matrice d'adjacence A, on peut réécrire ce système sous la forme

$$\pi(i) = \sum_{j=1}^{N} \pi(j) a(j,i), \quad \forall i = 1, \dots, N. \tag{2}$$

Autrement dit, le vecteur ligne $\pi = [\pi(1), \pi(2), \dots, \pi(N)]$ est solution de l'équation matricielle $\pi = \pi A$. Les rangs $\pi(i)$, $i = 1, \dots, N$, sont ainsi implicitement définis comme solutions d'un système linéaire de N équations.
Remarquons que vecteur nul est toujours une solution du système (2). Imposer la condition de normalisation $\sum_{i=1}^{N} \pi(i) = 1$ permet d'éliminer cette solution qui n'apporte aucune information sur l'importance des pages.
La définition (2) a deux inconvénients :

- Tous les hyperliens ont le même poids, de sorte qu'une page peut artificiellement accroître son pouvoir de vote en augmentant le nombre d'hyperliens distincts qu'elle contient. On voudrait au contraire que le pouvoir de vote d'une page ne dépende que de sa propre importance dans le graphe.
- Le système (2) admet rarement une solution non nulle. Par exemple, si l'on somme toutes les équations, on obtient

  $$\sum_{i=1}^{N} \pi(i) = \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \to i}}^{N} \pi(j) = \sum_{j=1}^{N} \sum_{\substack{i=1 \\ j \to i}}^{N} \pi(j) = \sum_{j=1}^{N} d(j)\pi(j).$$

  Ceci implique que $\pi(1) = \pi(2) = \ldots = \pi(N) = 0$ dès que les sommets ont tous un degré sortant supérieur ou égal à 2.

Pour ces deux raisons, on considère une deuxième définition du rang.

Définition formelle : 2ème essai On résout le premier problème en divisant chaque vote d'une page j par son degré sortant $d(j)$, de sorte que la somme de ses votes est égale à 1. On obtient le système suivant :

$$\pi(i) = \sum_{j=1}^{N} \pi(j) \frac{1}{d(j)} \mathbb{1}_{j \to i}, \quad \forall i = 1, \dots, N. \tag{3}$$

À nouveau, il s'agit d'un système linéaire de N équations à N inconnues, auquel on ajoute la condition de normalisation $\sum_{i=1}^{N} \pi(i) = 1$. On peut réécrire ce système sous forme matricielle $\pi = \pi P$ où la matrice $P = (p(i,j))_{i,j=1,\dots,N}$ est définie à partir de la matrice d'adjacence A par

$$p(i,j) = \frac{1}{d(i)} \mathbb{1}_{i \to j} = \begin{cases} \frac{1}{d(i)} & \text{si } i \to j, \\ 0 & \text{sinon.} \end{cases} \qquad \forall i, j = 1, \dots, N.$$

On rappelle que, pour une page i qui ne contient pas d'hyperlien sortant, on ajoute un arc $i \to i$ dans le graphe, de sorte que $p(i,i) = 1$ et $p(i,j) = 0$ pour chaque $j \neq i$. Les coefficients de P sont tous positifs et leur somme sur chaque ligne vaut 1. La matrice P est donc *stochastique à droite*.

Exemple 1 On considère le graphe de l'Exemple 1. La page 1 contient des hyperliens vers les pages 2 et 3, la page 2 contient un hyperlien vers la page 3 et la page 3 contient un hyperlien vers la page 1. Sa matrice d'adjacence est donnée par

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Les degrés sortants valent $d(1) = 2$ et $d(2) = d(3) = 1$. Le système satisfait par les rangs des pages (selon la définition (3)) s'écrit

$$\begin{cases} \pi(1) = \pi(3), \\ \pi(2) = \frac{1}{2}\pi(1), \\ \pi(3) = \frac{1}{2}\pi(1) + \pi(2). \end{cases}$$

La première équation montre que les pages 1 et 3 ont la même importance, tandis que la deuxième équation montre que la page 2 a deux fois moins d'importance que la page 1 (et donc aussi que la page 3). On obtient une unique solution normalisée : $\pi(1) = \frac{2}{5}$, $\pi(2) = \frac{1}{5}$ et $\pi(3) = \frac{2}{5}$.
Sous forme matricielle, le système s'écrit $\pi = \pi P$, avec

$$P = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Questions L'exemple précédent nous montre que le système (3) définit bien le rang de façon unique dans certains cas. Dans les deux prochaines parties, nous tenterons de répondre aux deux questions suivantes :

- *Le système* (3) *admet-il toujours une unique solution normalisée ?* Nous verrons dans la partie 3 que ce n'est pas toujours le cas, et nous donnerons une condition suffisante.
- *Si oui, peut-on calculer cette solution efficacement sur de grands graphes ?* Cette question sera traitée dans la partie 4. Nous considérerons un algorithme itératif pour calculer une valeur approchée du rang et nous donnerons une condition suffisante pour qu'il fonctionne.

