Aspects locaux de l'importance globale des pages web

Fabien Mathieu et Laurent Viennot

LIRMM, 34392 Montpellier Cedex 5 France
INRIA Rocquencourt, F-78153 Le Chesnay (France)

Lancé en 1998, le moteur de recherche *Google* classe les pages grâce à la combinaison de plusieurs facteurs dont le principal porte le nom de *PageRank*. Plus précisément, le classement des pages est fait en utilisant un indice numérique (le «PageRank») calculé pour chaque page. Nous allons montrer qu'il est possible de décomposer le PageRank en deux parties distinctes, que nous appellerons PageRank interne et PageRank externe. Ces deux PageRank jouent des rôles fondamentalement différents, et leur introduction permet de mieux comprendre comment fonctionne le PageRank à l'intérieur et à l'extérieur d'un site. Une première application est un algorithme local d'estimation du PageRank des pages d'un site. Nous allons également mettre en évidence des résultats quantitatifs sur la possibilité pour un site de «doper» son propre PageRank.

Keywords: PageRank, graphe du web, flot, chaînes de Markov

1 Introduction

Soit $G = (V,E)$ un graphe orienté fortement connexe apériodique sans boucle, et $\mathcal{S} = (S_1,\dots,S_k)$ une partition de G non triviale. Si G représente une partie du graphe du web, $\mathcal{S}$ peut par exemple être un découpage en «sites» de ce graphe (un élément de $\mathcal{S}$ regroupe les page d'un même site). $d^+(v)$ étant le degré sortant d'une page v de V, on considère la matrice A de V dans V définie par :

$$A = (a_{i,j})_{i,j\in V}, \text{ avec } a_{(i,j)} = \begin{cases} \dfrac{1}{d^+(i)} \text{ si } i \text{ pointe vers } j \\ 0 \text{ sinon} \end{cases}$$

D'après la théorie de Markov sur les processus stochastiques sans mémoire[SC96], il existe une unique distribution de probabilité P sur V vérifiant :

$$\forall v \in V,\ P(v) = \sum_{w\rightarrow v} \frac{P(w)}{d^+(w)} \tag{1}$$

Matriciellement, ce résultat peut s'écrire :

$$P = A^t P, \tag{2}$$

A^t étant la matrice transposée de A.

Cette distribution P est par définition le PageRank associé au graphe G. Introduit par [PBMW98] en 1998, le concept de PageRank a été popularisé par le moteur de recherche *Google*[Goo98]. L'objet de cette article est de décomposer le PageRank en deux parties distinctes, que nous appellerons PageRank entrant interne et PageRank entrant externe. Comme nous allons le voir, ces deux PageRank jouent des rôles fondamentalement différents, et leur introduction permet de mieux comprendre comment fonctionne le PageRank à l'intérieur et à l'extérieur d'un site. Une application directe est un algorithme local d'estimation du PageRank des pages d'un site.

Remarque : Le graphe du web n'est pas un graphe fortement connexe. Cependant, il existe de nombreuses techniques pour se ramener à un tel graphe, dont la liste non exhaustive suit :
– [PBMW98] propose de compenser le caractère sous-stochastique de la matrice A en renormalisant le vecteur P à chaque itération.
– [Hav99] rend A explicitement stochastique en éliminant itérativement les pages sans lien.
– Le facteur d'amortissement, introduit par [BP98] est utilisé par *Google* sur un graphe dont les feuilles ont été enlevées (sans itérer le procédé) et sont réinjectées après convergence de P. Ce procédé consiste à remplacer la matrice A par $d.A+\frac{1-d}{|V|}\mathbf{1}.\mathbf{1}^t$, où $\mathbf{1}$ est le vecteur ne contenant que des 1. On obtient ainsi un graphe fortement connexe pondéré (une pondération non-uniforme ne change rien à l'algorithme, l'essentiel étant de conserver une matrice stochastique).
– Enfin, [Abi01] rajoute une page virtuelle de *zap* qui pointe et est pointée par toute page.

Dans la suite de cet article sauf dans 2.3, 4.2 et 4.3, nous supposerons que le graphe correspondant à la matrice A est fortement connexe apériodique et sans boucle. PageRank désignera donc sans ambiguïté le vecteur de probabilité P vérifiant $P = A^tP$.

