LYCÉE CANTONAL

L'algorithme PageRank de Google

Paul Jolissaint

1 Introduction

Un moteur de recherche tel que Google doit faire trois choses :

1. Naviguer sur le web et localiser (si possible) toutes les pages ayant un accès public.
2. Indexer les données ci-dessus pour qu'elles puissent être retrouvées efficacement par mots-clés ou groupes de mots significatifs.
3. Classer l'importance de chaque page dans la base de données de sorte que lorsqu'un internaute fait une recherche et que le sous-ensemble des pages correspondantes dans la base de données a été trouvé, **les pages les plus importantes soient présentées en premier.**

Le but de l'algorithme PageRank est de résoudre le troisième problème ci-dessus, et nous allons expliquer comment on procède dans Google pour élaborer le classement des pages web.

Le texte ci-dessous doit beaucoup à l'article [1] ; pour une étude sensiblement plus approfondie de l'algorithme PageRank et une introduction aux moteurs de recherche sur l'internet, l'ouvrage [2] est très complet. Il contient même un chapitre consacré à l'algèbre linéaire et aux chaînes de Markov. Enfin, la référence [3] propose un algorithme de classification analogue à PageRank.

2 Modélisation

L'ensemble des pages web peut être représenté par un **graphe orienté** : chaque page est représentée par un point (appelé **sommet** du graphe) et on dessine une flèche d'un sommet A vers un sommet B si et seulement si la page A contient un lien hypertexte *vers* la page B. Une telle flèche s'appelle une **arête** (orientée).

Notons que le nombre de pages dans le web est estimé à plus de 60 milliards, et Google en répertorie environ 35 milliards. Ainsi, si la partie du web indexée par Google contient n pages, on convient de représenter chaque page par un numéro i compris entre 1 et n et on désignera par x_i le **score** de la page i. Il s'agit d'un nombre compris entre 0 et 1 qui doit indiquer l'importance de la page i au sens du point 3 ci-dessus : plus x_i est élevé, plus l'URL de la page i apparaît tôt dans la liste que propose le moteur de recherche lorsque la page i contient les mots-clés demandés. Comme première condition, on doit avoir $x_i > x_j$ si la page i est plus importante que la page j. Une approche simple consisterait à définir x_i de la façon suivante : notons m_i le nombre de liens vers la page i, alors x_i pourrait être défini par le rapport m_i/n.

Cette approche ignore toutefois un point crucial : *une page est importante si d'autres pages importantes pointent vers elle*. Ainsi, si la page j est importante (c'est-à-dire si x_j est grand) et s'il existe un lien de la page j vers la page i, cela donne d'autant plus d'importance à la page i, donc l'établissement d'un tel lien devrait augmenter x_i davantage que si on établit un lien

d'une page sans importance vers la page i. Toutefois, on veut éviter que des pages sans grande importance obtiennent un grand score par des moyens artificiels comme par exemple la création d'un grand nombre de pages fictives contenant chacune un lien vers la page en question. Ainsi, si la page j contient n_j liens vers d'autres pages, dont la page i, on veut alors que la contribution de la page j au score x_i de la page i soit égal à x_j/n_j et non pas égal à x_j.

Les concepteurs de Google, Larry Page et Sergey Brin, alors étudiants en informatique à Stanford, ont adopté en 1998 la définition [1] suivante de la suite des scores $x_1, x_2, \ldots, x_n$: rappelons que n_j désigne le nombre de liens contenus dans la page j pour tout j (donc issus de la page j vers d'autres pages, et on convient qu'un lien d'une page vers elle-même est ignoré), et soit $L_i \subset \{1,\ 2,\ \ldots,\ n\}$ l'ensemble des pages qui ont un lien vers la page i. On demande alors que pour tout i,

$$x_i = \sum_{j \in L_i} \frac{x_j}{n_j}$$

et que les composantes de $\vec{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ soient normalisées de sorte que $\sum_i x_i = 1$.

