République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Abderrahmane Mira de Béjaia

Faculté des Sciences Exactes

Département des Mathématiques

MÉMOIRE DE FIN D'ÉTUDE

En vue de l'obtention du Diplôme de Master en Mathématiques

Option : Statistique et Analyse Décisionnelle

Thème

Etude de l'algorithme Page Rank

Réalisé par :

M^{elle} ALLALI Sarra

Soutenu devant le jury composé de :

Présidente :M^{me} BOURAINE Louiza	**MCA**
Examinatrice :M^{me} KHELOUFI Karima	**MCA**
Rapporteur :M^{r} CHEMLAL Rezki	**MCB**

Promotion : 2015/2016

Remerciements

*T*out d'abord, je remercie ma famille, surtout mes parents qui m'ont épaulés, soutenus et suivis tout au long de ma vie.

A mon encadreur M^r **CHEMLAL** pour son encouragement et son suivi attentif pour la réalisation de ce travail.

Aux enseignants du département de mathématiques M^r **FARHI** , M^r et M^{me} **BOURAINE**, M^r **ARAB** et M^r **DAHMANI** pour leurs aides et leurs disponibilités pendant toute la réalisation de mon travail.

Je tiens à remercier également les membres de Jury M^{me} **BOURAINE** et M^{me} **KHELOUFI** pour avoir accepté d'évaluer mon travail.

A mes chères amis qui ont toujours été présents et fidèles.

Enfin, pour toute personne qui a contribué, de près ou de loin, à l'élaboration de ce mémoire.

Veuillez bien trouver ici l'expression de mes sincères remerciements.

Dédicaces

Ce modeste travail est dédié :

A mes chères parents.

A mon frère Lyes et mes sœurs.

A mes amis et collègues et tous ceux qui m'ont aidées.

Sarra

Table des matières

Introduction

Les premiers utilisateurs d'Internet se souviennent de moteurs de recherches comme AltaVista et Yahoo. De nos jours le moteur de recherche Google domine complètement la scène numérique. La suprématie présente de Google s'est établie à partir de 1998 par Sergey Brin et Lawrence Page.

L'efficacité de ses recherches est, en partie, due à l'utilisation d'un algorithme appelé Page Rank qui permet, à mots clefs fixés, de classer les pages web par ordre de pertinence. L'idée consiste à attribuer à chaque page un score proportionnel au nombre de fois où cette page serait visitée par un utilisateur qui, partant d'une page web quelconque, suivrait aléatoirement les hyperliens qui apparaissent au fil de sa visite.

L'algorithme Page Rank calcule un vecteur de n rangs, c'est-à-dire un vecteur de n réels positifs. Les pages web dont les rangs sont les plus élevés sont listées en premier par le moteur de recherche. Cet algorithme consiste à modéliser aussi la navigation sur Internet par une chaîne de Markov, la probabilité de suivre les liens successifs proposés par une page est considérée comme un paramètre sur lequel on peut agir.
L'algorithme Page Rank peut être analysé sur plusieurs aspects : Les propriétés et les théorèmes des chaînes de Markov nous a permis de déterminer la nature de cette probabilité, et nous a conduit au calcul de la distribution stationnaire de la chaîne de Markov étudié.

L'algorithme introduit également une probabilité d'abandon du suivi des liens hypertextes pour modéliser la possibilité d'abandonner sa navigation et de recommencer au hasard sur le web.

Le classement des pages est obtenus en calculant la distribution stationnaire de la

chaîne de Markov associée à une matrice stochastique irréductible du modèle.

Le mémoire est organisé en quatre chapitres, le premier chapitre est consacré au rappel des notions de base nécessaires pour la suite.
Le deuxième chapitre introduit le modèle Page Rank ainsi que l'étude de certaines propriétés. Le troisième chapitre est consacré à l'étude de la stabilité numérique de l'algorithme.

La particularité de l'algorithme Page Rank est de devoir gérer une masse de données très importante. En effet en 2008 l'entreprise Google affirmait référencer 40 milliard de pages web.

Les techniques de calculs et de stockage doivent être choisies de façon minutieuse, nous détaillons certaines techniques dans le chapitre 4.

Chapitre 1

Rappel et notions de base

1.1 Rappel d'algèbre linéaire et d'analyse numérique matricielle

1.1.1 Normes vectorielles et matricielles

Définition 1.1.1. Soient $\mathbf{E}$ un espace vectoriel sur $\mathbb{R}$, une norme sur $\mathbf{E}$ est une application de $\mathbf{E}$ dans $\mathbb{R}^+$ qui vérifie pour tout $x, y \in \mathbf{E}$ et tout α dans $\mathbb{R}$.

1. Homogeneities :$\forall \alpha \in \mathbb{R}\ \forall x \in \mathbf{E}\ \ \| \alpha x \| = | \alpha | \| x \|$.
2. Inégalité triangulaire : $\forall x, y \in \mathbf{E}\ \ \| x + y \| \leq \| x \| + \| y \|$.
3. Séparation : $\forall x \in \mathbf{E}\ \ \| x \| = 0_{\mathbb{R}} \Longleftrightarrow x = 0_{\mathbf{E}}$.

Exemples. Soit x un vecteur de $\mathbb{R}^n$, On a les normes suivantes :

1. $\| x \|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2}$.
2. $\| x \|_1 = \sum_{i=1}^{n} |x_i|$.
3. $\| x \|_\infty = \max_i |x_i|$.

On note $M_n(\mathbb{R})$ l'ensemble des matrices carrés d'ordre n à coefficients dans $\mathbb{R}$ ou n est un entier strictement positif.

Définition 1.1.2. On dit qu'une norme est une norme matricielle si elle vérifie les propriétés précédentes et qu'en plus elle vérifie :

$$\| A \cdot B \| \leq \| A \| \| B \|$$

Exemples. Soit A une matrice carrée, les normes matricielles classiques sont définies par :

1. $\| A \|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} | a_{ij} |$.
2. $\| A \|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} | a_{ij} |$.
3. $\| A \|_2 = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2}$.

Définition 1.1.3. Une norme matricielle et une norme vectorielle sont dites compatibles si on a pour toute matrice A et tout vecteur x l'inégalité :

$$\| Ax \| \leq \| A \| \| x \| .$$

***Remarque* 1.1.1.** Les normes matricielles 1 , 2 et infini sont respectivement compatibles avec les normes vectorielles 1, 2 et infini.

1.1.2 Matrice inversible

Définition 1.1.4. On dit qu'une matrice carrée A est inversible si et seulement si il existe une matrice B (de même format) telle que :

$$AB = BA = I.$$

B est alors appelée l'inverse de A , et est notée A^{-1}.

1.1.3 Matrice semblable

Définition 1.1.5. On dit que deux matrices A et $B \in M_n(\mathbb{R})$ sont semblables s'il existe $P \in M_n(\mathbb{R})$ une matrice inversible telle que :

$$B = P^{-1}AP.$$

1.1.4 Matrcice diagonale

Définition 1.1.6. Une matrice A de est dite diagonale si tous les éléments en dehors de la diagonale sont nuls $a_{ij} = 0$ pour $i \neq j$ et elle s'écrit sous la forme suivante :

$$\begin{pmatrix} \lambda_1 & & (0) \\ & \ddots & \\ (0) & & \lambda_n \end{pmatrix}$$

Proposition 1.1.1. *Une matrice est dite diagonalisable s'il existe une matrice $P \in M_n(\mathbb{R})$ tel que la matrice $P^{-1}AP$ est une matrice diagonale.*

1.1.5 Matrice triangulaire

Définition 1.1.7. Une matrice A de $M_n(\mathbb{R})$ est dite trigonalisable, si A est semblable à une matrice triangulaire supérieure. C'est-à-dire, s'il existe une matrice inversible P de $M_n(\mathbb{R})$ et une matrice triangulaire supérieure T à coefficients dans $\mathbb{R}$ telles que

$$A = PTP^{-1}.$$

1.1.6 Matrice orthogonale

Définition 1.1.8. On appelle matrice orthogonale une matrice dont les colonnes sont orthonormées. C'est à dire les matrices O telles que :

$$^tOO = I;$$

où ^tO désigne la transposée de O.

1.1.7 Matrice tridiagonale

Définition 1.1.9. Une matrice $A \in M_n(\mathbb{R})$ est dite matrice tridiagonale si pour tout coefficient a_{ij} on a :

$$a_{ij} = 0 \ \ pour\ tous\ (i,j)\ \ tels\ que \quad \mid i-j \mid > 1.$$

1.1.8 Conditionnement d'un système linéaire

Un système d'équations linéaires est dit mal conditionné si certaines équations sont quasi redondantes. Un système mal conditionné augmente les erreurs d'arrondi à cause de la soustraction de nombres proches.

Définition 1.1.10. On appelle nombre de condition ou conditionnement de la matrice A et on note $Cond(A)$ la valeur

$$Cond(A) = \| A \| \| A^{-1} \| .$$

Un système d'équations linéaires est bien conditionné si le conditionnement de sa matrice des coefficients est proche de 1 et mal conditionné si $Cond(A)$ est très grand par rapport à 1.

Proposition 1.1.2. *Soit A une matrice inversible. Soit b un vecteur non nul. Notons δA et δb de petites perturbations sur A et b respectivement . On a :*

1. *Si x et δx sont les solutions du système $Ax = b$ et $A(x + \delta x) = b + \delta b$ on a :*

$$\frac{\| \delta x \|}{\| x \|} \leq Cond(A) \frac{\| \delta b \|}{\| b \|}$$

2. *Si x et $x + \delta x$ sont les solutions du système $Ax = b$ et $(A + \delta A)(x + \delta x) = b$ on a :*

$$\frac{\| \delta x \|}{\| x + \delta x \|} \leq Cond(A) \frac{\| \delta A \|}{\| A \|}$$

Preuve

1. On commence par remmarquer

$$A\delta x = b \Longrightarrow \| \delta x \| \leq \| A^{-1} \| \| \delta b \| .$$

   Or on a également

$$\| b \| \leq \| A \| \| x \| \Longrightarrow \frac{1}{\| x \|} \leq \frac{\| A \|}{\| b \|}.$$

   D'où :

$$\frac{\| \delta x \|}{\| x \|} \leq Cond(A) \frac{\| \delta b \|}{\| b \|}.$$

2. On remarque que :

$$A\delta x + \delta A(x + \rho x) = 0 \Longrightarrow \| \delta x \| \leq \| A^{-1} \| \| \delta A \| \| x + \delta x \| .$$

   D'où on déduit

$$\frac{\| \delta x \|}{\| x + \delta x \|} \leq Cond(A) \frac{\| \delta A \|}{\| A \|}.$$

***Remarque* 1.1.2.** Nous utiliserons par la suite des preuves qui reposent sur cette technique.

1.1.9 Valeurs propres et vecteurs propres

Définition 1.1.11. Un vecteur colonne non nul x est un vecteur propre d'une matrice carrée A, s'il existe un scalaire λ tel que

$$Ax = \lambda x$$

λ est alors la valeur propre de A.