Dans la partie 5, nous introduirons une troisième définition du rang qui résoudra les problèmes soulevés dans les parties 3 et 4.

3 Existence et unicité du rang

Idée Écrire le rang π comme la distribution stationnaire d'une chaîne de Markov bien choisie. Sous certaines conditions, cela garantira l'existence et l'unicité de π.

Modèle du surfeur aléatoire On s'intéresse au parcours d'un individu (sans mémoire) qui surfe sur le web. On suppose qu'il saute de page en page de façon aléatoire, en suivant des hyperliens. Plus précisément, lorsque notre surfeur se trouve sur une page, il choisit un hyperlien de façon aléatoire et uniforme sur cette page et passe sur la page pointée par cet hyperlien.
On note $\mathbf{X}_0$ la page (potentiellement aléatoire) sur laquelle le surfeur se trouve initialement. Pour chaque $n \geq 1$, on note $\mathbf{X}_n$ la page sur laquelle le surfeur se trouve après le $n^{\text{ème}}$ saut. $(\mathbf{X}_n)_{n\in\mathbb{N}}$ est une chaîne de Markov homogène. Sa matrice de transition, notée $P = (p(i,j))_{i,j=1,\dots,N}$, est définie par

$$p(i,j) = \frac{1}{d(i)} \mathbb{1}_{i\to j} = \begin{cases} \frac{1}{d(i)} & \text{si } i \to j, \\ 0 & \text{sinon,} \end{cases} \qquad \forall i,j = 1,\dots,N.$$

Il s'agit de la matrice P définie dans la partie précédente. Si la chaîne de Markov admet une distribution stationnaire π, alors celle-ci satisfait l'équation matricielle $\pi = \pi P$, soit

$$\pi(i) = \sum_{j=1}^{N} \pi(j)p(j,i), \quad \forall i = 1, \dots, N,$$

avec la condition de normalisation

$$\sum_{i=1}^{N} \pi(i) = 1.$$

Ce système d'équations linéaires est identique au système (3) donné dans la partie précédente. Ainsi, si la chaîne de Markov définie par le parcours du surfeur admet une unique distribution stationnaire, alors les valeurs de cette distribution donnent les rangs des pages web.

Existence et unicité de π Le théorème suivant a été rappelé dans la leçon sur les chaînes de Markov.

Théorème 1. *Soit $(\mathbf{X}_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène, irréductible et récurrente. Alors $(\mathbf{X}_n)_{n\in\mathbb{N}}$ admet une unique distribution stationnaire π, et celle-ci vérifie $\pi(i) > 0$ pour chaque $i \in E$.*

La chaîne de Markov décrivant le parcours du surfeur vérifie-t-elle ces propriétés ?

Homogène Oui, car le choix du saut suivant dépend uniquement de la page sur laquelle se trouve le surfeur, et pas de l'étape à laquelle se produit le saut.

Irréductible La chaîne de Markov est irréductible si et seulement si son diagramme de transition est fortement connexe. Par définition du parcours du surfeur, ce diagramme de transition contient exactement les mêmes arcs que le graphe du web sous-jacent. Ainsi, la chaîne de Markov est irréductible si et seulement si le graphe du web est lui-même fortement connexe.

Récurrente Comme l'espace d'états $\{1, \dots, N\}$ est fini, la chaîne de Markov est automatiquement récurrente dès qu'elle est irréductible.

Au bout du compte, on a seulement besoin de vérifier si le graphe du web est fortement connexe. Par ergodicité, le rang d'une page sera alors égal à la fréquence à laquelle le surfeur aléatoire passe par cette page. Ce sera aussi l'inverse du nombre moyen de pas faits par le surfeur entre deux passages par cette page.

Exemple 1 (Existence et unicité) La graphe du web de l'Exemple 1 est fortement connexe. Le diagramme de transition de la chaîne de Markov associée est représenté en Figure 1. Le théorème 1 garantit donc qu'il existe une unique distribution stationnaire π, qui a été calculée dans la partie 2.

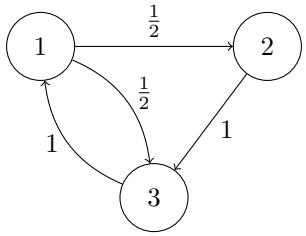


Figure 1 – Diagramme de transition de la chaîne de Markov décrivant le parcours d'un surfeur aléatoire sur le graphe de l'Exemple 1.