Le problème avec la définition ci-dessus est que le calcul de x_i fait intervenir les autres x_j qui sont a priori eux aussi inconnus. En fait, on peut calculer des valeurs approximatives des x_i en utilisant une méthode itérative : notons $x_i^{(k)}$ la valeur de x_i après la k-ième itération. On initialise le processus en posant $x_i^{(0)} = 1/n$ pour tout i (ce qui s'interprète en stipulant qu'au départ toutes les pages possèdent le même score), puis en définissant

$$x_i^{(k+1)} = \sum_{j \in L_i} \frac{x_j^{(k)}}{n_j}$$

pour tout $1 \leq i \leq n$ et pour tout $k \geq 0$. On simplifie les notations en introduisant la matrice A qui est définie comme suit : pour $1 \leq j \leq n$, on remplit la colonne j en posant $\frac{1}{n_j}$ dans la ligne i si $j \in L_i$ (c'est-à-dire s'il y a un lien de j vers i) et 0 sinon, et on note comme ci-dessus $\vec{x}^{(k)}$ le vecteur dont les composantes sont les $x_i^{(k)}$.

$$A = \begin{pmatrix} \ddots & \vdots & \ddots \\ \cdots & \frac{1}{n_j} & \cdots \\ \ddots & \vdots & \ddots \\ \cdots & \frac{1}{n_j} & \cdots \\ \cdots & 0 & \cdots \end{pmatrix}.$$

De la sorte, on a

$$A\vec{x} = \vec{x} \quad \text{et} \quad A\vec{x}^{(k)} = \vec{x}^{(k+1)}\ \forall k \geq 0.$$

Il semble que Google recalcule les scores une fois par semaine, en itérant un processus proche de celui décrit ci-dessus car ce dernier pose quelques problèmes que nous allons décrire plus loin.

En attendant, voici un exemple simple qui permet de se faire une idée de la situation :

1. Il semble que L. Page et S. Brin aient été (judicieusement) conseillés par le père de Brin qui est mathématicien, spécialiste de ce genre de définition. En effet, devenu riche, ce dernier a donné 2 millions de dollars à son université afin d'alimenter un fond pour créer un poste de professeur de mathématiques supplémentaire.

Exemple 1. Considérons le web donné par le graphe suivant :

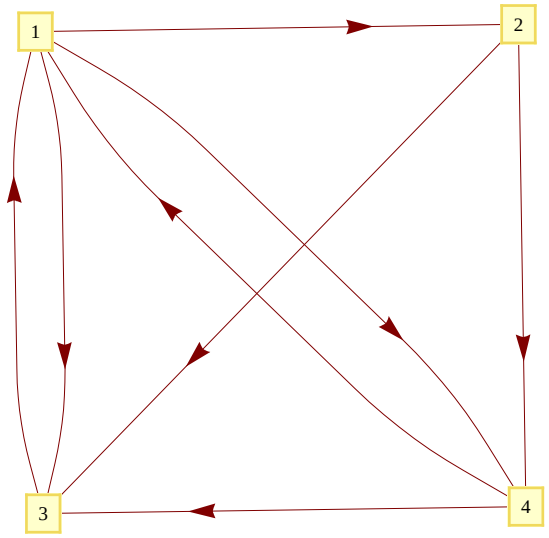


Fig. 1

Voici les valeurs correspondantes (rappelons que L_j représente l'ensemble des pages ayant un lien *vers* j, et que n_j est le nombre de flèches *issues de* j) :

j	n_j	L_j
1	3	{3,4}
2	2	{1}
3	1	{1,2,4}
4	2	{1,2}

Cela conduit au système linéaire :

$$\begin{cases} x_1 = x_3/1 + x_4/2 \\ x_2 = x_1/3 \\ x_3 = x_1/3 + x_2/2 + x_4/2 \\ x_4 = x_1/3 + x_2/2. \end{cases}$$

On pose comme ci-dessus $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$ et la matrice du système est

$$A = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

de sorte que $A\vec{x} = \vec{x}$.

En résumé, on cherche un vecteur $\vec{x}$ comme ci-dessus dont les composantes x_j sont positives, de somme 1, et qui satisfait l'équation $A\vec{x} = \vec{x}$. Ce problème est un problème de valeur et vecteur

propres : étant donné une matrice carrée A, trouver tous les nombres λ et tous les vecteurs $\vec{x} \neq \vec{0}$ tels que $A\vec{x} = \lambda\vec{x}$. Ceci fera l'objet du paragraphe suivant.