Définition 1.1.12. On appelle l'ensemble des valeurs propres d'une matrice carrée A le spectre de A et on le note $Sp(A)$.

Théorème 1.1.1. *Deux matrices semblables ont les mêmes valeurs propres.*

Proposition 1.1.3. *Si une matrice A d'ordre n^2 , admet n valeurs propres deux à deux distinctes, alors A est diagonalisable.*

1.1.9.1 Recherche des valeurs propres et vecteurs propres

La recherche des valeurs propres en cherchant les racines du polynôme caractéristique est un problème numériquement instable. C'est pour ça qu'on a recours aux méthodes numériques. Nous allons exposer ici trois méthodes classiques.

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice diagonalisable. On note $\lambda_1 \cdots \lambda_n$ les valeurs propres de A et $\nu_1 \cdots \nu_n$ des vecteurs propres unitaires associés.

1.1.9.2 Méthode de la puissance

La methode de la puissance permet d'obtenir la valeur propre de plus grand module $\mid \lambda_1 \mid$.
Elle consiste à se donner un vecteur initial $x^{(0)}$ tel que $\parallel x^{(0)} \parallel = 1$ et on construit une suite $(x)_{k\in\mathbb{N}}^{(k+1)}, (x)_{k\in\mathbb{N}}^{(k)}$ de la façon suivante :

$$\begin{cases} (x)^{(0)} : \parallel x^{(0)} \parallel = 1 \\ (y)^{(k+1)} = Ax^{(k)} \\ (x)^{(k)} = \frac{y^{(k)}}{\|y^{(k)}\|} \end{cases}$$

La valeur propre approchée est la dernière norme $\parallel y^{(k)} \parallel$ calculée et le vecteur propre approché associé est le vecteur $x^{(k)}$.

Théorème 1.1.2. *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice diagonalisable. On note $\lambda_1 \cdots \lambda_n$ les valeurs propres de A et $\nu_1 \cdots \nu_n$ des vecteurs propres unitaires associés qui vérifie :*

$$\mid \lambda_1 \mid > \mid \lambda_2 \mid > \mid \lambda_3 \mid \cdots > \mid \lambda_n \mid .$$

En suppose $\alpha_1 \neq 0$, il existe $C > 0$ tel que :

$$| \lambda^k - \lambda_1 | \leq C \left(\frac{| \lambda_1 |}{| \lambda_2 |} \right)^k .$$

Démonstration : [9]page 172.

1.1.9.3 Méthode QR

La méthode QR est utilisée pour approcher les valeurs propres d'une matrice A donnée. L'idée de base est de transformer la matrice A en une matrice semblable pour laquelle le calcul des valeurs propres est plus simple.

Si on construit une matrice semblable à une matrice triangulaire, on obtient toutes les valeurs propres.

Théorème 1.1.3. *Soit $A \in \mathbb{R}^{n\times n}$, on se donne une matrice orthogonale $Q^{(0)} \in \mathbb{R}^{n\times n}$ et on pose*

$$T^{(0)} = (Q^{(o)})^{-1} A Q^{(0)}$$

Les itérations de la méthode QR s'écrivent pour $k = 1, 2, \cdots$, jusqu'à convergence :

Déterminer $Q^{(k)}$ et $R^{(k)}$ telles que $Q^{(k)} R^{(k)} = T^{(k-1)}$.

Puis on pose

$$T^{(k)} = R^{(k)} Q^{(k)}$$

A chaque étape $k \leq 1$, on écrit la matrice $T^{(k-1)}$ sous la forme de produit d'une matrice orthogonale $Q^{(k)}$ et une matrice triangulaire superieure $R^{(k)}$. La second phase est un produit matricielle :

$$\begin{aligned} T^{(k)} &= R^{(k)} Q^{(k)} = (Q^{(k)})^{-1} (Q^{(k)} R^{(k)}) Q^{(k)} = (Q^{(k)})^{-1} T^{(k-1)} Q^{(k)} \\ &= (Q^{(0)} Q^{(1)} \cdots Q^{(k)})^{-1} A (Q^{(0)} Q^{(1)} \cdots Q^{(k)}) \end{aligned}$$

Concergence de la méthode QR

Soit $A \in \mathbb{R}^{n\times n}$ telle que

$$| \lambda_1 | > | \lambda_2 | > | \lambda_3 | \cdots > | \lambda_n |$$

Alors

$$\lim_{k\longrightarrow+\infty} T^{(k)} = \begin{pmatrix} \lambda_1 & t_{12} & \cdots & t_{1n} \\ 0 & \lambda_2 & t_{23} & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

Le taux de convergence est de la forme

$$\mid t^k_{i,i-1} \mid = O\left(\mid \frac{\lambda_i}{\lambda_{i-1}} \mid \right)^k \quad i = 2, \cdots, n \;\; et \; k \longrightarrow +\infty$$

Coût de la méthode QR

Une itération de la méthode QR demande de l'ordre $O(n^3)$ opérations.

1.1.9.4 Méthode de Givens-Householder

Définition 1.1.13. On appelle matrice de Householder associée au vecteur $v \neq 0$, la matrice

$$H(v) = I - 2\frac{vv^t}{\mid v \mid_2^2}$$

Par convention $H(0) = I$.

La méthode de Givens-Householder a été proposé en 1958, c'est l'association de deux algorithmes. La méthode de Householder transforme la matrice initiale A sous la forme tridiagonale symétrique, l'algorithme de Givens calcule les valeurs propres d'une matrice tridiagonale symétrique.

Le principe La méthode de Givens-Householder comprend deux étapes :

a) Réduction : On détermine, par la méthode de réduction de Householder, une matrice O orthogonale (obtenue comme produit de $n-2$ matrices de Householder) telle que la matrice O^tAO soit tridiagonale, i.e.

$$O^tAO = \begin{pmatrix} b_1 & c_1 & 0 & \cdots & \cdots \\ c_1 & b_2 & c_2 & \cdots & \cdots \\ 0 & c_2 & b_3 & c_3 & \cdots \\ \vdots & \cdots & \ddots & \ddots & \cdots \\ 0 & \cdots & \cdots & b_{n-1} & c_{n-1} \\ 0 & \cdots & \cdots & c_{n-1} & b_n \end{pmatrix} \quad avec \;\; b_i \in \mathbb{R} \; et \;\; c_i \in \mathbb{R}^*$$

b) Bissection : On est ainsi ramené au calcul des valeurs propres d'une matrice symétrique tridiagonale, qui s'effectue par la méthode de bissection de Givens.

Coût de la méthode de Givens-Householder

Une itération de la méthode de Givens-Householder demande de l'ordre $O(\frac{3}{2}n^2)$ opérations.

1.2 Théorème de point fixe dans $\mathbb{R}^n$

Définition 1.2.1. On appelle point fixe d'une fonction $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ tout vecteur $r \in \mathbb{R}^n$ vérifiant $f(r) = r$.

Définition 1.2.2. Une fonction $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ est dite contractante de rapport $k < 1$ (par rapport a la norme $\| \, . \, \|$) si elle vérifie

$$\| f(x) - f(y) \| \leq k \, \| x - y \| \quad pour \;\; tout \;\; x, y \in \mathbb{R}^n.$$

Théorème 1.2.1. *Si $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ est une fonction contractante de rapport $k < 1$, alors :*

1. *Il existe un et un seul point $r \in \mathbb{R}^n$ verifiant $f(r) = r$.*
2. *Pour toute valeur initiale $x_0 \in \mathbb{R}^n$ la suite iterative $x_{m+1} = f(x_m)$ converge vers r.*
3. *On a $\| x_m - r \| \leq k^m \, \| x_0 - r \|$, la convergence vers r est donc au moins aussi rapide que celle de la suite geometrique $\frac{k^m}{1-k}$ vers 0. Pour le calcul sur ordinateur on a l'estimation de l'écart.*

$$\| x_m - r \| \leq \frac{k}{1-k} \, \| x_m - x_{m-1} \| \, .$$

1.3 Processus stochastiques et chaînes de Markov à temps discret

Cette section est destinée à un rappel des propriétés de bases des chaînes de Markov qui constituent un élément important dans le modèle Page Rank.

1.3.1 Propriétés et caractéristiques des processus stochastiques

1.3.1.1 Variable aléatoire

Définition 1.3.1. Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité.

L'application $X : \Omega \longrightarrow \mathbb{R}$ est appelée variable aléatoire si

$$\forall\ B \in \mathbb{B}_{\mathbb{R}}\ \ X^{-1}(B) = \{w \in \omega\ \ X(w) \in B\} \in \mathcal{F}.$$

1.3.1.2 Espace d'état

Définition 1.3.2. On appelle espace des états l'ensemble S où les variables aléatoires prennent leurs valeurs.

1.3.1.3 Processus stochastique

Définition 1.3.3. Soit un espace d'état S. On appelle processus stochastique l'ensemble de variables aléatoires $(X_t)_{t\in T}$ dans S où le paramètre t désigne généralement le temps, et X_t représente la position du processus à l'instant t.

***Remarque* 1.3.1.** Si $t \in \mathbb{N}$ fini où infini dénombrable, le processus est dit à temps discret. Si $t \in \mathbb{R}$ le processus est dit continu.

1.3.1.4 Processus stationnaire

Définition 1.3.4. On appelle un processus $(X_t)_{t\in T}$,un processus strictement stationnaire si les fonctions de répartitions des familles de variables aléatoires $(X_{t_1+h}, X_{n_2+h}, \ldots, X_{n_k+h})$ et $(X_{n_1}, X_{n_2}, \ldots, X_{n_k})$ soient les mêmes.

1.3.2 Chaîne de Markov à temps discrèt

Définition 1.3.5. Une chaîne de Markov est un processus stochastique $(X_t)_{t\in T}$ qui peut être soit discret où continu, satisfaisant la propriété de Markov suivante :

$$\begin{aligned} P_{ij}^{(n)} &= \mathbb{P}(X_{n+1}=j|X_n=i, X_{n-1}=i_{n-1}, \cdots, X_0=i_0) \\ &= \mathbb{P}(X_{n+1}=j|X_n=i) \end{aligned}$$

Le processus markovien est un processus sans mémoire car l'état futur du processus ne dépend que de l'état présent de ce processus et non pas de ceux du passés.