Exemple 3 (Rang nul) Que se passe-t-il si la chaîne de Markov n'est pas irréductible ? Considérons le graphe du web de l'Exemple 3. Ce graphe n'est pas fortement connexe : il contient deux composantes fortement connexes, la première constituée des pages 1, 2 et 3 et la seconde des pages 4, 5 et 6. Il existe un chemin des pages 1, 2 et 3 vers les pages 4, 5 et 6 mais la réciproque est fausse. En termes de chaînes de Markov, on dit que les états 1, 2 et 3 sont *transitoires* alors que les états 4, 5 et 6 sont *récurrents*. Le système d'équations linéaires associé s'écrit

$$\begin{cases} \pi(1) = \pi(3), & \pi(4) = \pi(6), \\ \pi(2) = \frac{1}{2}\pi(1) + \frac{1}{3}\pi(4), & \pi(5) = \frac{1}{3}\pi(4), \\ \pi(3) = \frac{1}{2}\pi(1) + \pi(2), & \pi(6) = \frac{1}{3}\pi(4) + \pi(5). \end{cases}$$

Les trois dernières équations imposent $\pi(4) = \pi(5) = \pi(6) = 0$. En injectant ce résultat dans les autres équations, on est ramené au système de l'Exemple 1, de sorte que $\pi(1) = \frac{2}{5}$, $\pi(2) = \frac{1}{5}$ et $\pi(3) = \frac{2}{5}$. Concrètement, cela signifie que les pages 1, 2 et 3 absorbent tous les votes sans les redistribuer aux pages 4, 5 et 6.

Exemple 4 (Absence d'unicité) Si la chaîne de Markov n'est pas irréductible, il n'y a pas non plus de garantie d'unicité de la distribution stationnaire. On considère le graphe du web de l'Exemple 4. Le système d'équations linéaires s'écrit

$$\begin{cases} \pi(1) = \pi(3), & \pi(4) = \frac{1}{2}\pi(6), & \pi(7) = \pi(9), \\ \pi(2) = \frac{1}{2}\pi(1) + \frac{1}{3}\pi(4), & \pi(5) = \frac{1}{3}\pi(4), & \pi(8) = \frac{1}{2}\pi(6) + \frac{1}{2}\pi(7), \\ \pi(3) = \frac{1}{2}\pi(1) + \pi(2), & \pi(6) = \frac{1}{3}\pi(4) + \pi(5), & \pi(9) = \frac{1}{2}\pi(7) = \pi(8). \end{cases}$$

À nouveau, les équations 4, 5 et 6 donnent $\pi(4) = \pi(5) = \pi(6) = 0$. En injectant ce résultat dans les autres équations, on obtient deux systèmes de la forme de celui de l'Exemple 1, vérifiés par les rangs des pages 1, 2 et 3 d'une part et des pages 7, 8 et 9 d'autre part. En utilisant la condition de normalisation, on déduit qu'il existe $p \in [0, 1]$ tel que $\pi(1) = \frac{2}{5}p$, $\pi(2) = \frac{1}{5}p$, $\pi(3) = \frac{2}{5}p$, $\pi(7) = \frac{2}{5}(1-p)$, $\pi(8) = \frac{1}{5}(1-p)$ et $\pi(9) = \frac{2}{5}(1-p)$. N'importe quel réel p entre 0 et 1 convient, donc il n'y a pas unicité de la distribution stationnaire.

4 Algorithme itératif pour le calcul du rang

Idée On suppose dans cette partie que le graphe du web est fortement connexe, et donc que sa chaîne de Markov associée admet une unique distribution stationnaire, égale au rang des pages. On va utiliser une autre propriété des chaînes de Markov pour calculer une valeur approchée du rang sans avoir à résoudre le système linéaire. Pour cela, on devra supposer que le graphe du web est apériodique.

Convergence vers la distribution limite On applique le résultat suivant :

Théorème 2. *Soit $(\mathbf{X}_n)_{n\in\mathbb{N}}$ une chaîne de Markov homogène, irréductible, récurrente* ***et apériodique*** *sur un espace d'états E. La distribution stationnaire π de $(\mathbf{X}_n)_{n\in\mathbb{N}}$ est aussi sa distribution limite, dans le sens où $\pi(i) = \lim_{n\to+\infty} \mathbb{P}(\mathbf{X}_n = i)$ pour chaque $i \in E$.*

Ainsi, si la chaîne de Markov définie par le parcours du surfeur est apériodique en plus d'être irréductible, alors on obtient une valeur approchée du rang en calculant itérativement les probabilités $\mathbb{P}(\mathbf{X}_n = i)$ pour chaque $i = 1, \ldots, N$. Cette idée est formalisée dans le paragraphe suivant.