Pour terminer, mentionnons que l'équation de l'exemple 1 admet une solution : c'est un multiple du vecteur $\vec{v} = \begin{pmatrix} 12 \\ 4 \\ 9 \\ 6 \end{pmatrix}$, donc $\vec{x} = \begin{pmatrix} \frac{12}{31} \\ \frac{4}{31} \\ \frac{9}{31} \\ \frac{6}{31} \end{pmatrix} \approx \begin{pmatrix} 0.387 \\ 0.129 \\ 0.290 \\ 0.194 \end{pmatrix}$.

Remarque importante. Dans la description ci-dessus, on n'a pas tenu compte des "pages cul-de-sac" (en anglais : "dangling nodes") : ce sont les pages qui n'ont aucun lien vers une autre page, donc ce sont les pages j pour lesquelles $n_j = 0$. Brin et Page ont résolu le problème en remplaçant chaque colonne correspondant à une page cul-de-sac par la colonne $\begin{pmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{pmatrix}$.
On interprète cette modification ainsi : lorsqu'un internaute visite une page cul-de-sac, la probabilité de visiter les autres pages du web en quittant celle-ci est uniforme.

Exemple 2. Soit le graphe :

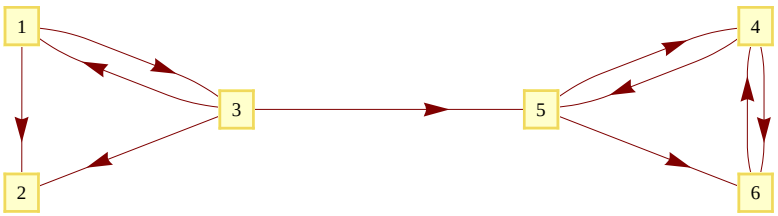


Fig. 2

Les données sont

j	n_j	L_j
1	2	{3}
2	0	{1,3}
3	3	{1}
4	2	{5,6}
5	2	{3,4}
6	1	{4,5}

la matrice de départ est

$$\begin{pmatrix} 0 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 1/3 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1 \\ 0 & 0 & 1/3 & 1/2 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}$$

et la matrice A modifiée et qui sera utilisée est

$$A = \begin{pmatrix} 0 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 1/3 & 0 & 0 & 0 \\ 1/2 & 1/6 & 0 & 0 & 0 & 0 \\ 0 & 1/6 & 0 & 0 & 1/2 & 1 \\ 0 & 1/6 & 1/3 & 1/2 & 0 & 0 \\ 0 & 1/6 & 0 & 1/2 & 1/2 & 0 \end{pmatrix}.$$

Pour terminer le paragraphe, discutons quelques propriétés de la matrice A. A priori, le calcul de $A\vec{x}$ requiert un nombre de calculs grosso modo proportionnel à n^2 puisqu'il s'agit d'un produit matriciel avec une matrice d'ordre n. Or, heureusement, la matrice A est une matrice *creuse*, c'est-à-dire qu'elle possède un grand nombre de coefficients nuls puisque la plupart des pages ne contiennent que quelques dizaines de liens vers d'autres pages, au maximum. Donc les $|L_i|$ et les n_j sont bornés par une constante C relativement petite. On verra l'importance de cette propriété dans le paragraphe 5 en particulier.

Les équations ci-dessus posent un certain nombre de questions :
- l'équation $A\vec{x} = \vec{x}$ avec $\sum_i x_i = 1$ admet-elle une solution avec $x_i > 0$ pour tout i ?
- si une solution à l'équation ci-dessus existe, est-elle unique ?
- dans le calcul de la suite $(\vec{x}^{(k)})$ qui est censée approcher $\vec{x}$, le processus converge-t-il ? et sous quelles conditions ?

Afin de garantir l'existence et l'unicité de $\vec{x}$, et la convergence du processus, Brin et Page ont modifié l'équation ci-dessus : le vecteur de score $\vec{x}$ est la solution de l'équation

$$\boxed{\vec{x} = ((1-m)A + mS)\,\vec{x}}$$

où $0 < m < 1$ est un paramètre et S est la matrice $n \times n$ dont tous les coefficients sont égaux à $1/n$. On verra que l'équation ci-dessus admet toujours une unique solution $\vec{x}$ telle que $\sum_i x_i = 1$ et que tous les x_i sont strictement positifs. En 2005, Google utilisait la valeur $m = 0.15$. Nous ignorons si c'est encore la valeur utilisée aujourd'hui ; de plus, les diverses valeurs des x_i ne sont pas publiques.