1.3.2.1 Matrice stochastique

Définition 1.3.6. Soit une matrice carrée non nulle P d'ordre n. On dit que la matrice P est stochastique si la somme de chaque ligne est égale à 1.
i.e : $\forall(i,j)\in S$ On a

$$\sum_{j=1}^{n} P_{ij} = 1, \quad 0 \leq P_{ij} \leq 1$$

1.3.2.2 Matrice anti stochastique

Définition 1.3.7. Soit une matrice carrée non nulle A d'ordre n n. On dit que la matrice A est anti stochastique si la somme de chaque colonne est égale à 1.
i.e : $\forall(i,j)\in S$ On a

$$\sum_{j=1}^{n} P_{ij} = 1, \quad 0 \leq P_{ij} \leq 1$$

Proposition 1.3.1. *Soit une matrice stochastique $P_{n\times n}, \lambda_1, \lambda_2, \cdots, \lambda_n$ valeurs propres de P alors on a :*

1. *P admet 1 comme valeur propre.*
2. *Tout autre valeur propre est de module inférieur à 1.*

$$1 > |\ \lambda_1\ | \geq |\ \lambda_2\ | \geq \cdots \geq |\ \lambda_n\ |$$

3. *Il existe un vecteur propre π de P, associé à la valeur propre 1,qui définit une distribution de probabilité*

$$\pi = (\pi_1, \pi_2, \ldots, \pi_n)$$

avec $\pi_k \geq 0$ et $\sum_{k=1}^{n} \pi_k = 1$

Démonstration

1. Soit P une matrice stochastique d'ordre n qui s'écrit sous la forme

$$P = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & & & \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{pmatrix}$$

Supposons un vecteur v non nul

$$v = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

soit la multiplication suivante

$$\begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & & & \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

car

$$\sum_{j=1}^{n} P_{ij} = 1$$

donc $P.v = 1.v$ ce qui montre que 1 est une valeur propre de P.

2. Soit P une matrice stochastique, $\lambda \in \mathbb{R}$ une valeur propre de P.

Soit v un vecteur non nul associé à la valeur propre λ

$$v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

$$P.v = \lambda.v \Longleftrightarrow \begin{cases} P_{11}v_1 + P_{12}v_2 + \cdots + P_{1n}v_n = \lambda v_1 \\ P_{21}v_1 + P_{22}v_2 + \cdots + P_{2n}v_n = \lambda v_2 \\ \vdots \\ P_{n1}v_1 + P_{n2}v_2 + \cdots + P_{nn}v_n = \lambda v_n \end{cases}$$

On a $\forall i \in \{1, \cdots, n\}$

$$\lambda x_i = P_{i1}x_1 + P_{i2}x_2 + \cdots + P_{in}x_n$$

En introduisant la valeur absolue, par l'inégalité triangulaire on obtient

$$\begin{aligned} \mid \lambda \mid\mid x_i \mid &\leq P_{i1} \mid x_1 \mid + P_{i2} \mid x_2 \mid + \cdots + P_{in} \mid x_n \mid \\ &\leq P_{i1} \max \mid x_k \mid + P_{i2} \max \mid x_k \mid + \cdots + P_{in} \max \mid x_k \mid \\ &\leq \max_{1 \leq k \leq n} \mid x_k \mid (P_{i1} + P_{i2} + \cdots + P_{in}) \end{aligned}$$

Donc

$$\forall i \in \{1, \cdots, n\} \quad \mid \lambda \mid\mid x_i \mid \leq \max \mid x_k \mid . \tag{1.1}$$

or

$$\max_{1 \leq i \leq n} \mid x_i \mid = \mid x_{k_0} \mid .$$

Posons dans (1.1) $i = k_0$

Donc

$$\mid \lambda \mid\mid x_{k_0} \mid \leq \mid x_{k_0} \mid .$$

D'où

$$\mid \lambda \mid \leq 1.$$

Corollaire 1.3.1. *Toute matrice anti stochastique vérifie la proposition 1.3.1 .*

Preuve :La preuve se repose sur le fait que la transposée d'une matrice anti stochastique est une matrice stochastique, et la matrice transposée possède exactement les même valeurs propre de sa matrice associée.

1.3.2.3 Matrice de transition

Définition 1.3.8. Une matrice de transition d'une chaîne de Markov est une matrice stochastique.

$$P = P_{ij} = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1n} \\ P_{21} & P_{22} & \cdots & P_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ P_{n1} & P_{n2} & \cdots & P_{nn} \end{pmatrix}$$

qui vérifie les propriétés suivantes :

$$\sum_{j=0}^{n} P_{ij} = 1 \qquad \forall i, j \in S.$$

$$0 \leq P_{ij} \leq 1 \qquad \forall i, j \in S.$$

$\mathbb{P}(X_{t+1} = j | X_t = i)$ est une probabilité de transition à une étape.

Définition 1.3.9. Soit une chaîne de Markov à espace d'état S.
La probabilité de transition $P_{ij}(t) = \mathbb{P}(X_{t+1} = j | X_t = i)$ est la probabilité d'être dans l'état j à l'instant $t+1$ sachant que la chaîne était dans l'état i à l'instant t, donc c'est la probabilité de se déplacer de l'état i à l'état j en une transition.

De ce fait le calcul de la probabilité en n transitions est comme suit :

$$P_{ij}^{(n)} = \sum_{k \in S} P_{ik}^{(n-1)} P_{kj}. \tag{1.2}$$

***Remarque* 1.3.2.** Si $P_{ij}(t) = P_{ij}$, $(X_t)_{t \in T}$ est dit homogène.

1.3.2.4 Graphe de transition

Définition 1.3.10. Un graphe $G = (X, U)$ est déterminé par la donnée, d'un ensemble X dont les éléments sont appelés des sommets où des nœuds , d'un ensemble U dont les éléments sont des couples ordonnés de sommets appelés des arcs.

Graphiquement, les sommets sont représentés par des points et $u = (i, j)$ sera représenté par une flèche allant du point i vers le point j.

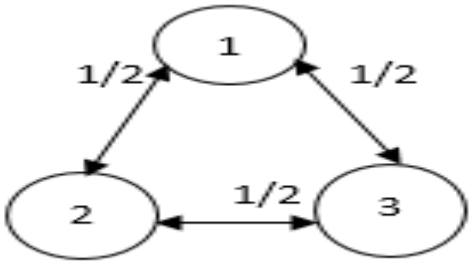

FIGURE 1.1 – Graphe de transition d'une chaîne de Markov

***Remarque* 1.3.3.** Un graphe de transition est déterminé par, un ensemble de sommets qui représente les états de la chaîne de Markov et un ensemble d'arcs qui représente les probabilités de transitions strictement positives.

1.3.2.5 La loi de probabilité d'une chaîne de Markov

Nous introduisons maintenant les probabilités d'état

$$\pi_i(n) = P(X_n = i) \qquad (n = 0, 1, ...et \;\; i = 1, 2, ...).$$

La distribution de X_n peut alors être écrite sous forme du vecteur ligne

$$\pi(n) = (\pi_1(n), \pi_2(n), ...)$$

où la somme des termes vaut 1.Pour calculer $\pi(n)$, il faut connaitre soit la valeur prise par X_0, c'est-à-dire l'état initial du processus, soit sa distribution initiale $\pi(0)$.

D'après le théoréme des probabilités totales,on a alors

$$\pi_k(n) = \sum_{i \in S} \pi_i(0) P_{ik}^{(n)}. \tag{1.3}$$

1.3.2.6 Le régime transitoire et le régime permanent

L'étude de régime transitoire d'un phénomène aléatoire est de déterminer les probabilités d'état $\pi_i(n) = P(X_n = i)$ en fonction de nombre n des transitions. En terme techniques, ceci revient à étudier le régime transitoire du phénomène aléatoire en question.

L'étude de régime permanent c'est l'étude de phénomène lorsque $n \longrightarrow \infty$, dans ce cas on peut dire qu'il y a une sorte de disparition du hasard, le processus devient prévisible. Une fois on observe plusieurs fois le phénomène pendant une certaine durée, la répartition des probabilités d'états est égale.

Dans un processus permanent on constate souvent que la distribution $\pi(n)$ converge vers une distribution limite si $n \longrightarrow \infty$. Contrairemenr au régime transitoire, le régime permanent n'est pas influencer par le choix de la distribution initiale.

En terme formels, on dit qu'une chaîne de Markov converge vers π possède une distribution limite π si

$$\lim_{n \longrightarrow \infty} \pi(n) = \pi.$$

1.3.3 Existence d'une distribution limite

La recherche de condition assurant la convergence de $\pi(n)$ vers une distribution limite constitue un chapitre important en théorie des chaîne de Markov. De nombreux théorèmes limites ont été établis.

Théorème 1.3.1. *Si la matrice de transition P est telle qu'une au moins de ses puissances n'a que des termes strictement positifs,alors*

$$\pi(n) \longrightarrow \pi.$$

quelque soit la distribution initiale $\pi(0)$,

$$\pi = \pi P^*.$$

lorsque $n \longrightarrow \infty$. π est un vecteur de probabilité strictement positif, et P^* est une matrice dont toutes les lignes sont identiques au vecteur limite π. En plus

$$\pi P^* = \pi.$$

Théorème 1.3.2. *Si la valeur propre 1 de P est simple (c'est-à-dire de multiplicité 1) et si toutes autres valeurs propre de P est de module strictement inférieur à 1, alors*

$$P^n \longrightarrow P^*.$$

$$P^*\pi = \pi.$$

1.3.4 Distribution stationnaire d'une chaîne de Markov

Définition 1.3.11. On appelle distribution stationnaire d'une chaîne de Markov associé à la matrice de transition P la distribution de probabilité π_i telle que

$$\forall i, \quad \pi_i = \sum_{j \in S} P_{ij}\pi_j \quad et \quad \sum_{i \in S} \pi_i = 1.$$

1.3.5 Classifications des états

1.3.5.1 Irréductibilité

Définition 1.3.12. Un état j est accessible à partir d'un état i s'il existe un entier m tel que $P_{ij}^{(m)} > 0$.
Deux états i et j mutuellement accessible sont appelés communicants, on écrit : $i \longleftrightarrow j$; il existe donc deux entiers m et n tels que $P_{ij}^{(m)} > 0$ et $P_{ji}^{(n)} > 0$.
La propriété état communicants définie sur l'ensemble des états de la chaîne une relation d'équivalence.
A partir d'une relation d'équivalence, on définit sur l'ensemble des états, des classes d'équivalences.

Définition 1.3.13. Une chaîne de Markov est dite irréductible si elle ne possède qu'une seul classe d'équivalence. C'est-à-dire tous ses états communiquent entre eux.

1.3.5.2 Temps moyen de retour

Définition 1.3.14. Soit un espace d'état S, soit un état $i \in S$

La variable aléatoire T_i définie par :

$$T_i = \inf\{n \geq 1, X_n = i\}$$

est appelée "temps moyen du premier retour" à i lorsque la chaîne part de i.