2 PageRank interne, PageRank externe

2.1 Notations

Pour v dans V, nous noterons $S(v)$ l'élément de $\mathcal{S}$ qui contient v. Nous allons également définir la fonction $\delta_{\mathcal{S}}$ dans $V\times V$ par :

$$\delta_{\mathcal{S}}(v,w)=\begin{cases}1 \text{ si S(v)=S(w)}\\ 0 \text{ sinon}\end{cases}$$

On pose également $A_{\mathcal{S}}=(a_{v,w}\delta_{\mathcal{S}}(v,w))_{v,w\in V}$, matrice diagonale par blocs selon $\mathcal{S}$.

Nous définirons le degré interne d_i^+ (resp. degré externe d_e^+) d'un sommet v comme son degré sortant mesuré sur le graphe induit par $S(v)$ (resp. $\{v\}\cup(V\setminus S(v))$).

Nous allons maintenant introduire différents PageRank, calculés à partir du PageRank sur G noté P et défini par la relation (2) :
– Le PageRank entrant interne P_{ei} (resp. entrant externe P_{ee}) d'une page v est la probabilité de venir en v à partir d'une page de $S(v)$ (resp. $E\setminus S(v)$), soit :

$$P_{ei} = A_{\mathcal{S}}^tP \quad (3)$$
$$P_{ee} = (A-A_{\mathcal{S}})^tP = P-P_{ei}, \text{ puisque } A^tP=P \quad (4)$$

– Le PageRank sortant interne P_{si} (resp. sortant externe P_{se}) d'une page v est égal au produit $P(v)\frac{d_i^+(v)}{d^+(v)}$ (resp. $P(v)\frac{d_e^+(v)}{d^+(v)}$). Matriciellement, cela donne :

$$P_{si} = (A_{\mathcal{S}}\mathbf{1}).\times P$$
$$P_{se} = ((A-A_{\mathcal{S}})\mathbf{1}).\times P,$$

où $.\times$ désigne le produit terme à terme.

2.2 Lois de conservation

Nous allons écrire les lois de conservation des PageRank. Tout d'abord, d'après les définitions, on a :

$$P = P_{ee}+P_{ei} = P_{se}+P_{si} \quad (5)$$

Plus intéressantes sont les lois de conservation externe et interne. En effet, à $S\in\mathcal{S}$ fixé, on constate que

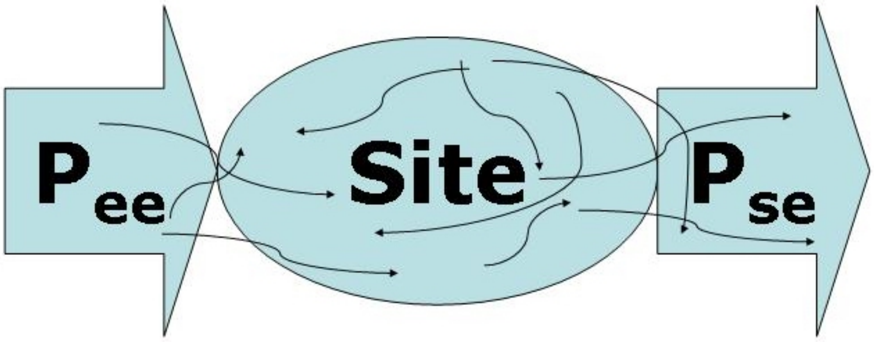

FIG. 1: Conservation du PageRank externe : $\sum_{v\in S} P_{ee}(v) = \sum_{v\in S} P_{se}(v)$

$$\sum_{v\in S} P(v) = \sum_{v\in S}\sum_{w\to v} \frac{P(w)}{d^+w} = \sum_{(w,v)\in E\cap V\times S} \frac{P(w)}{d^+w} \tag{6}$$

$$= \sum_{(w,v)\in E\cap S^2} \frac{P(w)}{d^+w} + \sum_{(v,w)\in E\cap(S\times V\setminus S)} \frac{P(w)}{d^+w} \tag{7}$$

$$= \sum_{w\in S} P_{si}(w) + \sum_{v\in S} P_{ee}(v) \tag{8}$$

On déduit des relations (5) et (8) les lois de conservation externe

$$\sum_{v\in S} P_{ee}(v) = \sum_{v\in S} P_{se}(v) \tag{9}$$

et interne

$$\sum_{v\in S} P_{ei}(v) = \sum_{v\in S} P_{si}(v). \tag{10}$$

La relation (9) indique qu'un site distribue autant de PageRank (sortant externe) qu'il en reçoit (entrant externe). Si l'on voit le PageRank comme un flot de surfeurs aléatoires, il y a donc une conservation du flot associé au PageRank externe sur le graphe quotient $G/\mathcal{S}$ (voir la figure 1). C'est en partant de ce constat que nous allons déduire un calcul de PageRank intra-sites et inter-sites.