3 Valeurs et vecteurs propres

Soit n un entier positif et soit A une matrice carrée $n \times n$.

Définition 3.1 *Un nombre λ est une **valeur propre** de A s'il existe un vecteur $\vec{x} \neq \vec{0}$ tel que*

$$A\vec{x} = \lambda\vec{x}.$$

*Un vecteur non nul $\vec{x}$ qui satisfait l'égalité $A\vec{x} = \lambda\vec{x}$ est appelé **vecteur propre** associé à la valeur propre λ. On note $V_\lambda(A)$ l'ensemble des vecteurs $\vec{v}$ tels que $A\vec{v} = \lambda\vec{v}$ et on l'appelle le **sous-espace propre** associé à λ.*

Il faut imposer $\vec{x} \neq \vec{0}$ car il est toujours vrai que $A\vec{0} = \lambda\vec{0}$ pour tout nombre λ. En particulier, $\vec{0}$ appartient à chaque sous-espace propre. $V_\lambda(A)$ est un sous-espace vectoriel de $\mathbb{R}^n$.

Rappelons le résultat suivant :

Théorème 3.2 *Un nombre λ est une valeur propre de A si et seulement si*

$$\det(A - \lambda I) = 0$$

où I désigne la matrice identité.

Preuve. Dire que λ est une valeur propre revient à dire que la matrice $A-\lambda I$ envoie *au moins deux* vecteurs différents (à savoir $\vec{x} \neq \vec{0}$ et $\vec{0}$) sur $\vec{0}$. Elle n'est donc pas inversible, et par conséquent son déterminant est nul. Réciproquement, si $\det(A - \lambda I) = 0$ alors la matrice $A - \lambda I$ n'est pas inversible. On démontre qu'elle n'est alors pas *injective*, ce qui signifie qu'il existe au moins deux vecteurs différents $\vec{v}$ et $\vec{w}$ tels que $A\vec{v} - \lambda\vec{v} = A\vec{w} - \lambda\vec{w}$. Par linéarité, cela s'écrit aussi $A(\vec{v} - \vec{w}) - \lambda(\vec{v} - \vec{w}) = 0$, donc $\vec{x} = \vec{v} - \vec{w}$ est non nul et satisait $A\vec{x} = \lambda\vec{x}$. Par suite, λ est une valeur propre de A. □

Corollaire 3.3 *La matrice A et sa transposée ${}^tA$ ont les mêmes valeurs propres.*

Preuve. Pour toute matrice carrée B, on a $\det({}^tB) = \det(B)$. Ainsi, $\det({}^tA-\lambda I) = \det(A-\lambda I)$, et $\det(A - \lambda I)$ et $\det({}^tA - \lambda I)$ s'annulent pour les mêmes valeurs de λ. □

La matrice A du paragraphe précédent possède une propriété partagée par les matrices de Google et qui joue un rôle crucial dans la résolution de l'équation $A\vec{x} = \vec{x}$: ses coefficients sont tous positifs ou nuls et la somme des composantes de chaque colonne vaut 1. Une telle matrice est dite **stochastique** (par rapport à ses colonnes). Cette propriété nous assure que l'équation $A\vec{x} = \vec{x}$ admet (au moins) une solution non nulle.

Corollaire 3.4 *Toute matrice stochastique par rapport à ses colonnes admet 1 comme valeur propre.*

Preuve. Par le corollaire 3.3, il suffit de démontrer que la transposée ${}^tA$ de A admet 1 comme valeur propre. Or la matrice ${}^tA$ est stochastique par rapport à ses lignes, c'est-à-dire que la somme des coefficients de chaque ligne de ${}^tA$ vaut 1. En prenant $\vec{v} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$, le vecteur ayant toutes ses composantes égales à 1, on voit immédiatement que ${}^tA\vec{v} = \vec{v}$. □

Le théorème 3.2 est malheureusement inutilisable dans le problème de Google vu la taille des matrices de Google. Il faut donc un autre moyen pour trouver $\vec{x}$. De plus, même si le corollaire 3.4 affirme l'existence d'un tel vecteur, il ne dit rien sur son unicité. Afin de pallier ces défauts, comme on l'a vu dans le paragraphe précédent, on utilise une matrice stochatique modifiée définie comme suit pour calculer le vecteur de score.