1.3.5.3 Réccurence et transiance

Définition 1.3.15. Un état i est dit récurrent si partant de i on y revient presque surement en temps fini :

$$\mathbb{P}(T_i < +\infty | X_0 = i) = 1$$

Dans le cas $\mathbb{P}(T_i = +\infty / X_0 = i) > 0$ $\quad i$ est dit transitoire. i.e si avec une probabilité positive on le quitte pour ne jamais y revenir.
On dit qu'un état i est récurrent positif lorsque le temps moyen de retour en i est fini :

$$\mu_i = E(T_i | X_0 = i) < +\infty$$

Dans le cas où $\mu_i = E(T_i | X_0 = i) = +\infty$ $\ i$ est dit récurent nul.

1.3.5.4 Périodicité et appériodicité

Définition 1.3.16. On appelle période d'un état i l'entier $d(i)$ définie par :

$$d(i) = PGCD\{n \geq 1, \quad P_{ii}^{(n)} > 0\}$$

lorsque $d(i) = 1 \quad i$ est dit apériodique.

1.3.5.5 Ergodicité

Définition 1.3.17. Un état i qui est à la fois récurent positif et apériodique est dit ergodique.

Soit (X_n) une chaîne de Markov irréductible à espace d'état discret S ,i un état de la chaîne de Markov avec $i \in S$.

Théorème 1.3.3. *a) Si i est un état transitoire (où récurent nul) alors*

$$\lim_{n \longrightarrow \infty} P_{ii}^{(n)} = 0$$

b) Si i est un état ergodique (récurent positif et apériodique) alors

$$\lim_{n \longrightarrow \infty} P_{ii}^{(n)} = \frac{1}{\mu_i} \quad avec \quad \mu_i = E[T_i | X_0 = i].$$

Théorème 1.3.4. *a) Si i est un état transitoire (où récurent nul) alors*

$$\lim_{n \longrightarrow \infty} P_{ij}^{(n)} = 0, \quad \forall \ j \in S.$$

b) Si i est un état ergodique (récurent positif et apériodique) alors

$$\lim_{n \longrightarrow \infty} P_{ij}^{(n)} = \frac{P(T_j < \infty | X_0 = i)}{\mu_j}.$$

1.3.6 Théorème d'existence et d'unicité de la distribution stationnaire

Théorème 1.3.5. *Une chaîne de Markov irréductible possède une distribution stationnaire si et seulement si ses états sont tous ergodiques.*

Alors cette distribution coïncide avec la distribution limite solution de :

$$\pi_i = \sum_{j \in S} \pi_i P_{ji} \quad \forall i \in S$$

$$\sum_{i \in S} \pi_i = 1 \quad \forall i \in S \ \pi_i > 0$$

Preuve

Soit (X_n) une chaîne de Markov irréductible.

$\Longrightarrow$) par contraposé, si tous les états sont transitoires (où reccurent nuls) alors

$\forall i, j \in S, \lim_{n \longrightarrow \infty} P_{ij}^{(n)} = 0 \Longrightarrow$ il ne peut exister une distribution stationnaire π_j car

$$\forall j \in S : \pi_j = \sum_{j \in S} \pi_i P_{ij}^{(n)}$$

(Propriétés de la distribution stationnaire)

et quand $n \longrightarrow \infty, \forall j \in S : \pi_j = 0$ (contradiction)

D'où il existe une distribution stationnaire dans une chaîne de Markov irréductible alors tous les états sont ergodiques.

$\Longleftarrow$)considérons une chaîne de Markov irréductibles dont tous les états sont ergodiques, d'aprés le théorème 1.3.4

$$\forall i \in S \lim_{n \longrightarrow \infty} P_{ij}^{(n)} = \frac{1}{\mu_j} = \pi_j$$

Montrons que $\pi_j = \{frac1\mu_j\}$ est une distribution stationnaire.

π_j est une distribution de probabilitée

en effet $\forall j \in S \frac{1}{\mu_j} > 0$

Soit M un nombre entier arbitraire

On a

$$P_{ij}^{(n+m)} = \sum_{k \in S} P_{ik}^{(m)} P_{kj}^{(n)} \geq \sum_{k \in S} P_{ik}^{(m)} P_{kj}^{(n)}$$

Quand $m \longrightarrow \infty$ on aura

$$\pi_j \geq \sum_{k=1}^{M} \pi_k P_{kj}^{(n)}$$

et quand $M \longrightarrow \infty$

$$\pi_j \geq \sum_{k=1}^{\infty} \pi_k P_{kj}^{(n)}$$

Montrons que cette relation est une égalitée.

Raisonons par l'absurde supposons qu'il existe $j \in S$ tel que

$$\pi_j > \sum_{k=1}^{\infty} \pi_k P_{kj}^{(n)}$$

alors

$$\sum_{j \in S} \pi_j > \sum_{j \in S} \sum_{k \in S} \pi_k P_{kj}^{(n)}$$

or

$$\begin{aligned} \sum_{j \in S} \sum_{k \in S} \pi_k P_{kj}^{(n)} &= \sum_{j \in S} P_{kj}^{(n)} \sum_{k \in S} \pi_k \\ &= \sum_{k \in S} \pi_k (\sum_{j \in S} P_{kj}^{(n)} = 1) \end{aligned}$$

alors

$$\sum_{j \in S} \pi_j > \sum_{k \in S} \pi_k$$

(contradiction)

D'où $\forall j \in S$

$$\pi_j = \sum_{k \in S} \pi_k P_{kj}^{(n)}$$

Quand $n \longrightarrow \infty$

$$\lim_{n \longrightarrow \infty} P_{kj}^{(n)} = \pi_j$$

et

$$\pi_j = (\sum_{k \in S} \pi_k) \pi_j.$$

d'où

$$\sum_{k \in S} \pi_k = 1 \quad car \quad \pi_j > 0.$$

Donc $\{\frac{1}{\mu_j}\} = \{\pi_j\}$ est une distribution de probabilité.

Montrons qu'elle est unique.

Supposons qu'il existe une autre distribution stationnaire π_j^* alors par définition

$$\forall i \in S, \pi_i^* = \sum_{j \in S} \pi_j^* P_{ji}^n.$$

et quand

$$n \longrightarrow \infty \quad \pi_i^* = \sum_{j \in S} \pi_j^* \frac{1}{\mu_i} = \frac{1}{\mu_i}.$$

Donc

$$\forall i \in S : \{\pi_i^*\} = \{\frac{1}{\mu_i}\}.$$

D'où l'unicité.

***Remarque* 1.3.4.** Ce théorème est très important en pratique lorsque l'espace d'état S est fini.

En effet, il est facile de voir si une chaîne de Markov est irréductible. Si elle est alors sa distribution limite se calculera en résolvant le système linéaire suivant :

$$\sum_{i \in S} \pi_i P_{ij} = \pi_j$$
$$\sum_{j \in S} \pi_j = 1$$

1.3.7 Etat absorbant

Définition 1.3.18. Un état i est dit absorbant si le processus ne peut plus quitter cet état une fois qu'il est rentré, en d'autre terme si $P_{ii} = 1$. Une chaîne de Markov est dite absorbante si elle comprend en moins un état absorbant.

1.3.8 Résumé

Propriétés d'une chaîne de Markov :

1. Aucune chaine finie n'est transiente.
2. Dans une chaîne de Markov finie, il n'existe pas d'état récurrent nul.
3. Toute chaîne de Markov finie, possède au moins une classe récurrente positive, donc au moins une distribution stationnaire portée par cette classe.
4. Une chaîne de Markov transiente ne possède pas de distribution stationnaire.

Chapitre 2

Le modèle Page Rank

2.1 Principe de base de l'algorithme Page Rank

Les documents hypertextes fournissent des liens les uns vers les autres, ainsi on peut considérer le web comme un immense graphe dont les sommets sont les pages web qui seront numérotés de 1 à n et les liens sont les arcs.

Le principe des classements des pages web est basée sur un système de citations.

Plus une page reçoit de liens externes plus elle est importante. Une page en émettant beaucoup de liens diminue son importance.

Si on note $\mu(i)$ le poids où l'importance d'une page i, on peut exprimer le principe de base de l'algorithme par l'idée d'un vote où chaque page vote en fonction de son poids.

Ainsi si une page émet des liens vers un certain nombre de page on estime qu'elle vote pour ces pages en distribuant son propre poids à égalité entre chaque page.

Soit i une page web, on note ℓ_i le nombre des liens émis par la page (i) et $\mu(i)$ le poid de la page web i .

Le poid d'une page web peut ainsi être modélisé par la formule suivante :

$$\mu(i) = \sum_{j \longrightarrow i} \frac{1}{\ell_j} \mu(j) \quad i = 1, \cdots, n. \tag{2.1}$$

où $j \longrightarrow i$ si la page j pointe un lien vers la page i.

Soit la matrice $H = (h_{ij})$ la matrice definies par :

$$h_{ij} = \begin{cases} \frac{1}{\ell_j} & si \quad j \longrightarrow i \quad * \\ 0 & sinon \end{cases}$$

On peut écrire la formule $(*)$ pour la Fig 2.1 sous forme matricielle :

$$H = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

2.2 Un premier exemple.

Nous supposons que nous disposons d'une toile de cinq pages, notre espace d'état est $S = \{1, 2, ..., 5\}$.

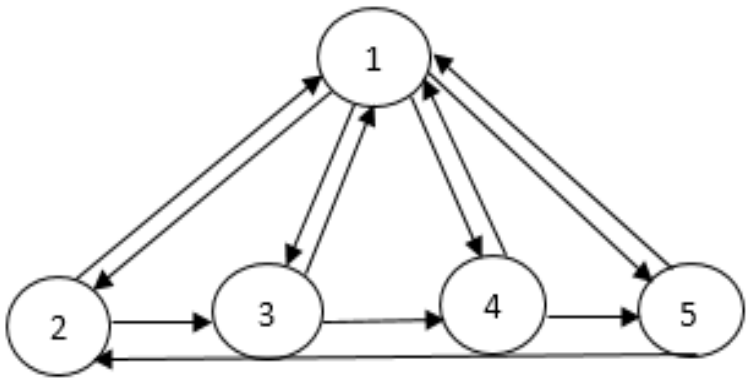

FIGURE 2.1 – Une toile de cinq pages et ses liens

Les arêtes tracés entre les pages indiquent par exemple que :

- Les liens de la page 1 pointent vers les pages 2,3,4 et 5.
- Les liens de la pages 5 pointent vers les pages 1 et 2.
- Les liens de la page 3 pointent vers les pages 1 et 4.

2.2.1 Marche aléatoire sur le web

Supposons qu'une personne navigue sur notre toile en cliquant sur les liens qui lui sont offerts. Par exemple s'il se trouve à la page 3 il aura le choix de cliquer sur le lien vers la page 4 où sur le lien vers la page 1 . Les évènements étant supposés équiprobables, il y a 50% de chances de cliquer sur la page 1 et autant de cliquer sur la page 4. Ainsi après un clic le surfeur va se trouver soit sur la page 1 soit sur la page 4. A partir de là on peut calculer la probabilité conditionnelle de se trouver sur chaque page. Voir Fig 2.2.