Algorithme On note π_0 la distribution initiale de la chaîne de Markov, i.e., $\pi_0(i) = \mathbb{P}(\mathbf{X}_0 = i)$ pour chaque $i = 1, \ldots, N$. On peut par exemple supposer la distribution initiale uniforme sur l'ensemble des

pages. De même, pour chaque $n \geq 1$, on note π_n la distribution de la chaîne de Markov après le $n^{\text{ème}}$ saut, i.e., $\pi_n(i) = \mathbb{P}(\mathbf{X}_n = i)$ pour chaque $i = 1, \ldots, N$. Pour chaque $n \in \mathbb{N}$, on a

$$\mathbb{P}(\mathbf{X}_{n+1} = i) = \sum_{j=1}^{N} \mathbb{P}(\mathbf{X}_n = j)\mathbb{P}(\mathbf{X}_{n+1} = i | \mathbf{X}_n = j), \quad \forall i = 1, \ldots, N,$$

soit

$$\pi_{n+1}(i) = \sum_{j=1}^{N} \pi_n(j)p(j,i), \quad \forall i = 1, \ldots, N.$$

Pour chaque $n \in \mathbb{N}$, on assimile la distribution π_n au vecteur ligne $[\pi_n(1), \pi_n(2), \ldots, \pi_n(N)]$. On peut alors réécrire le système ci-dessus sous forme matricielle :

$$\pi_{n+1} = \pi_n P, \quad \forall n \in \mathbb{N}. \tag{4}$$

Le théorème 2 montre que $\lim_{n \to +\infty} \pi_n(i) = \pi(i)$ lorsque la chaîne de Markov est irréductible et apériodique. On peut donc calculer une valeur approchée de π et commençant avec une distribution initiale arbitraire π_0 puis appliquant l'équation (4) récursivement.

Remarque. Nous ne nous attardons pas ici sur la vitesse de convergence vers la limite, même cette question est très importante pour l'efficacité de l'algorithme PageRank. Nous reviendrons rapidement dessus dans la partie 5, après avoir introduit une troisième définition du rang.

Exemple 1 (Irréductible et apériodique) On revient au graphe du web de l'Exemple 1. On a vu que la chaîne de Markov associée était bien homogène, irréductible et récurrente. Elle est également apériodique. En effet, la période de l'état 1 est égale à 2 et celle de l'état 2 est égale à 3. Le PGCD de 2 et 3 étant égal à 1, cela confirme que la chaîne est apériodique. On peut donc utiliser l'algorithme ci-dessus pour obtenir une valeur approchée du rang calculé dans la partie 2. Pour chaque $n \in \mathbb{N}$, on a

$$\begin{cases} \pi_{n+1}(1) = \pi_n(3), \\ \pi_{n+1}(2) = \frac{1}{2}\pi_n(1), \\ \pi_{n+1}(3) = \frac{1}{2}\pi_n(1) + \pi_n(2). \end{cases}$$

On écrit les premières étapes de l'algorithme avec une distribution initiale uniforme sur l'ensemble des pages :

Étape	0	1	2	3	4	5	6	7	8
$\pi_n(1)$	$\frac{1}{3}$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{4}{12}$	$\frac{5}{12}$	$\frac{10}{24}$	$\frac{9}{24}$	$\frac{20}{48}$	$\frac{19}{48}$
$\pi_n(2)$	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{3}{12}$	$\frac{2}{12}$	$\frac{5}{24}$	$\frac{5}{24}$	$\frac{9}{48}$	$\frac{10}{48}$
$\pi_n(3)$	$\frac{1}{3}$	$\frac{3}{6}$	$\frac{2}{6}$	$\frac{5}{12}$	$\frac{5}{12}$	$\frac{9}{24}$	$\frac{10}{24}$	$\frac{19}{48}$	$\frac{19}{48}$

Avec une distribution unimodale, concentrée sur la première page :

Étape	0	1	2	3	4	5	6	7	8	9	10
$\pi_n(1)$	1	0	$\frac{1}{2}$	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{4}{8}$	$\frac{3}{8}$	$\frac{6}{16}$	$\frac{7}{16}$	$\frac{12}{32}$	$\frac{13}{32}$
$\pi_n(2)$	0	$\frac{1}{2}$	0	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{3}{16}$	$\frac{3}{16}$	$\frac{7}{32}$	$\frac{6}{32}$
$\pi_n(3)$	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{7}{16}$	$\frac{6}{16}$	$\frac{13}{32}$	$\frac{13}{32}$

Dans les deux cas, l'algorithme semble bien converger vers la distribution stationnaire calculée en partie 2. La vitesse de convergence dépend de la distribution initiale choisie.