Rappelons que S est la matrice $n \times n$ dont tous les coefficients sont égaux à $1/n$, et que m désigne un nombre réel compris entre 0 et 1 (qui valait $m = 0.15$ en 2005). On remplace la matrice de Google A par la suivante :

$$M = (1 - m)A + mS$$

et on cherche à résoudre l'équation $M\vec{x} = \vec{x}$. Cela revient à ajouter du hasard pur dans le calcul du vecteur de score ; cela semble raisonnable dans le sens que lorsqu'on surfe sur internet, il est fréquent de cliquer sur des liens au hasard, de temps en temps. En 2005, les concepteurs de Google ont sans doute estimé à 15% la part de hasard dans les recherches sur internet.

La matrice M est encore stochastique par rapport à ses colonnes, mais en plus, *tous* ses coefficients sont positifs. On va démontrer dans le paragrahe suivant que pour une telle matrice, $V_1(M)$ est de dimension 1, donc qu'il existe un unique vecteur $\vec{x}$ dont la somme des composantes vaut 1, et de plus que toutes ses composantes sont positives. Avant cela, reprenons l'exemple du paragraphe 2.

Exemple. En prenant la matrice

$$A = \begin{pmatrix} 0 & 0 & 1 & 1/2 \\ 1/3 & 0 & 0 & 0 \\ 1/3 & 1/2 & 0 & 1/2 \\ 1/3 & 1/2 & 0 & 0 \end{pmatrix}$$

on obtient, avec la valeur $m = 0.15$,

$$M = \begin{pmatrix} 0.0375 & 0.0375 & 0.8875 & 0.4625 \\ 0.3208 & 0.0375 & 0.0375 & 0.0375 \\ 0.3208 & 0.4625 & 0.0375 & 0.4625 \\ 0.3208 & 0.4625 & 0.0375 & 0.0375 \end{pmatrix}$$

qui donne les scores $x_1 \approx 0.368$, $x_2 \approx 0.142$, $x_3 \approx 0.288$ et $x_4 \approx 0.202$. Cela donne le même classement que pour A, mais les scores sont légèrement différents.

4 Analyse de la matrice M

On considère une matrice $M = \begin{pmatrix} m_{1,1} & \dots & m_{1,n} \\ \vdots & \ddots & \vdots \\ m_{n,1} & \dots & m_{n,n} \end{pmatrix}$ qui est stochastique par rapport à ses colonnes et dont tous les coefficients $m_{i,j} > 0$; on dit qu'une telle matrice est **positive**. On rappelle que le sous-espace propre associé à la valeur propre 1 est

$$V_1(M) = \{\vec{x} \in \mathbb{R}^n | M\vec{x} = \vec{x}\}.$$

Proposition 4.1 *Soit M comme ci-dessus. Alors tout vecteur propre dans $V_1(M)$ a ses composantes soit toutes positives, soit toutes négatives.*

Preuve. Observons d'abord que si $\vec{y}$ est un vecteur à n composantes $y_1,\ y_2,\ \dots\ y_n$, alors $|\sum_i y_i| \leq \sum_i |y_i|$ et que l'inégalité est stricte si et seulement si certains y_i sont positifs et d'autres négatifs. Supposons alors par l'absurde qu'il existe $\vec{x} \in V_1(M)$ dont les composantes ont des signes mélangés (certaines positives et d'autres négatives). De l'équation $\vec{x} = M\vec{x}$, on déduit que $x_i = \sum_j m_{i,j}x_j$ pour tout i, et comme $m_{i,j} > 0$ pour tous i et j, on a :

$$|x_i| = \left|\sum_{j=1}^{n} m_{i,j}x_j\right| < \sum_j m_{i,j}|x_j|.$$

En sommant ces inégalités sur i de 1 à n, puis en intervertissant les sommes sur i et sur j on obtient

$$\sum_{i=1}^{n} |x_i| < \sum_{i=1}^{n}\sum_{j=1}^{n} m_{i,j}|x_j| = \sum_{j=1}^{n}\left(\sum_{i=1}^{n} m_{i,j}\right)|x_j| = \sum_{j=1}^{n} |x_j|$$

qui donne une contradiction. □

Proposition 4.2 *Soient $\vec{v}$ et $\vec{w}$ deux vecteurs linéairement indépendants dans $\mathbb{R}^n$. Alors il existe une valeur réelle s telle que le vecteur $\vec{x} = \vec{v} + s\vec{w}$ ou le vecteur $\vec{x} = s\vec{v} + \vec{w}$ admet des composantes positives et des composantes négatives.*