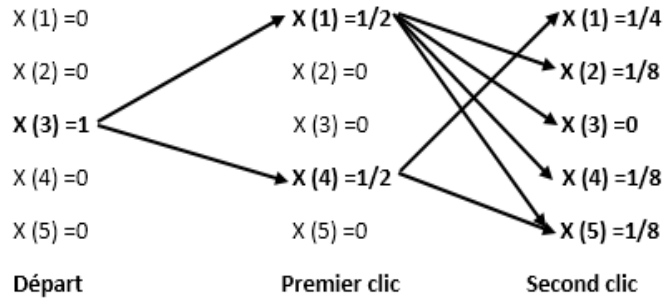

FIGURE 2.2 – Les deux premiers pas de la marche aléatoire sur un réseau de 5 pages.

La Fig 2.2 répond à cette question pour les deux premiers clics d'un promeneur commençant à la page 3. Cette page offre deux liens, et le promeneur ne pourra aller qu'aux pages 1 et 4. Ainsi après le premier clic, il se trouvera à la page 1 avec une probabilité $\frac{1}{2}$ et à la page 4 avec une probabilité $\frac{1}{2}$ c'est ce qui est indiqué dans la colonne médiane de la Fig 2.2 par les deux relations

$$X(1) = \frac{1}{2}, \quad X(4) = \frac{1}{2}.$$

alors que les autres relations

$$X(2) = 0, \quad X(3) = 0, \quad X(5) = 0.$$

indiquent qu'après le premier clic, le promeneur ne pourra pas être aux pages 2,3 et 4 car aucun lien ne mène de la page 3 vers ces dernières.

Les chemins de la Fig 2.2 indiquent les probabilités de visiter chacune des cinq pages.

On peut également procéder au même calcul en supposant à chaque fois des pages de départ différentes.

On veut calculer les probabilités de visites des cinq pages après deux clics X_1 et X_2 sachant que le promeneur commence par la page 3. Ceci est exprimé par une probabilité conditionnelle P_{ij}.

Où $P(X_1 = 1|X_0 = 3)$ désigne la probabilité que le promeneur se trouve à la page 1 après le premier clic ($X_1 = 1$)s'il se trouvait en 3 au départ ($X_0 = 3$).Ainsi, on peut modélisé la navigation sur Internet par une chaîne de Markov. Voir Fig 2.4.

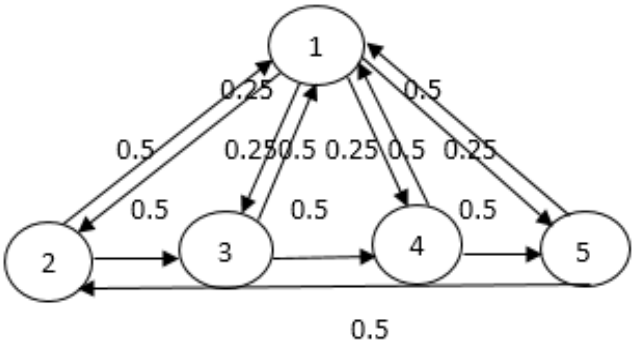

FIGURE 2.3 – Graphe d'une chaîne de Markov associée à un réseau de 5 pages.

Le calcul des probabilités pour les deux prémières itérations est :

$$\begin{cases} P(X_1 = 1|X_0 = 3) = \frac{1}{2} \\ P(X_1 = 2|X_0 = 3) = 0 \\ P(X_1 = 3|X_0 = 3) = 0, \\ P(X_1 = 4|X_0 = 3) = \frac{1}{2} \\ P(X_1 = 5|X_0 = 3) = 0 \end{cases}$$

et

$$\begin{cases} P(X_2 = 1|X_0 = 3) = \frac{1}{4} \\ P(X_2 = 2|X_0 = 3) = \frac{1}{8} \\ P(X_2 = 3|X_0 = 3) = \frac{1}{8} \\ P(X_2 = 4|X_0 = 3) = \frac{1}{8} \\ P(X_2 = 5|X_0 = 3) = \frac{3}{8} \end{cases}$$

D'une façon générale il suffit d'écrire la matrice de transition pour la toile représentée dans la Fig 2.3

$$S = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix}$$

Si on connait la distribution initiale et la matrice de transition d'une chaîne de Markov on peut alors calculer la distribution des probabilités des états.
Soit alors le vecteur initial supposé au départ $\pi(0) = (0, 0, 1, 0, 0)$ et la matrice de transition cité au-dessus.

Le vecteur de probabilité calculé pour la première itération π_1 est simplement $\pi_1 = P\pi_0^t$ est donc

$$\pi_1 = \begin{pmatrix} p(x_1 = 1) \\ p(x_1 = 2) \\ p(x_1 = 3) \\ p(x_1 = 4) \\ p(x_1 = 5) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ \frac{1}{2} \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

De la même façon, le second clic donnera lien à un vecteur de probabilité $\pi_2 = P\pi_1^t$ obtenu :

$$\pi_2 = \begin{pmatrix} p(x_2 = 1) \\ p(x_2 = 2) \\ p(x_2 = 3) \\ p(x_2 = 4) \\ p(x_2 = 5) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \frac{1}{2} \\ 0 \\ 0 \\ \frac{1}{2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{8} \\ \frac{1}{8} \\ \frac{3}{8} \end{pmatrix}$$

En observant le graphe de la Fig 2.3 on constate que tous les états communiquent entre eux donc la chaîne de Markov est irréductible.

Tous les états de la chaîne sont récurrents positifs i.e le temps moyen de premier retour pour chaque état est fini

$$\mu_i = E(T_i | X_0 = i) < +\infty.$$

De plus tous les états sont apériodiques

$$d(1) = d(2) = \cdots d(5) = PGCD\{2, 3, 4, 5, 6\} = 1.$$

Donc tous les états sont ergodiques.

D'après le théorème 1.3.6, il existe une unique distribution stationnaire

$$\pi = (\pi_1, \cdots, \pi_5).$$

Le calcul sous Maple nous a permis de calculer la distribution stationnaire de la Fig 2.3, les résultats sont obtenus ainsi :

$\alpha = 0.15$	
Méthode de la puissance	
$\pi_{(0)}$	$\pi_{(n)}$
0	0.2279
0	0.1930
1	0.1930
0	0.1930
0	0.1930

TABLE 2.1 – Vecteur de classement des pages web de la Fig 2.1.

2.3 Un modèle de navigation sur le web pour un problème d'absorption

Le graphe suivant est une légère variante de l'exemple donné à la Fig 2.3, où la page 5 n'émet pas de liens. Quand le surfeur arrivera à la page 5 il n'aura plus de choix de liens sortants.

La page 5 devient ainsi une page importante ce qui ne reflète pas vraiment la structure de la toile.

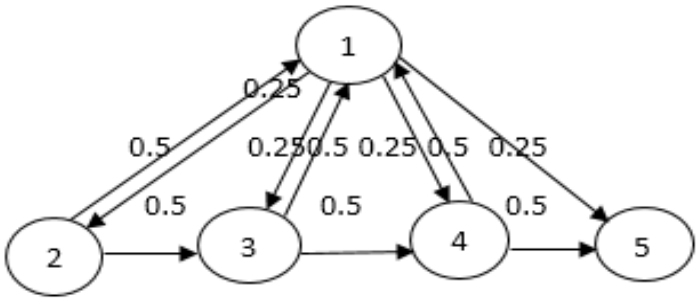

FIGURE 2.4 – Graphe d'une chaîne de Markov absorbante

La matrice de transition de cette toile est alors

$$A = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Le vecteur de classement calculé sous Maple est

$\alpha = 0.15$	
Méthode de la puissance	
$\pi_{(0)}$	$\pi_{(n)}$
0	0.2121
0	0.1779
1	0.1913
0	0.1923
0	0.2263

TABLE 2.2 – Vecteur de classement de la chaîne de Markov absorbante Fig 2.4 .

Dans ce cas la méthode de la puissance ne converge pas vers le bon vecteur propre.On constate que l'ordre le plus important est atribué à la page 5 , ce qui ne correspond pas vraiment à la structure des liens observée.

Dans ce genre de cas on introduit un vecteur ligne équiprobable pour que la chaîne de Markov soit non absorbante.

$$A = \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \end{pmatrix}$$

Le résultat obtenus pour le calcul du vecteur de classement après élimination de l'état absorbant est :

$\alpha = 0.15$	
Méthode de la puissance	
$\pi_{(0)}$	$\pi_{(n)}$
0	0.2898
0	0.1159
1	0.1739
0	0.2028
0	0.2173

TABLE 2.3 – Vecteur de classement associé à une chaîne de Markov non absorbante.

2.4 L'algorithme Page Rank avec une probabilité d'abandon

Le modèle de la marche aléatoire expliqué précédemment suppose implicitement qu'une personne qui navigue sur le net va nécessairement choisir un lien parmi ceux proposés par la page où il se trouve.

Ceci n'est évidemment pas réaliste car une personne qui navigue peut décider soit de commencer sur une nouvelle page complètement différente soit tout simplement d'arrêter sa recherche.

Pour tenir compte de cette situation on introduit alors la probabilité abandonner la navigation selon la structure des liens et de recommencer la navigation sur une page aléatoire sur la toile.

Soit S la matrice stochastique associée à une chaîne de Markov et $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ l'application définie par :

$$G(x) = \alpha Sx + (1 - \alpha)e.$$

avec :

α est la probabilité qu'un utilisateur continue à suivre la structure des liens du web.

$1 - \alpha$ est la probabilité qu'il abandonne et recommence à nouveau sur une nouvelle page choisie au hasard.

Le paramètre α est fixé par les programmeurs de Google.

$e = (\frac{1}{n}, \frac{1}{n}, \cdots, \frac{1}{n})^t$ est un vecteur qui représente la probabilité de se retrouver sur l'une des pages formant la toile.

2.5 Matrice de téléportation

On peut réécrire la formule précédente comme un problème de point fixe par rapport à x.

Soit la matrice E définie par :

$$\begin{pmatrix} \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \end{pmatrix}$$

On va montrer que pour tout vecteur stochastique v on a $Ev = e$

alors

$$Ev = \begin{pmatrix} \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \sum_{i=1}^{n} v_i \\ \vdots \\ \vdots \\ \frac{1}{n} \sum_{i=1}^{n} v_i \end{pmatrix}$$

comme v_i est stochastique $\sum_{i=1}^{n} v_i = 1$ alors

$$Ev = e = \begin{pmatrix} \frac{1}{n} \\ \vdots \\ \vdots \\ \frac{1}{n} \end{pmatrix}$$

La matrice de téléportation permet ainsi par produit vectoriel de simuler une sorte de processus de saut qui permet de sauter d'une page web vers une autre sans respecter la structure des liens.