Exemple 2 (Irréductible et périodique) On considère le graphe de l'Exemple 2. La chaîne de Markov associée est irréductible. Le théorème 1 garantit l'existence et l'unicité de la distribution stationnaire. La résolution du système linéaire donne directement $\pi(1) = \pi(4) = \frac{1}{3}$ et $\pi(2) = \pi(3) = \frac{1}{6}$.
La chaîne de Markov est périodique (de période 2) donc le théorème 2 ne s'applique pas. On peut voir numériquement que la distribution de la chaîne de Markov ne converge pas nécessairement vers une distribution limite. Pour chaque $n \in \mathbb{N}$, on a

$$\begin{cases} \pi_{n+1}(1) = \pi_n(4), & \pi_{n+1}(3) = \pi_n(2), \\ \pi_{n+1}(2) = \frac{1}{2}\pi_n(1), & \pi_{n+1}(4) = \frac{1}{2}\pi_n(1) + \pi_n(3). \end{cases}$$

Si le surfeur commence sur la page 1 avec probabilité 1, on obtient :

Étape	0	1	2	3	4	5	6	7	8
$\pi_n(1)$	1	0	$\frac{1}{2}$	0	$\frac{3}{4}$	0	$\frac{5}{8}$	0	$\frac{11}{16}$
$\pi_n(2)$	0	$\frac{1}{2}$	0	$\frac{1}{4}$	0	$\frac{3}{8}$	0	$\frac{5}{16}$	0
$\pi_n(3)$	0	0	$\frac{1}{2}$	0	$\frac{1}{4}$	0	$\frac{3}{8}$	0	$\frac{5}{16}$
$\pi_n(4)$	0	$\frac{1}{2}$	0	$\frac{3}{4}$	0	$\frac{5}{8}$	0	$\frac{11}{16}$	0

Il ne semble pas y avoir de convergence vers la distribution stationnaire car la masse de la distribution alterne entre les pages 1 et 3 d'une part et 2 et 4 d'autre part.

Exemple 3 (Non irréductible) Que se passe-t-il si l'on applique le même algorithme alors que la chaîne de Markov n'est pas irréductible ? Pour le voir, on revient au graphe de l'Exemple 3. Pour chaque $n \in \mathbb{N}$, on a

$$\begin{cases} \pi_{n+1}(1) = \pi_n(3), & \pi_{n+1}(4) = \pi_n(6), \\ \pi_{n+1}(2) = \frac{1}{2}\pi_n(1) + \frac{1}{3}\pi_n(4), & \pi_{n+1}(5) = \frac{1}{3}\pi_n(4), \\ \pi_{n+1}(3) = \frac{1}{2}\pi_n(1) + \pi_n(2), & \pi_{n+1}(6) = \frac{1}{3}\pi_n(4) + \pi_n(5). \end{cases}$$

En partant d'une distribution uniforme sur toutes les pages, on obtient :

Étape	0	1	2	3	4	5	6	7
$\pi_n(1)$	$\frac{1}{6}$	$\frac{6}{36}$	$\frac{9}{36}$	$\frac{48}{216}$	$\frac{57}{216}$	$\frac{402}{1296}$	$\frac{363}{1296}$	$\frac{2568}{7776}$
$\pi_n(2)$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{5}{36}$	$\frac{43}{216}$	$\frac{32}{216}$	$\frac{227}{1296}$	$\frac{249}{1296}$	$\frac{1297}{7776}$
$\pi_n(3)$	$\frac{1}{6}$	$\frac{9}{36}$	$\frac{8}{36}$	$\frac{57}{216}$	$\frac{67}{216}$	$\frac{363}{1296}$	$\frac{428}{1296}$	$\frac{2583}{7776}$
$\pi_n(4)$	$\frac{1}{6}$	$\frac{6}{36}$	$\frac{8}{36}$	$\frac{24}{216}$	$\frac{28}{216}$	$\frac{144}{1296}$	$\frac{104}{1296}$	$\frac{624}{7776}$
$\pi_n(5)$	$\frac{1}{6}$	$\frac{2}{36}$	$\frac{2}{36}$	$\frac{16}{216}$	$\frac{8}{216}$	$\frac{56}{1296}$	$\frac{48}{1296}$	$\frac{208}{7776}$
$\pi_n(6)$	$\frac{1}{6}$	$\frac{8}{36}$	$\frac{4}{36}$	$\frac{28}{216}$	$\frac{24}{216}$	$\frac{104}{1296}$	$\frac{104}{1296}$	$\frac{496}{7776}$

La probabilité semble s'amasser vers les pages 1, 2 et 3 alors que les probabilités des pages 4, 5 et 6 diminuent.