Preuve. Comme ils sont linéairement indépendants, ils sont tous deux non nuls. Posons $d = \sum_i v_i$. Si $d = 0$, alors $\vec{v}$ possède des composantes de chaque signe, et il suffit de prendre $s = 0$ et donc de poser $\vec{x} = \vec{v}$. Si $d \neq 0$, posons $s = -\dfrac{\sum_i w_i}{d}$ et $\vec{x} = s\vec{v} + \vec{w}$. Comme $\vec{v}$ et $\vec{w}$ sont linéairement indépendants, on a $\vec{x} \neq \vec{0}$ et par construction,

$$\sum_i x_i = s\sum_i v_i + \sum_i w_i = -\sum_i w_i + \sum_i w_i = 0.$$

Par ci-dessus, $\vec{x}$ a des composantes positives et des composantes négatives. □

Corollaire 4.3 *Si M est une matrice stochastique et positive, alors $V_1(M)$ est de dimension 1 et formé des multiples d'un vecteur $\vec{x}$ dont les composantes $x_i > 0$ pour tout i et telles que $\sum_i x_i = 1$.*

Preuve. Le fait que $V_1(M)$ n'est pas réduit à $\{\vec{0}\}$ suit du corollaire 3.4 du paragraphe précédent. Supposons par l'absurde que $\dim(V_1(M)) \geq 2$, et soient $\vec{v}$ et $\vec{w}$ deux vecteurs linéairement indépendants dans $V_1(M)$. Par la proposition 4.2, on peut fabriquer un vecteur $\vec{x} = \alpha\vec{v} + \beta\vec{w}$ qui possède des composantes positives et des composantes négatives. De plus, $\vec{x} \in V_1(M)$ car

$$M\vec{x} = \alpha M\vec{v} + \beta M\vec{w} = \alpha\vec{v} + \beta\vec{w} = \vec{x}.$$

Cela contredit la proposition 4.1 puisque les vecteurs non nuls de $V_1(M)$ ont leurs composantes soit toutes négatives, soit toutes positives. □

Enfin, il reste à montrer comment on procède pratiquement pour trouver le vecteur $\vec{x}$ du corollaire 3.3. Cela fera l'objet du paragraphe suivant.

5 Calcul du vecteur de score

Soit M une matrice stochastique par rapport à ses colonnes et positive. L'idée pour calculer le vecteur $\vec{x}$ tel que $\sum_i x_i = 1$ et $M\vec{x} = \vec{x}$ consiste à choisir un vecteur initial arbitraire $\vec{x}^{(0)}$ à composantes positives, puis à définir la suite de vecteurs $\vec{x}^{(1)}$, $\vec{x}^{(2)}$, $\ldots$, $\vec{x}^{(k)}$, $\ldots$ par $\vec{x}^{(k)} = M\vec{x}^{(k-1)}$, autrement dit $\vec{x}^{(k)} = M^k\vec{x}^{(0)}$. Il s'agit alors de démontrer que cette suite converge vers le vecteur cherché.

Notons V le sous-espace vectoriel de $\mathbb{R}^n$ formé des vecteurs $\vec{w}$ dont la somme des composantes $\sum_i w_i = 0$. C'est, pour le produit scalaire standard, l'orthogonal du vecteur

$$\vec{e} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

De plus, définissons la **1-norme** suivante pour les éléments de $\mathbb{R}^n$:

$$\|\vec{v}\|_1 := \sum_{i=1}^{n} |v_i|.$$

Proposition 5.1 *Soit M une matrice stochastique par rapport à ses colonnes et positive. Alors, pour tout $\vec{v} \in V$, on a $M\vec{v} \in V$ et*

$$\|M\vec{v}\|_1 \leq c\|\vec{v}\|_1$$

où

$$0 < c = \max_{1\leq j\leq n} |1 - 2 \cdot \min_{1\leq i\leq n} m_{i,j}| < 1.$$

Preuve. Pour $\vec{v} \in V$ fixé, posons $\vec{w} = M\vec{v}$, de sorte que $w_i = \sum_j m_{i,j}v_j$ et alors

$$\sum_i w_i = \sum_i \sum_j m_{i,j}v_j = \sum_j v_j \left(\sum_i m_{i,j} \right) = \sum_j v_j = 0$$