Remarque : S'inspirant de l'interprétation ci-dessus, il existe une preuve plus simple de (9), obtenue en formalisant rigoureusement le PageRank en termes de théorie des flots. On peut en effet voir le PageRank comme un flot stationnaire. Le flot est conservé en tout sommet et donc aussi en tout ensemble de sommets, d'où la relation (9).

2.3 Facteur d'amortissement et flot

Jusqu'à la prochaine section, nous supposerons toujours que le graphe est sans feuille, mais nous n'imposerons plus qu'il soit fortement connexe apériodique.

Comme il a été dit, le facteur d'amortissement consiste à substituer à la matrice A la matrice $d.A + \frac{1-d}{|V|}\mathbf{1}\mathbf{1}^t$. Aux transitions classiques ($d.A$) s'ajoutent maintenant des transitions dites de *zap*, censées modéliser l'action de se déplacer n'importe où sur le web en un coup sans cliquer (*Bookmarks*, adresse tapée à la main, utilisation d'un moteur de recherche...). Plutôt que de séparer le flot de *zap* en flot interne et externe, il apparaît plus intéressant d'introduire les notions de PageRank induit, noté P_{ind} et de PageRank dissipé P_{dis}. Nous allons donc avoir six définitions correspondant à trois types de flot :

Flot	entrant	sortant
interne	$P_{ei} = dA^t_{\mathcal{S}}P$	$P_{si} = d(A_{\mathcal{S}}\mathbf{1}).\times P$
externe	$P_{ee} = d(A - A_{\mathcal{S}})^t P$	$P_{se} = d((A - A_{\mathcal{S}})\mathbf{1}).\times P$
de *zap*	$P_{ind} = \frac{1-d}{\lvert V\rvert}\mathbf{1}$	$P_{dis} = (1-d)P$

FIG. 2: Les différents flots dans un PageRank avec *zap*

Si l'on considère que, par convention, l'intégralité du flot de *zap* est un flot de type «externe»[†], l'équation de conservation interne est inchangée. En revanche, on a une nouvelle équation de conservation du flot externe :

$$\sum_{v\in S} (P_{ee}(v) + P_{ind}(v)) = \sum_{v\in S} (P_{se}(v) + P_{dis}(v)),$$

où encore en abrégé

$$P_{ee}(S) + P_{ind}(S) = P_{se}(S) + P_{dis}(S) \tag{11}$$

Stabilité du PageRank : L'équation (11) permet de mettre en évidence la stabilité du flot «classique» au niveau d'un site. En effet, comme $P_{ind}(S) = (1-d)\frac{|S|}{|V|}$ et $P_{dis}(S) = (1-d)P(S)$, on en déduit que si un site a un PageRank $P(S)$ supérieur (resp. inférieur) à la moyenne (qui est de $\frac{|S|}{|V|}$ pour un site de taille $|S|$), alors $P_{es}(S)$ est inférieur (resp. supérieur) à $P_{ee}(S)$. Plus clairement, un site «riche» en PageRank sera systématiquement «avare» en donnant moins qu'il ne reçoit (hors *zap*), et réciproquement. Grâce à l'amortissement, on voit apparaître une rétro-action qui empêche les phénomènes de sur-amplification que nous verrons en 4.2.

3 Calcul local de l'importance globale

3.1 Relation entre PageRank externe et PageRank

Les relations (3) et (4)nous permettent d'écrire $A^t_{\mathcal{S}}.P = P - P_{ee}$, d'où

$$P = (Id - A^t_{\mathcal{S}})^{-1} P_{ee} \tag{12}$$

où Id est la matrice identité sur V.

Notons que $(Id - A^t_{\mathcal{S}})$ est inversible sous nos hypothèses. En effet, comme G est fortement connexe, il existe forcément des liens entre sites différents. La matrice $A_{\mathcal{S}}$ est donc strictement inférieur à A, *i.e.* strictement sous-stochastique. Son rayon spectral est donc strictement inférieur à 1, ce qui assure que $(Id - A^t_{\mathcal{S}})$ est inversible.