2.6 Existence d'une distribution stationnaire pour l'algorithme Page Rank

Le théorème suivant à pour objet de décrire les propriétés de la matrice de Google.

Théorème 2.6.1. *Soit le modèle Page Rank définie par l'application suivante*

$$G(\alpha) = \alpha S + (1-\alpha)E.$$

avec S est la matrice des connexions, E est la matrice de téléportation.
Si la matrice S est stochastique alors G l'est aussi.

Preuve

- S est stochastique alors

$$\forall\ i,j \in \mathbf{N}\ s_{ij} \in [0,1]\ \ et\ \ \sum_{j=1}^{n} s_{ij} = 1\ \ \forall i$$

On a alors

$$G_{ij} = \alpha s_{ij} + (1-\alpha)\frac{1}{n}\ \ \forall i,j \in \{1,\cdots,n\}$$

$\forall i,j \in \{1,\cdots,n\}$ on a

$$\begin{aligned} 0 \leq s_{ij} &\leq 1 \\ 0 \leq \alpha s_{ij} &\leq \alpha \end{aligned}$$

$$\Longrightarrow\ 0 \leq \alpha s_{ij} + (1-\alpha)\frac{1}{n} \leq \alpha\frac{1}{n}$$

d'où $\forall\ i,j \in \mathbf{N}\ 0 \leq G_{ij} \leq 1$

- On a

$$\begin{aligned} \sum_{j=1}^{n} G_{ij} &= \sum_{j=1}^{n} (\alpha s_{ij} + (1-\alpha)\,\frac{1}{n} \\ &= \alpha \sum_{j=1}^{n} s_{ij} + \sum_{i=1}^{n} \frac{1}{n} \end{aligned}$$

Comme on a

$$\sum_{j=1}^{n} s_{ij} = 1$$

et

$$\sum_{i=1}^{n} \frac{1}{n} = 1$$

alors

$$\sum_{j=1}^{n} G_{ij} = \alpha + 1 - \alpha$$

$\Longrightarrow \sum_{j=1}^{n} G_{ij} = 1.$

2.7 Le modèle Page Rank en tant que problème de point fixe

Nous avons déjà prouvé précédemment que que la chaîne de Markov associée à la matrice de transition définie par l'application

$$G(\alpha) = \alpha S + (1 - \alpha)E.$$

admet une distribution stationnaire et qu'on a calculé pour les Fig 2.1 et Fig 2.3.

En terme d'analyse numérique, cette distribution stationnaire est un point fixe de l'application $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$,

$$G_\alpha(x) = \alpha S x + (1 - \alpha)e. \tag{2.2}$$

Où S est la matrice des connexions,α est la probabilité de suivre les liens offerts, x est un vecteur dans $\mathbb{R}^n$ et e représente le vecteur équiprobable.

$$e = \begin{pmatrix} \frac{1}{n} \\ \vdots \\ \\ \vdots \\ \frac{1}{n} \end{pmatrix}$$

Proposition 2.7.1. *Soit $S \in \mathbb{R}^{n\times n}$ une matrice stochastique et $G : \mathbb{R}^n \longrightarrow \mathbb{R}^n$, l'application définie par*

$$G_\alpha(x) = \alpha S x + (1 - \alpha)e.$$

avec α une constante $\alpha \in]0, 1[$, alors G est contractante de rapport α, par conséquent elle admet un point fixe unique qui est la distribution stationnaire de $G(\alpha)$.

Preuve

Soient deux vecteurs x *et* $y \in \mathbb{R}^n$,

On a

$$\begin{aligned} G_\alpha x - G_\alpha y &= (\alpha S x + (1-\alpha)e) - (\alpha S y - (1-\alpha)e) \\ &= \alpha S(x-y) \end{aligned}$$

Ceci permet de calculer la norme

$$\begin{aligned} \| \alpha S(x-y) \| &= | \alpha | \| S(x-y) \| \\ &= | \alpha | \sum_{i=1}^{n} | \sum_{j=1}^{n} s_{ij}(x_i - y_i) | \\ &\leq \alpha \sum_{i=1}^{n} | \sum_{j=1}^{n} s_{ij}(x_i - y_i) | \\ &\leq \alpha \sum_{i=1}^{n} | (x_i - y_i) | \sum_{j=1}^{n} s_{ij} \\ &\leq \alpha \| (x-y) \| \end{aligned}$$

D'où

$$\| Gx - Gy \| \leq \alpha \| x - y \| .$$

Chapitre 3

Etude de l'algorithme Page Rank

3.1 Estimation de la probabilité d'abandon

Le paramètre α représente la probabilité qu'un utilisateur de moteur de recherche Google suit les liens qui lui sont offert et $1-\alpha$ est la probabilité d'abondon avec $0 < \alpha < 1$.

La probabilité $1 - \alpha$ a été fixé par les deux fondateurs de Google à 0.85. Des études statistiques ont montrés qu'en moyenne un utilisateur clique sur 6 liens successif avant de quitter et recommencer à nouveau. Le calcul de cette probabilité est basique :

Soit l'évènement A : un utilisateur suit 6 liens parmi n liens qui lui sont offert.

Soit l'évènement B : un utilisateur quitte la recherche et recommence à nouveau après avoir suivi 6 liens.

Alors

$$P(B) = 1 - P(A).$$

Or que

$$P(A) = \frac{n(n-1)(n-2)(n-3)(n-4)(n-5)}{n!} = A_n^6.$$

On a alors

$$P(B) = 1 - A_n^6.$$

Pour un ensemble de pages estimé à 40 pages web on aura

$$P(B) = 0.99.$$

3.2 Sensibilité de l'algorithme aux conditions initiales.

Dans l'algorithme Page Rank, le paramètre $1-\alpha$ est placé habituellement à 0.85 mais il peut théoriquement varier entre 0 et 1 et la matrice des connexions est toujours choisi de tel sorte qu'elle soit stochastique.

D'après le modèle de Page Rank

$$G(\alpha) = \alpha S + (1-\alpha)E.$$

G dépend de α et S.

L'algorithme Page Rank met à jours régulièrement ces deux paramètres dans sa formule.

Nous allons explorer la sensibilité aux variations de chaque paramètre.

Il faut noter qu'on s'attend à ce qu'un classement reste stable. En effet il serait anormal si l'algorithme proposait des réponses complètement différentes pour de petits changements.

3.2.1 Sensibilité par rapport à La probabilité d'abandon $1-\alpha$

Nous allons commencer par l'étude de la sensibilité de l'algorithme par rapport à la probabilité d'abandon.

Soit l'application de l'algorithme Page Rank suivante :

$$G(\alpha) = \alpha S + (1-\alpha)E.$$

Soient $\pi(\alpha) = (\pi_1(\alpha), \cdots, \pi_n(\alpha))$ le vecteur de classement (où le vecteur de distribution stationnaire) associé à G, S la matrice des connexions associé et π_1 et π_2 deux vecteurs de classements associés à deux valeurs α_1 et α_2 respectivement.

$\delta\pi$ représente la variation de π tel que $\delta\pi = \pi_1 - \pi_2$ et $\delta\alpha$ la variation de α tel que $\delta\alpha = \alpha_1 - \alpha_2$.

Alors on a

$$\begin{aligned} \pi_1 - \pi_2 &= (\alpha_1 S + (1-\alpha_1)E)\pi_1 - (\alpha_2 S + (1-\alpha_2)E) \\ &= \alpha_1 S + (\pi_1 - \pi_2) + S\pi_2(\alpha_1 - \alpha_2) - (\alpha_1 - \alpha_2)E \end{aligned}$$

Ce qui donne

$$\delta\pi = \alpha_1 S\delta\pi + S\pi_2\delta\alpha - \delta\alpha E.$$

Par l'inégalité triangulaire on obtient

$$\| \delta\pi \| \leq \alpha_1 \| S \| \| \delta\pi \| + \| S \| \| \pi_2 \| \| \delta\alpha \| + \| \delta\alpha \|$$

On a π est stochastique alors $\sum_{i=1}^{n} \pi_i = 1 \Longrightarrow \| \pi_1 \| \leq 1$ et comme S est stochastique on a la $\sum_{j=1}^{n} s_{ij} = 1$.
D'où

$$\| \delta\pi \| \leq \frac{2 \| \delta\alpha \|}{1-\alpha}.$$

Donc si on prend α trop proche de 1 le problème est sensible aux conditions initiales.

On illustre cette section par quelques exemples de pratique qui justifie le choix de α.

On donne une matrice des connexions qui correspond à huit pages du web.

$$S = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{7} & 0 & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ 0 & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & \frac{1}{5} \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Le tableau suivant montre les valeurs propres et les vecteurs propres associés à la matrice des connexions S fixe et celles de la matrice de Google ainsi que le classement des pages selon leurs importance pour une valeur $1-\alpha$ fixé à 0.85 tout en changeant la distribution initiale.

$\pi_1(0)$	$\pi(n)$	rang	$\pi_2(0)$	$\pi(n)$	rang
$\frac{1}{8}$	0.1379	1	$\frac{1}{7}$	0.1250	**1**
$\frac{1}{8}$	0.1233	3	$\frac{1}{7}$	0.1250	**1**
$\frac{1}{8}$	0.1323	4	$\frac{1}{7}$	0.1250	**1**
$\frac{1}{8}$	0.1158	8	$\frac{1}{7}$	0.1249	**2**
$\frac{1}{8}$	0.1201	5	$\frac{1}{7}$	0.1249	**2**
$\frac{1}{8}$	0.1172	7	$\frac{1}{7}$	0.1249	**2**
$\frac{1}{8}$	0.1179	6	$\frac{1}{7}$	0.1249	**2**
$\frac{1}{8}$	0.1351	2	0	0.1250	**1**

TABLE 3.1 – Vecteurs de classement des huit pages web pour $1-\alpha$ fixé à 0.99 et différentes distributions initiales

Voici un autre exemple pour le calcul des vecteurs de classement pour la même matrice de l'exemple précédent et le classement des pages selon leurs importance pour une même distribution initiale tout en variant la probabilité d'abandon $1-\alpha$.

$1-\alpha=0.4$		$1-\alpha=0.6$		$1-\alpha=0.8$		$1-\alpha=0.99$	
$\pi(n)$	rang	$\pi(n)$	rang	$\pi(n)$	rang	$\pi(n)$	rang
0.1912	1	0.1644	1	0.1427	1	0.1250	**1**
0.1091	4	0.1173	4	0.1225	4	0.1250	**1**
0.1570	3	0.1453	3	0.1348	3	0.1250	**1**
0.0871	8	0.1002	8	0.1127	8	0.1249	**2**
0.1002	5	0.1102	5	0.1183	5	0.1249	**2**
0.0912	**6**	0.1033	**7**	0.1145	7	0.1249	**2**
0.0894	**7**	0.1037	**6**	0.1153	6	0.1249	**2**
0.1745	2	0.1552	2	0.1388	2	0.1250	**1**

TABLE 3.2 – Vecteurs de classement des huit pages web pour différentes valeurs de α

3.2.2 Sensibilité par rapport à la matrice des connexions

La matrice des connexions représente le rapport des liens entre les pages web. Dans cette section on s'intéresse à la sensibilité de l'algorithme par rapports aux changements de cette matrice.