5 Amortissement

Idée On résout du même coup les problèmes d'irréductibilité et d'apériodicité en ajoutant un *amortissement* qui redistribue une partie du rang de chaque page de façon uniforme dans tout le graphe. Comme dans les parties 2 et 3, on considère deux définitions alternatives : une première comme solution d'un système d'équations linéaires et une seconde comme distribution stationnaire d'une chaîne de Markov.

Définition formelle : 3ème (et dernier) essai Le vote $\pi(j)$ de chaque page j est distribué de la façon suivante :

— une fraction $\alpha < 1$ de son vote est réparti uniformément entre les pages pointées par ses hyperliens, comme expliqué dans la partie 2 ;

— la fraction $1 - \alpha$ restante est répartie de façon uniforme entre toutes les pages du web, y compris la page j elle-même.

La constante α, appelée *facteur d'amortissement*, est la même pour toutes les pages. Sa valeur est souvent fixée à $0,85$. Nous reviendrons dessus plus tard. Le système d'équations linéaires s'écrit maintenant :

$$\pi(i) = \sum_{j=1}^{N} \pi(j) \left(\alpha \underbrace{\frac{1}{d(i)} \mathbb{1}_{i \to j}}_{p(j,i)} + (1-\alpha)\frac{1}{N} \right), \quad \forall i = 1, \ldots, N.$$

En utilisant la propriété de normalisation, on peut réécrire ce système sous la forme

$$\pi(i) = \alpha \left(\sum_{j=1}^{N} \pi(j) \frac{1}{d(j)} \mathbb{1}_{j \to i} \right) + (1-\alpha)\frac{1}{N}, \quad \forall i = 1, \ldots, N. \tag{5}$$

On a en particulier

$$\pi(i) \geq (1-\alpha)\frac{1}{N}, \quad \forall i = 1, \ldots, N,$$

ce qui veut dire que chaque page est garantie d'avoir un rang supérieur ou égal à $(1-\alpha)\frac{1}{N}$.
En écriture matricielle, on obtient

$$\pi = \pi M \quad \text{avec} \quad M = \alpha P + (1-\alpha)\frac{1}{N} E$$

où P est la matrice définie en partie 2 et E est la matrice $N \times N$ dont tous les coefficients sont égaux à 1, ou encore

$$\pi = \alpha \pi P + (1-\alpha)\frac{1}{N} e$$

où e est le vecteur ligne de dimension N dont tous les coefficients sont égaux à 1.

Modèle du surfeur aléatoire et algorithme itératif À nouveau, on souhaite écrire le rang π que l'on vient d'introduire comme la distribution stationnaire d'une chaîne de Markov (qui, cette fois, sera toujours irréductible et apériodique). Pour cela, on modifie le modèle du surfeur aléatoire de la partie 3. On suppose que le surfeur s'ennuie parfois de suivre des hyperliens, auquel cas il "saute" sur une page choisie au hasard de façon uniforme dans tout le web. Plus précisément, à chaque saut, le surfeur choisit entre deux alternatives :

— avec probabilité $\alpha < 1$, il suit un hyperlien choisi de façon aléatoire et uniforme dans la page où il se trouve actuellement, comme décrit dans la partie 3 ;

— avec probabilité $1 - \alpha$, le surfeur passe sur une page choisie de façon aléatoire et uniforme *parmi toutes les pages du web*, y compris la page sur laquelle il se trouve.

La matrice de transition $M = (m(i,j))_{i,j=1,\ldots,N}$ de la chaîne de Markov qui décrit le parcours du surfeur est maintenant donnée par :

$$m(i,j) = \alpha \frac{1}{d(i)} \mathbb{1}_{i \to j} + (1-\alpha)\frac{1}{N}, \quad \forall i, j = 1, \ldots, N.$$

Autrement dit, on a $M = \alpha P + (1-\alpha)\frac{1}{N}E$. Comme dans la partie 4, on note π_0 la distribution initiale de la chaîne de Markov et, pour chaque $n \geq 1$, π_n sa distribution après le $n^{\text{ème}}$ saut. Cette distribution vérifie la récurrence suivante : pour chaque $n \in \mathbb{N}$,

$$\pi_{n+1}(i) = \sum_{j=1}^{N} \pi_n(j) \left(\alpha \frac{1}{d(j)} \mathbb{1}_{j \to i} + (1-\alpha)\frac{1}{N} \right) = \alpha \left(\sum_{j=1}^{N} \pi_n(j) \frac{1}{d(j)} \mathbb{1}_{j \to i} \right) + (1-\alpha)\frac{1}{N}, \quad \forall i = 1, \ldots, N.$$

soit, en écriture matricielle,

$$\pi_{n+1} = \pi_n M = \alpha \pi_n P + (1-\alpha)\frac{1}{N},$$

Remarque. On dit qu'on applique la *méthode itérative* si l'on utilise la seconde formule, obtenue après avoir appliqué la condition de normalisation. Lorsqu'on utilise la première formule, on dit qu'on applique la *méthode de puissance*.