puisque la matrice M est stochastique et que $\sum_j v_j = 0$. Ainsi, $\vec{w} = M\vec{v} \in V$. Pour prouver l'inégalité, notons $e_i = \text{sgn}(w_i) = \pm 1$ le signe de w_i, de sorte que

$$\|\vec{w}\|_1 = \sum_i e_i w_i = \sum_i e_i \left(\sum_j m_{i,j}v_j \right).$$

Observons que les e_i n'ont pas tous le même signe car $\vec{w} \in V$ (sauf bien sûr au cas où $\vec{w} = \vec{0}$, mais alors l'inégalité est trivialement vraie). Posons encore $a_j = \sum_i e_i m_{i,j}$ et intervertissons la double somme ci-dessus pour obtenir

$$\|\vec{w}\|_1 = \sum_j v_j \left(\sum_i e_i m_{i,j} \right) = \sum_j a_j v_j.$$

Puisque les signes des $e_i = \pm 1$ varient, que $e_i + 1 = 0$ ou 2 et que $\sum_i m_{i,j} = 1$, avec $0 < m_{i,j} < 1$, on obtient d'une part

$$a_j + 1 = \sum_i (e_i + 1)m_{i,j} \geq 2 \cdot \min_i m_{i,j}$$

et d'autre part, puisque $e_k - 1 = -2$ ou 0 et que $m_{k,j} \geq \min_i m_{ij}$ et

$$a_j - 1 = \sum_k (e_k - 1)m_{k,j} \leq -2 \cdot \min_i m_{i,j}.$$

Cela implique immédiatement que

$$-1 < -1 + 2 \cdot \min_i m_{i,j} \leq a_j \leq 1 - 2 \cdot \min_i m_{i,j} < 1.$$

Ainsi, $|a_j| \leq |1 - 2 \cdot \min_i m_{i,j}| \leq c < 1$. On a finalement

$$\|\vec{w}\|_1 = \sum_j a_j v_j \leq \left| \sum_j a_j v_j \right| \leq \sum_j |a_j||v_j| \leq c \sum_j |v_j| = c\|\vec{v}\|_1,$$

ce qui prouve la proposition. □

Voici enfin le résultat qui garantit la convergence de la suite des $\vec{x}^{(k)}$:

Théorème 5.2 *Soit M une matrice comme dans la proposition 2. Alors elle admet un unique vecteur $\vec{q} \in V_1(M)$ à composantes toutes positives et tel que $\|\vec{q}\|_1 = 1$. Il peut être calculé par*

$$\vec{q} = \lim_{k\to\infty} M^k \vec{x}^{(0)}$$

à partir de n'importe quel vecteur initial $\vec{x}^{(0)}$ à composantes positives et tel que $\|\vec{x}^{(0)}\|_1 = 1$.

Preuve. On sait déjà que $\vec{q}$ existe et est unique, par le paragraphe précédent. Une fois le vecteur intial $\vec{x}^{(0)}$ choisi tel que $x_j^{(0)} > 0$ pour tout j et $\|\vec{x}^{(0)}\|_1 = 1$, posons $\vec{v} = \vec{x}^{(0)} - \vec{q}$ de sorte que $\vec{x}^{(0)} = \vec{q} + \vec{v}$ et $\vec{v}$ appartient à V, c'est-à-dire que la somme des composantes de $\vec{v}$ vaut 0. On a alors pour tout k :

$$M^k \vec{x}^{(0)} = M^k \vec{q} + M^k \vec{v} = \vec{q} + M^k \vec{v}$$

car $M\vec{q} = \vec{q}$ donc $M^k \vec{q} = \vec{q}$ pour tout $k > 0$ et

$$M^k \vec{x}^{(0)} - \vec{q} = M^k \vec{v}.$$

Par une récurrence évidente, on obtient, par la proposition 5.1 : $\|M^k \vec{v}\|_1 \leq c^k \|\vec{v}\|_1$ pour tout k, et comme $0 < c < 1$, on a