Une preuve sera établie pour faire l'objet.

Soit le modèle de Page Rank donné par la formule suivante :

$$G(\alpha) = \alpha S + (1-\alpha)E.$$

Soient S_0, S_1 deux matrices de connexions, π_0, π_1 deux vecteurs de classements associés à G_0 et G_1 respectivement.

Soit $\delta\pi = \pi_1 - \pi_0$ une variation de π.

On a alors

$$\begin{aligned} \pi_1 - \pi_0 &= (\alpha S_1 - (1-\alpha)E)\pi_1 - (\alpha S_0 - (1-\alpha)E)\pi_0 \\ &= \alpha S_1\pi_1 - (1-\alpha)E\pi_1 - \alpha S_0\pi_0 - (1-\alpha)E\pi_0 \end{aligned}$$

et on aussi

$$(1-\alpha)E\pi_1 - (1-\alpha)E\pi_0 = 0.$$

car π_1 et π_0 sont des vecteurs de probabilité et

$$\sum_{j=1}^{n} \pi_i = 1.$$

alors

$$\begin{pmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_n \end{pmatrix} = \begin{pmatrix} \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix}$$

$$\begin{aligned} \pi_1\pi_0 &= \alpha S_1\pi_1 - \alpha S_0\pi_0 \\ &= \alpha S_1\pi_1 + \alpha S_0\pi_1 - \alpha S_0\pi_1 - \alpha S_0\pi_0 \\ &= \alpha(S_1 - S_0)\pi_1 + \alpha S_0(\pi_1 - \pi_0) \end{aligned}$$

$$\begin{aligned} \Longrightarrow (\pi_1 - \pi_0) - \alpha S_0(\pi_1 - \pi_0) &= \alpha(S_1 - S_0)\pi_1 \\ \Longrightarrow (Id - \alpha S_0)(\pi_1 - \pi_0) &= \alpha(S_1 - S_0)\pi_1 \\ \Longrightarrow \pi_1 - \pi_0 &= \alpha(Id - \alpha S_0)^{-1}(S_1 - S_0)\pi_1((Id - \alpha S_0)\ est\ inversble) \\ \Longrightarrow \delta\pi &= \alpha(Id - \alpha S_0)^{-1}\delta S\pi_1 \end{aligned}$$

on aura

$$\| \delta\pi \| \leq \alpha \| Id - \alpha S_0 \|^{-1} \| \delta S \| \| \pi_1 \| .$$

on aussi

$*$ π est stochastique donc $\sum_{j=1}^{n} \pi_i = 1 \Longrightarrow \| \pi_1 \| \leq 1$

$*$ $\| Id - \alpha S_0 \| = 1 - \alpha$ car

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} - \alpha \begin{pmatrix} s_{11} & \cdots & \cdots & s_{1n} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ s_{n1} & \cdots & \cdots & s_{nn} \end{pmatrix}$$

comme $\sum_{j=1}^{n} s_{ij} = 1$ alors

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \cdots & 0 \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} - \begin{pmatrix} \alpha \\ \vdots \\ \vdots \\ \alpha \end{pmatrix} = \begin{pmatrix} 1-\alpha \\ \vdots \\ \vdots \\ 1-\alpha \end{pmatrix}$$

d'où le résultat.

donc

$$\| \delta\pi \| \leq \frac{\alpha}{1-\alpha} \| \delta S \| .$$

Soit deux matrices de connexions S_1 et S_2 où S_2 est une légère modification de la matrice S_1 où on a ajouté un lien de la page 8 vers la page 7.

$$S_1 = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{7} & 0 & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ 0 & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & \frac{1}{5} \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad S_2 = \begin{pmatrix} 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{7} & 0 & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} & \frac{1}{7} \\ 0 & \frac{1}{5} & 0 & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & \frac{1}{5} \\ 0 & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & 0 & 0 & \frac{1}{6} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Voici un exemple qui montre la sensibilité du vecteur de classement par rapport aux changements de la matrice des connexions S_1 et S_2 lorsque $(1-\alpha) \longrightarrow 1$.

S_1		S_2	
$\pi(n)$	rang	$\pi(n)$	rang
0.1250	1	0.1220	5
0.1250	1	0.1280	1
0.1250	1	0.1280	1
0.1249	2	0.1249	3
0.1249	2	0.1068	6
0.1249	2	0.1041	7
0.1249	2	0.1173	5
0.1250	2	0.1248	4

TABLE 3.3 – Vecteurs de classements associés aux deux matrice S_1 et S_2 pour une valeur $1-\alpha = 0.99$.

3.3 Vitesse de convergence de l'algorithme

La vitesse de convergence de la méthode de la puissance dépend de la valeur propre sous dominante de G, $\lambda_2(G) = \alpha$, plus $\alpha^k \longrightarrow 0$ plus la vitesse de convergence de la méthode de la puissance est rapide.
Le vecteur de classement est calculé au bout de 50 à 100 itérations [2].

α	Nombre d'ittération
0.5	34
0.75	81
0.8	104
0.85	142
0.9	219
0.95	449
0.99	2292
0.99	23015

TABLE 3.4 – Effet du paramètre α et le nombre d'itération nécessaire à une précision de 10^{-10} voir [2]

3.4 Propriété spectrale des valeurs propres du modèle Page Rank

Nous allons à présent étudier la propriété spectrale du modèle Page Rank renforcer par une preuve.

Proposition 3.4.1. *Soit le modèle Page Rank donné par la formule :*

$$G(\alpha) = \alpha S + (1 - \alpha)E.$$

Si le spectre de la matrice stochastique S est $\{1, \lambda_2, \cdots, \lambda_n\}$ alors le spectre de G est $\{1, \alpha\lambda_2, \cdots, \alpha\lambda_n\}$.

Preuve

Soit S une matrice stochastique de spectre $Sp = \{1, \lambda_2, \cdots, \lambda_n\}$.

Posons $G = \alpha S + (1 - \alpha)E$.

S est stochastique donc $S.v = v$ où le vecteur colonne v est un vecteur propre de S pour la valeur propre 1.

Posons

$$A = (v, T).$$

où v est le vecteur propre associé à la matrice S et T une matrice $n \times (n-1)$ choisi de sorte que A soit inversible.

Notons

$$A^{-1} = \begin{pmatrix} v' \\ T' \end{pmatrix}$$

où v' est un vecteur ligne et T' une matrice $(n-1) \times n$. Effectuons le produit par bloc

$$I_n = A^{-1}A = \begin{pmatrix} v' \\ T' \end{pmatrix} (v, T) = \begin{pmatrix} v'v & v'T \\ T'v & T'T \end{pmatrix}$$

on déduit

$$v'v = 1, \quad v'T = \begin{pmatrix} 0 & \cdots & 0 \end{pmatrix}, \quad T'v = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}, \quad T'T = I_{n-1}$$

Considérons

$$A^{-1}SA = \begin{pmatrix} v' \\ T' \end{pmatrix} S \begin{pmatrix} v & T \end{pmatrix} = \begin{pmatrix} v' \\ T' \end{pmatrix} \begin{pmatrix} Sv & ST \end{pmatrix} = \begin{pmatrix} v' \\ T' \end{pmatrix} \begin{pmatrix} v & ST \end{pmatrix} = \begin{pmatrix} v'v & v'ST \\ T'v & T'ST \end{pmatrix}$$

par identification

$$A^{-1}SA = \begin{pmatrix} 1 & * \\ 0 & T'ST \end{pmatrix}$$

La matrice $A^{-1}SA$ étant semblable à S et a exactement les mêmes valeurs propres $1, \lambda_2, \cdots, \lambda_n$.

Calculons maintenant

$$A^{-1}GA = \begin{pmatrix} v' \\ T' \end{pmatrix} G \begin{pmatrix} v & T \end{pmatrix} = \alpha \begin{pmatrix} 1 & * \\ 0 & T'ST \end{pmatrix} + (1-\alpha) \begin{pmatrix} v'Ev & v'ET \\ T'Ev & T'ET \end{pmatrix}$$

On obtient

$$A^{-1}GA = \begin{pmatrix} 1 & * \\ 0 & \alpha(T'ST) \end{pmatrix}$$

D'où les valeurs propres de G sont $1, \alpha\lambda_2, \cdots, \alpha\lambda_n$.

Conclusion

Pour que l'algorithme Page Rank assure la convergence vers le bon vecteur de classement, le paramètre α doit être choisi de sorte à respecter les trois contraintes suivantes :

- On a
$$\| \delta\pi \| \leq \frac{2 \| \delta\alpha \|}{1-\alpha}$$
Alors plus $\alpha \longrightarrow 0$ plus le modèle est numériquement stable .
- Le paramètre α représente la probabilité qu'un navigateur suit les liens qui lui sont offert , donc α doit être autour de cette probabilité qui est estimé à .
- La vitesse de convergence de méthode de la puissance dépend de α, plus $\alpha^k \longrightarrow 0$ plus la vitesse de convergence est rapide, donc α doit être choisi proche de 0.

Chapitre 4

Optimisation des opérations de calcul du vecteur propre

En 2008 l'entreprise Google affirmait indexer 40 milliards de pages. Cette taille imposante oblige les concepteurs à utiliser un ensemble de techniques et d'astuces visant à réduire le coût en calcul au maximum.

Nous proposons dans cette section certaines techniques utilisées pour minimiser le coût de calcul du vecteur propre.

4.1 Optimisation de l'espace de stockage

L'algorithme Page Rank se base sur le calcul du vecteur propre de la matrice G qui est dense et de très grande taille. Le stockage d'une matrice de cette taille est extrêmement coûteux même avec des installations de pointe. L'espace de stockage nécessaire pour la matrice de Google d'ordre n^2 est $2.56.10^{42}$ éléments .

La matrice S elle par contre est creuse et permet d'obtenir la matrice G en vertu de la formule :

$$G(\alpha) = \alpha S + (1 - \alpha)E.$$

Une étude statistique faite par la compagnie Google sur les pages web montre qu'en moyenne une page web émet 10 liens, ce qui signifie que la matrice S contient $10n$ informations pertinentes ainsi l'espace de stockage est 4.10^{11} éléments.

Le gain réalisé en utilisant cette technique de stockage est important et permet d'alléger les appels de l'algorithme.