Connaissant le PageRank extérieur P_{ee} auquel il est soumis, il est donc en théorie possible de déduire le PageRank de toutes les pages d'un site donné S avec la seule connaissance locale des liens internes au site.

Remarque : $(Id - A^t_{\mathcal{S}})^{-1} = \sum_{k=0}^{\infty} (A^t_{\mathcal{S}})^k$ est une matrice diagonale par blocs suivant $\mathcal{S}$. Elle peut s'interpréter sur chaque site comme étant la matrice de transition de tous les chemins internes possibles.

3.2 Matrice du PageRank externe

Il s'agit ici de traduire l'intuition de la figure 1 par une relation de conservation ne faisant intervenir que Pe. D'après les relations (4) et (12), on peut écrire :

$$P_{ee} = (A - A_{\mathcal{S}})^t P = (A - A_{\mathcal{S}})^t (Id - A^t_{\mathcal{S}})^{-1} P_{ee} \tag{13}$$

On obtient alors la matrice de transition du PageRank externe, notée A_e :

[†] C'est en toute rigueur faux si on considère qu'il est possible de *zapper* sur une page interne au site. On peut malgré tout supposer qu'on colorie toutes les transitions de *zap* de sorte à pouvoir les considérer comme externes dans tous les cas.

$$A_e^t = (A - A_S)^t (Id - A_S^t)^{-1}$$

Montrons que A_e est stochastique, *i.e.* vérifions que la somme des entrées d'une colonne de A_e^t vaut 1. On constate tout d'abord que

$$\begin{aligned} A_e^t &= \sum_{k=0}^{\infty} \left(A^t (A_S^t)^k - (A_S^t)^{k+1} \right) = A^t + \sum_{k=1}^{\infty} \left(A^t (A_S^t)^k - (A_S^t)^k \right) \\ &= A^t + A^t M - M, \text{ avec } M = \sum_{k=1}^{\infty} (A_S^t)^k \end{aligned}$$

Considérons la somme s_w de la colonne w de $A^t M$.

$$s_w = \sum_{u \in V} \sum_{v \in V} A_{u,v}^t M_{v,w} = \sum_{v \in V} \left(\sum_{u \in V} A_{u,v}^t \right) M_{v,w} = \sum_{v \in V} M_{v,w}$$

La somme de chaque colonne de $A^t M - M$ est nulle et A_e^t est donc stochastique par ses colonnes comme A^t.

3.3 Algorithme semi-distribué de calcul de PageRank

Les équations (12) et (13) nous permettent d'implémenter un algorithme semi-distribué calculant le PageRank :

- Chaque site S calcule à partir du bloc de A_S dont il dispose un bloc de la matrice de relation PageRank externe/PageRank $(Id - A_S^t)^{-1}$.
- Les coefficient de A_e sont confiés à un organisme central.
- Le dispositif central calcule le PageRank P'_e associé à A_e qui donne l'ensemble des PageRank externes ($P'e$ est défini par $A_e^t P'_e = p'_e$).
- Chaque site obtient alors son propre PageRank grâce à la relation $P' = (Id - A_S^t)^{-1} P'e$ appliquée sur son bloc.

Vérifions que le PageRank P' ainsi obtenu est bien égal, à un facteur de normalisation près, au PageRank P du graphe G. On constate que

$$\begin{aligned} A^t P' &= A^t (Id - A_S^t)^{-1} P'_e \\ &= (A^t - A_S)(Id - A_S^t)^{-1} P'_e + A_S^t (Id - A_S^t)^{-1} P'_e \\ &= A_e^t P'_e + \left((Id - A_S^t)^{-1} - (Id - A_S^t)(Id - A_S^t)^{-1} \right) P'_e \\ &= P'_e + \left((Id - A_S^t)^{-1} - Id \right) P'_e \\ &= P'_e + P' - P'_e = P' \end{aligned}$$

La valeur propre dominante de A étant simple, on a bien $P = P'$, après normalisation.