Exemple 1 On illustre cette troisième définition sur le graphe de l'Exemple 1. Le diagramme de transition de la chaîne de Markov décrivant le parcours du surfeur aléatoire avec amortissement est représenté en Figure 2. On constate que ce diagramme est complet, ce qui garantit du même coup l'irréductibilité (on peut relier deux états arbitraires en suivant l'arc qui les relie) et l'apériodicité (chaque état a une boucle, donc sa période est 1).

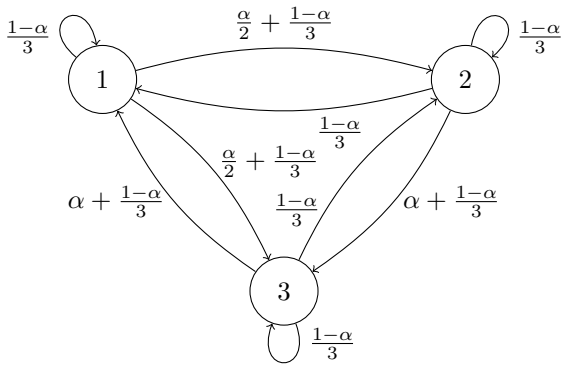


Figure 2 – Diagramme de transition de la chaîne de Markov décrivant le parcours d'un surfeur aléatoire, avec amortissement, sur le graphe de l'Exemple 1.

Le rang est défini par le système d'équations linéaires suivante :

$$\begin{cases} \pi(1) = \alpha\pi(3) + (1-\alpha)\frac{1}{3}, \\ \pi(2) = \alpha\frac{1}{2}\pi(1) + (1-\alpha)\frac{1}{3}, \\ \pi(3) = \alpha\left(\frac{1}{2}\pi(1) + \pi(2)\right) + (1-\alpha)\frac{1}{3}. \end{cases}$$

Sous forme matricielle, ce système s'écrit $\pi = \pi M$ avec

$$M = \alpha \times \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} + (1-\alpha) \times \frac{1}{3} \times \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{1-\alpha}{3} & \frac{\alpha}{2} + \frac{1-\alpha}{3} & \frac{\alpha}{2} + \frac{1-\alpha}{3} \\ \frac{1-\alpha}{3} & \frac{1-\alpha}{3} & \alpha + \frac{1-\alpha}{3} \\ \alpha + \frac{1-\alpha}{3} & \frac{1-\alpha}{3} & \frac{1-\alpha}{3} \end{bmatrix}$$

Propriétés Les remarques de l'exemple précédent restent valables sur n'importe quel graphe du web. La chaîne de Markov vérifie donc toutes les hypothèses souhaitées :

Homogène Oui, pour les mêmes raisons qu'avant.

Irréductible Le diagramme de transition de la chaîne de Markov est maintenant complet (dès que $\alpha < 1$). Il est donc fortement connexe même si le graphe du web ne l'est pas.

Récurrente Oui, car la chaîne de Markov est irréductible et son espace d'états $\{1, \ldots, N\}$ est fini.

Apériodique Tous les nœuds ont des boucles, donc en particulier leur période est égale à 1.

On peut donc appliquer les résultats des théorèmes 1 et 2 : la chaîne de Markov admet une unique distribution stationnaire π, et celle-ci est aussi sa distribution limite. Le rang affecté par l'algorithme PageRank est donc bien défini, quelle que soit la structure du graphe du web, et on peu le calculer de façon itérative.

Influence du facteur d'amortissement La constante α détermine le poids que l'on donne à la structure de graphe du web par rapport à un graphe complet. On comprend mieux son impact à travers le modèle du surfeur aléatoire, où elle donne la "probabilité d'ennui" du surfeur. Le nombre d'hyperliens que le surfeur suit avant de s'ennuyer suit une loi géométrique de paramètre $1 - \alpha$, c'est-à-dire que la probabilité que le surfeur suive consécutivement k hyperliens avant de sauter sur une page choisie au hasard dans tout le web est égale à $(1 - \alpha)\alpha^k$. Le nombre moyen d'hyperliens suivis avant une interruption est donc égal à

$$\sum_{k \in \mathbb{N}} k(1-\alpha)\alpha^k = \alpha(1-\alpha) \sum_{k \geq 1} k\alpha^{k-1} = \alpha(1-\alpha)\frac{1}{(1-\alpha)^2} = \frac{\alpha}{1-\alpha} = \frac{1}{1-\alpha} - 1.$$

Ce nombre croît de 0 à $+\infty$ lorsque α croît de 0 à 1. Aux deux extrêmes :

— Lorsque $\alpha = 1$, le surfeur suit systématiquement des hyperliens. On est ramené à la chaîne de Markov décrite dans la partie 3. Sa matrice de transition est donnée par $M = P$ et on n'a aucune garantie d'irréductibilité ni d'apériodicité.