4.2 Optimisation du nombre d'opérations nécessaires pour le produit matriciel

4.2.1 Méthode naïve

Si on applique la méthode de la puissance on est censé effectuer le produit de la matrice G par un vecteur à chaque itération.

Pour une matrice de taille n le coût de cette opération est de n opérations de multiplications et $n-1$ opérations d'additions. ainsi le nombre d'opération est de $2n^2 - n$.

4.2.2 Formule optimisée

Il est possible de tirer profit de la formule de l'algorithme pour réduire le coût des opérations en calcul :

$$G(\alpha)v = \alpha Sv + (1-\alpha)Ev$$

Alors Si on applique la puissance pour la formule optimisée on est censé effectuer le produit de la matrice S par un vecteur à chaque itération plus l'addition du vecteur équiprobable e .

Comme la matrice S contient en moyenne 10 éléments non nuls par ligne. Le produit de la matrice S par un vecteur de taille n nécessite 10 opérations de multiplication et 9 opérations d'additions. Ainsi le nombre d'opération calculé pour une seule itération est de $(20n)$ opérations .

On peut contourner le produit de la matrice de téléportation en faisant la remarque suivante :

$$Ev = e \quad \forall\ v\ \textit{stochastique.}$$

On a une matrice de téléportation E et un vecteur stochastique v ,alors

$$Ev = \begin{pmatrix} \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \\ \vdots & & & \vdots \\ \vdots & & & \vdots \\ \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} v_i \\ \vdots \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} v_i \end{pmatrix}$$

comme $\sum_{i=1}^{n} v_i = 1$ alors

$$Ev = e = \begin{pmatrix} \frac{1}{n} \\ \vdots \\ \vdots \\ \frac{1}{n} \end{pmatrix}$$

4.2.3 Comparaison entre la méthode naïve et la formule optimisée

Nous allons à présent comparer le coût en calcul des deux méthodes précédentes. Nous avons utilisé comme repère un processeur ayant une vitesse d'un pétaflops [1] C'est-à-dire effectuant 10^{15} opérations par secondes.En utilisant la matrice G le nombre d'opérations est évaluée à environ de $3.2.10^{21}$ opérations par itération, ainsi il faudrait environ 37 jours. Mais en utilisant la matrice des connexions S , le nombre d'opérations nécessaires est évalué à environ 8.10^{11} opérations, avec une même machine il faudrait $(8.10^{-4}s)$ pour effectuer une itération.

4.3 Optimisation des opérations de calcul par le test d'arrêt

Le critère d'arrêt usuel utilisé par plusieurs méthodes numériques est :

$$\| x^{k+1} - x^k \| < \epsilon.$$

Ce test nécessite le calcul du vecteur en entier même si plusieurs composantes ne peuvent plus être améliorées.

Pour tenir compte de cette situation l'algorithme Page Rank utilise un critère différent qui permet de bloquer le calcul d'une composante dès qu'elle atteint une certaine précision.

$$| x_i^{k+1} - x_i^k | < \epsilon.$$

c'est-à-dire lors de calcul du vecteur de classement elle contrôle les valeurs calculée à chaque fois une valeur atteinte vérifie ce dérnier critère la valeur sera fixée et l'algorithme continue le calcul pour le restes des valeurs jusqu'à l'obtention du vecteur propre.

1. Début juin 2008, le super calculateur IBM Roadrunner est le premier à franchir la barre symbolique du petaflops. Puis, en novembre 2008, c'est au tour du supercalculateur Jaguar de Cray. En avril 2009, c'étaient les deux seuls supercalculateurs à avoir dépassés le petaflops.(source Page processeur Wikipédia 18/12/2010)

Kamvar et Haveliwala [7] ont prouvé que ce critère d'arrêt permet de réduire 17% du temps de calcul du vecteur Page Rank.

Voici un exemple qui calcule le vecteur de classement pour la matrice citée dans le chapitre précédent tout en précisant le fonctionnement du critère d'arrêt.

	Itération 1		Itération 2		Itération 3		Itération 4	
x_i	x^k	x^{k+1}	x^k	x^{k+1}	x^k	x^{k+1}	x^k	x^{k+1}
x_1	0.1250	0.1611	0.1611	0.1689	0.1689	0.1705	0.1705	0.1706
x_2	0.1250	0.1221	0.1221	0.1161	0.1161	0.1157	0.1157	0.1155
x_3	0.1250	0.1470	0.1470	0.1470	0.1470	0.1479	0.1479	0.14809
x_4	0.1250	0.0974	0.0974	0.0977	**0.0977**	**0.0977**	**0.0977**	**0.0977**
x_5	0.1250	0.1114	0.1114	0.1086	0.1086	0.1080	0.1080	0.1079
x_6	0.1250	0.1020	0.1020	0.1008	0.1008	0.1004	**0.1004**	**0.1004**
x_7	0.1250	0.1049	0.1049	0.1016	0.1016	0.1006	0.1006	0.1004
x_8	0.1250	0.1536	0.1536	0.1590	0.1590	0.1595	0.1595	0.1597

TABLE 4.1 – Calcul du vecteur de classement par la méthode de la puissance.

Si on observe le tableau on constate que la valeur x_4 n'évolue pas à partir de la troisième itération, et la valeur x_6 à partir de la quatrième itération, donc il n'est pas nécessaire de continuer le calcul qui représente ainsi des opérations de calcul en plus et qui sont inutiles.

Conclusion

Dans ce travail, nous avons étudié l'algorithme Page Rank de la compagnie Google qui a fait son succès.

La navigation sur le web est modélisée par une chaîne de Markov qui a été forcée pour qu'elle soit irréductible et apériodique ce qui permet de calculer une distribution stationnaire qui représente le vecteur de classement des pages.

L'algorithme dépend d'une probabilité qui représente la probabilité de suivre ou pas les liens proposés. Cette probabilité est utilisée comme paramètre par Google afin de rendre l'algorithme stable numériquement.

La taille imposante du web impose des choix judicieux en matière d'optimisation des opérations de calcul. Sans quoi l'algorithme serait lent et perdrait son efficacité.

L'algorithme Page Rank fait la chasse aux opérations superflues et élimine tout ce qui peut être éliminé comme calculs inutiles.

Nous avons proposés trois techniques parmi celle utilisés par Google pour optimiser les calculs.

Bibliographie

[**1**] Alan Ruegg, Processus stochastique avec applications aux phénomènes d'attente et de fiabilité, Edition Press Polythecnique Romande 1989.

[**2**] Amy N. Langville and Carl D. Meyer, Google's Page Rank and Beyond : The Science of Search Engine Rankings,Edition Springer 2006.

[**3**] Amy N. Langville and Carl D. Deeper Inside PageRank 2004.

[**4**] Christiane Rousseau Yvan Saint Aubin,Mathématiques et Technologie, Edition Springer 2008.

[**5**] Michael Eisermann, Comment fonctionne google,, Instiut Fourier, Université de Grenoble I, France www.fourier.ujf-grenoble.fr/ eiserm.

[**6**] Michel Pierre et Antoine Henrot, Analyse numérique, ,Ecole des Mines de Nancy,www-fourier.ujf-grenoble.fr/ parisse 2014.

[**7**] Pascal Azerad ,Mathématiques du Web, Université de Montpellier, www.mon.univ-montp2.fr, 28 mars 2014.

[**8**] Yen-Yu Chen, Qingqing Gan, and Torsten Suel. "I/O-Efficient Techniques for Computing PageRank." In Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02), pp. 549— 557. New York : ACM Press, 2002.

Annexe

Programme utilisé permettant de calculer le vecteur des classements

restart ;

with(LinearAlgebra) :

Choisir "n" le nombre de pages web. La variable "link" est une liste de paires $[i, j]$, où "$[i, j]$" signifie la page "i" pointe un lien à la page "j".

$n := 8$:

links :=

$A := Matrix(n, n)$:

L := nops(links) :

for i from 1 to n do

for k from 1 to L do

if links $[k, 1] = i$ then $A[links[k, 2], i] := 1$; fi :

od :

od :

for i from 1 to n do

$s := add(A[j, i], j = 1..n)$:

if$s <> 0$ then for j from 1 to n do $A[j, i] := evalf(A[j, i]/s)$; od : fi ;

if $s = 0$ then for j from 1 to n do $A[j, i] := 1.0/n$; od : fi ;

od :
A ;
Eigenvalues(A) ;
Remarque : I est la base des complexes.
Vecteur propre de A la matrice des connexions
$r := 0$:
$eigvA := array(1..n)$:
$Q := Eigenvectors(A)$:
$eigs := Q[1]$:
for j from 1 to n do
if evalf$(abs(eigs[j] - 1.0)) < 1.0e - 6 then r := r + 1; eigvA[r] := Column(Q[2], j)$: fi ;
od :
printf $V_1(A)$ est de dimension
for j from 1 to r do
$s := add(eigvA[j][k], k = 1..n)$:
$eigvA[j] := map(q- > Re(q), eigvA[j]/s)$;
$print(eigvA[j])$;
od :
$V_1(A)$ est de dimension 1
m est la probabilité d'abondan
$m := 0.85$:
$M := evalf((1 - m)_* A + m_* Matrix([seq([seq(1/n, j = 1..n)], k = 1..n)]))$;
Calcul du spectre de la matrice Google.
Eigenvalues(M) ;
Calcul du vecteur propre de la matrice de Google par la commande de Maple.
$Q2 := Eigenvectors(M)$:
$eigs := Q2[1]$:
for j from 1 to n do
if evalf$(abs(eigs[j] - 1.0)) < 1.0e - 6 then eigvM := Column(Q2[2], j)$: fi ;
od :
$s := add(eigvM[k], k = 1..n)$:
$eigvM := evalf(map(qq- > Re(qq), eigvM/s), 16)$;
Méthode de la puissance

Calcul du vecteur propre de la matrice de Google par la méthode de la puissance avec le vecteur initial qui contient 1/n comme composante.

$s := Vector(n)$:
for j from 1 to n do
$s[j] := 1./n$
od :
$x := s$;
for k from 1 to 1000 do
$xnew := (1 - m) * A.x + m * s$:
$xnew := xnew/Norm(xnew, 1)$:
if $Norm(x - xnew, 1) < 1.0e - 3$ then break ; fi ;
for i from 1 to n do
print $(x[i], xnew[i], k)$;
od :
$x := xnew$;
od :
k;
x;
eigvM ;
$x := s$;
for k from 1 to 1000 do
$xnew := (1 - m) * A.x + m * s$:
$xnew := xnew/Norm(xnew, 1)$:
for j from 1 to n do
if $abs(x[j] - xnew[j]) > 0.001$ then
print $(x[j], xnew[j])$;
$x[j] := xnew[j]$;
fi ;
od :
od ;
k ;
x ;
eigvM ;

Table des figures

Liste des tableaux