3.4 Estimation du PageRank d'un site

La question se pose de savoir si un site web peut, à partir des seules données locales dont il dispose, estimer l'importance de ses pages. Nous allons voir qu'il est possible, à un facteur de normalisation prêt, d'avoir une estimation du PageRank. Cela est important pour un moteur de recherche local qui pourra ainsi estimer l'importance des pages du site comme s'il avait une vision globale du graphe du web. D'après (12), il suffit d'estimer le PageRank entrant externe. Or, d'après [PBMW98], le PageRank est une modélisation du comportement statistique des surfeurs naviguant sur le web. Il est donc naturel d'estimer le PageRank d'une page donnée par le nombre de hits par unité de temps que cette page reçoit en moyenne. Plus spécifiquement, le PageRank entrant externe devrait être proportionnel au débit moyen de hits provenant de l'extérieur du site. Chaque site peut donc avoir à un facteur près une estimation du rang entrant externe de ses pages en analysant les fichiers de logs de son serveur web.

Il suffit ensuite de calculer $P = (Id - A^t_S)^{-1} P_{ee}$ pour le bloc correspondant au site. Remarquons d'ailleurs que si l'on possède une estimation de P_{ee}, il n'est pas nécessaire d'inverser explicitement la matrice $(Id - A^t_S)$. Il suffit de résoudre l'équation $P = A^t_S P + P_{ee}$ sur S de manière itérative par une méthode de type Jacobi, en choisissant un vecteur P_0 initial (uniforme ou égal à la précédente estimation du PageRank) et en utilisant la convergence de :

$$P_{n+1} = A^t_S P_n + P_{ee}$$

Le rayon spectral de A_S étant strictement inférieur à 1, un théorème de point fixe nous assure une convergence géométrique. Qui plus est, les résultats empiriques rapportés dans [PBMW98] indiquent une convergence rapide en pratique de ce genre d'algorithme appliqué au web.

Remarque : Des méthodes bien connues des numériciens permettent d'accélérer encore la convergence pour ce genre de problème, comme par exemple la méthode de Gauss-Seidel ou l'emploi d'un facteur de relaxation[AK98]. Lors de la réalisation effective de notre algorithme, ce sera a priori ce genre de méthodes qui sera mise en œuvre.

Une alternative plus simple consisterait à estimer l'importance des pages du site par leur fréquentation. Cependant, cette méthode a le défaut de sous-estimer l'importance des pages peu fréquentées et des nouvelles pages. D'autre part, pour être plus proche de la définition du PageRank dans notre méthode, on pourrait aussi estimer le PageRank entrant externe d'une page par le nombre de référeurs externes répertoriés dans les fichiers de logs pour cette page (le degré entrant d'une page donne souvent une bonne idée de son PageRank).

Notre méthode permet de plus de moduler manuellement l'estimation des PageRank : l'administrateur du site peut pour son estimation augmenter arbitrairement le PageRank externe de certaines pages qui lui paraissent plus importantes que ce que reflètent les fichiers de logs.

4 Altérations locales du PageRank

Notre façon de décomposer le PageRank permet de proposer quelques résultats sur la possibilité pour un site de contrôler son propre PageRank. Tout vient du fait que l'on peut en première approximation, supposer que si un site S a peu de pouvoir sur le PageRank externe qu'il reçoit P_{ee}, il en est autrement du PageRank interne, qui ne dépend comme nous l'avons vu que de P_{ee} et du graphe restreint au site.

4.1 Facteur d'amplification

Soit S un site, $P(S) := \sum_{v \in S} P(v)$ et $P_{ee}(S) := \sum_{v \in S} P_{ee}(v)$. On peut définir le coefficient d'amplification de S par $\alpha(S) = \frac{P(S)}{P_{ee}(S)}$. Ce facteur dépend à la fois de S et de P_{ee}, mais la connaissance de S nous fournit un encadrement de $\alpha(S)$.

En effet, si on appelle $\omega = \min_{v \in S} \frac{d_i(v)}{d(v)}$ et $\Omega = \max_{v \in S} \frac{d_i(v)}{d(v)}$, on obtient l'encadrement suivant :

$$\frac{1}{1-\omega} \leq \alpha(S) \leq \frac{1}{1-\Omega} \tag{14}$$

Preuve : Pour tout vecteur élémentaire e_v, $v \in S$, on a $\|A_S(e_v)\|_1 = \frac{d_i(v)}{d(v)}$, d'où l'on déduit par additivité $\omega \|(X)\|_1 \leq \|A_S X\|_1 \leq \Omega \|X\|_1$ pour tout vecteur X positif défini sur le site S.