— Lorsque $\alpha = 0$, on oublie complètement la structure de graphe du web. À chaque saut, le surfeur choisit une page de façon aléatoire et uniforme dans tout le web, sans tenir compte des hyperliens de la page sur laquelle il se trouve. La matrice de transition de la chaîne de Markov qui décrit son parcours est $M = \frac{1}{N}E$.

On peut de plus montrer que l'algorithme PageRank converge d'autant plus vite que α est petit. Cependant, comme on vient de le voir, choisir α "trop" petit donne peu d'importance aux hyperliens. La valeur $\alpha = 0,85$ est généralement considérée comme un bon compromis.

Références

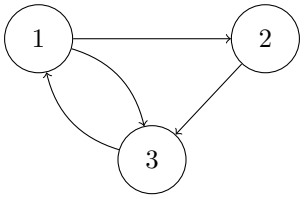
[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International Conference on World Wide Web 7*, WWW7, pages 107–117, Amsterdam, The Netherlands, The Netherlands, 1998. Elsevier Science Publishers B. V.

[2] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5) :604–632, Sept. 1999.

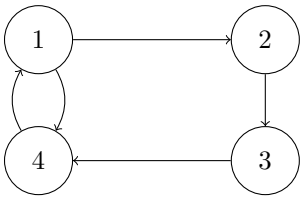
[3] F. Mathieu. *Web Graphs, PageRank-like Measurements.* Theses, Université Montpellier II - Sciences et Techniques du Languedoc, Dec. 2004. Le chapitre II.4 fait de nombreux rappels sur les chaînes de Markov. Plusieurs intuitions données dans ce cours sont tirées du chapitre 2.5.

[4] L. Page. Method for node ranking in a linked database. United States Patent US 6,285,999 B1., dated from Sep. 4, 2001. Submitted on Jan. 9, 1998.

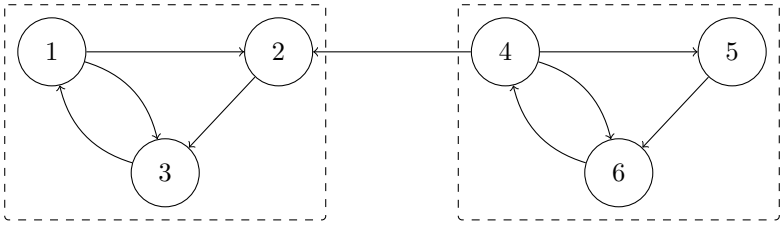
Exemples



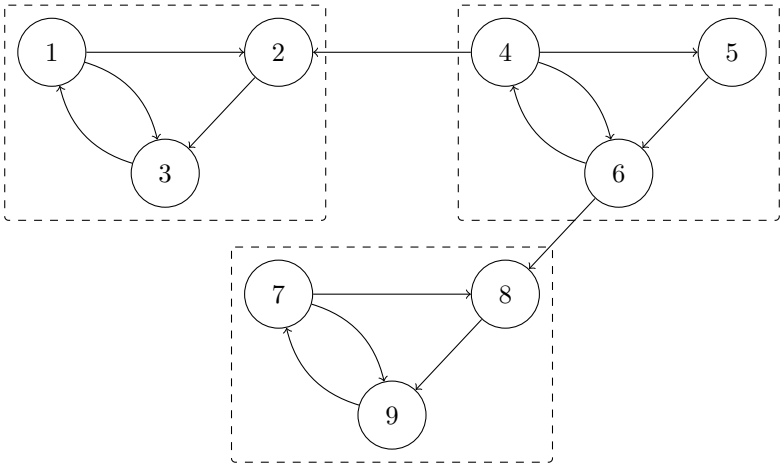
Exemple 1 – Exemple-jouet utilisé dans le brevet [4]. Ce graphe du web est fortement connexe et apériodique.



Exemple 2 – Exemple de graphe fortement connexe et périodique (de période 2).



Exemple 3 – Un exemple de graphe du web avec deux composantes fortement connexes, chacune apériodique.



Exemple 4 – Un exemple de graphe du web avec trois composantes fortement connexes, chacune apériodique.