$$\lim_{k\to\infty} \|M^k \vec{v}\|_1 = 0$$

ce qui permet de conclure que

$$\vec{q} = \lim_{k\to\infty} M^k \vec{x}_0.$$

□

Dans le cas de la matrice de Google, le calcul direct de la suite $(\vec{x}^{(k)})_{k\geq 1}$ par récurrence comme présenté ci-dessus serait malgré tout très fastidieux puisque chaque coefficient de $\vec{x}^{(k)}$ nécessiterait des milliards de multiplications et d'additions. En fait, on rend le calcul raisonnable et efficace grâce à la forme spéciale de la matrice M ; comme on l'a vu, $M = (1-m)A + mS$ et A est une matrice creuse. De plus, on observe la particularité suivante de S : Soit $\vec{v} \in \mathbb{R}^n$ un vecteur dont la somme des composantes vaut 1. Alors

$$S\vec{v} = \begin{pmatrix} 1/n & \dots & 1/n \\ \vdots & \dots & \vdots \\ 1/n & \dots & 1/n \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} \frac{v_1+\dots+v_n}{n} \\ \vdots \\ \frac{v_1+\dots+v_n}{n} \end{pmatrix} = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix} =: \vec{u}.$$

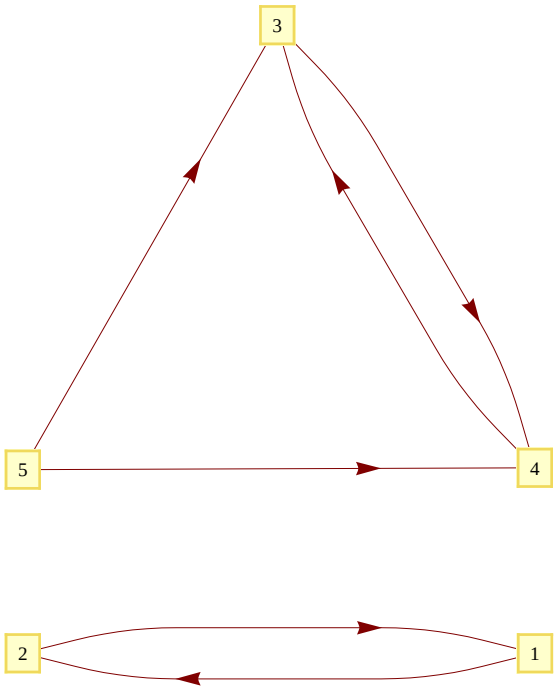
Par suite, quel que soit k, on a $mS\vec{v}^{(k)} = m\vec{u}$ qui est indépendant de k et ainsi la récurrence est en réalité :

$$\vec{x}^{(k+1)} = (1-m)A\vec{x}^{(k)} + m\vec{u}$$

qui peut se calculer plus facilement puisque chaque composante de $\vec{x}^{(k+1)}$ ne fait intervenir que quelques dizaines de composantes de $\vec{x}^{(k)}$. On peut démontrer qu'une cinquantaine d'itérations suffisent pour obtenir un vecteur de score satisfaisant.

6 Exercices

Exercice 1. Écrire la matrice A du web suivant :



Exercice 2. Considérons la matrice

$$A = \begin{pmatrix} 0.5 & 0.3 \\ 0.5 & 0.7 \end{pmatrix}.$$

Calculer le vecteur de score de $M = (1-m)A + mS$ pour $0 \leq m < 1$. Calculer également la valeur de c dans la proposition 5.1 en fonction de m et en déduire la rapidité de la convergence de la suite $M^k \vec{x_0}$ vers $\vec{q}$.

Exercice 3. Dans le calcul de $\vec{q}$ par itérations (théorème 5.2), en notant comme dans la preuve $\vec{x}_0$ un vecteur initial et $\vec{v} = \vec{v}_0 - \vec{q}$, démontrer que

$$\|M^k \vec{v}\|_1 \leq 2 \cdot c^k$$

pour tout k. Si on veut chaque composante de $\vec{q}$ avec 3 décimales exactes, démontrer que l'on peut prendre

$$k > \frac{\ln(2000)}{\ln(1/c)}$$

et estimer k pour diverses valeurs de $0 < c < 1$.

7 Références

[1] K. Bryan and T. Leise. *The $ 25,000,000,000 eigenvector ; the linear algebra behind Google*, preprint, 10 pages.

[2] A. Longville and C. Meyer. *Google's PageRank and Beyond*, Princeton University Press, Princeton, 2006.

[3] Y. Ollivier and P. Senellart. *Finding Related Pages Using Green Measures : An Illustration with Wikipedia*, preprint, 10 pages.