La première partie de (14) s'obtient alors ainsi :

$$\begin{aligned} P(S) &= \sum_{v\in S} P(v) = \left\| \sum_{k\in\mathbb{N}} (A_S^t)^k (P_{ee}) \right\|_1 \\ &\geq \sum_{k=0}^{\infty} \omega^k \|P_{ee}\|_1 = \frac{1}{1-\omega} P_{ee}(S) \end{aligned}$$

et la deuxième se trouve de la même manière.

On voit quelles peuvent être les conséquences d'un tel système d'amplification : un site peut augmenter arbitrairement son amplification. Dans le cas limite où aucune page ne pointe vers l'extérieur[‡], on a un phénomène de «court-circuit». On retrouve alors le fait bien connu que si il existe des sous-parties fortement connexes, celles-ci vont absorber tout le PageRank qu'elles reçoivent jusqu'à l'assécher.

Heureusement, nous allons voir que dans une certaine mesure l'ajout d'un facteur d'amortissement permet d'atténuer cet effet.

4.2 Amortissement et amplification

Nous nous plaçons à nouveau dans les hypothèses de 2.3. En particulier, la matrice de transition est du type $d.A + \frac{1-d}{|V|}\mathbf{1}\mathbf{1}^t$; les résultats précédents restent valables en remplaçant A par dA et P_{ee} par le PageRank reçu total $P_{ee}+P_{ind}$. Le nouvel encadrement des valeurs possibles du nouveau facteur d'amplification $\alpha'(S) = \frac{P(S)}{P_{ee}(S)+P_{ind}(S)}$ est alors :

$$\frac{1}{1-d\omega} \leq \alpha'(S) \leq \frac{1}{1-d\Omega} \tag{15}$$

En effet, en s'inspirant de la démonstration de (14), on peut écrire que :

$$\begin{aligned} P(S) &= \sum_{v\in S} P(v) = \left\| \sum_{k\in\mathbb{N}} (dA_S^t)^k (P_{ee} + \frac{1-d}{|V|}\mathbf{1}) \right\|_1 \\ &\leq \sum_{k=0}^{\infty} (d\Omega)^k (\|P_{ee}\|_1 + (1-d)\frac{\|\mathbf{1}\|_1}{|V|}) \\ &\leq \frac{1}{1-d\Omega}(P_{ee}(S) + (1-d)\frac{|S|}{|V|}) \end{aligned}$$

la borne inférieure se trouve de la même manière.

Il n'est pas impossible *a priori* pour un site d'avoir $\omega = \Omega = 0$ (site sans lien interne) ou $\omega = \Omega = 1$ (site sans lien externe). Le facteur d'amplification varie donc entre 1 et $\frac{1}{1-d}$. La valeur empirique pour d étant autour de $0,85$, on en déduit que suivant sa politique d'hyperliens, un site de $|S|$ pages recevant un PageRank entrant externe «classique» fixé peut voir son PageRank total fluctuer jusqu'à un facteur $\frac{20}{3}$...

4.3 Amplification d'une page donnée

Quand un internaute lance une requête sur *Google*, le résultat lui est renvoyé sous forme de pages web triées suivant de multiples critères, dont l'un est le PageRank. Si un site veut se faire connaître, l'important n'est pas tant d'avoir un fort PageRank total que de concentrer ce PageRank sur quelques pages, voire une seule. La question que l'on se pose est donc : étant donné un site S de taille $n+1$ et un flot entrant P_{ee}, comment maximiser le PageRank d'une page fixée $v_0 \in S$? En fait, le problème n'est pas compliqué une

[‡] Un site réel n'est pas tenu de respecter les hypothèses de cet article. En particulier, beaucoup de sites commerciaux ne laissent pas de sortie à leurs visiteurs.

fois que l'on a vu que la structure optimale est celle en rayons, où toutes les pages filles $v \neq v_0$ pointent et sont pointées par v_0[§]. Il n'est alors guère difficile de majorer le PageRank de v_0 :

$$P(v_0) \leq \frac{P_{ee}(S)}{1-d^2} + \frac{1+nd}{(1+d)\,|V|}, \tag{16}$$

le cas d'égalité étant atteint si $P_{ee}(S) = P_{ee}(v_0)$.

L'équation (16) met en évidence quelques stratégies pour améliorer son classement dans *Google*. Par exemple :

– Si le propriétaire d'un site en rayons sans retour (v_0 pointe vers les pages filles sans être pointé par elles)[¶] rajoute les liens de retour, il peut voir sa page d'accueil augmenter son PageRank jusqu'à un facteur $\frac{1}{1-d^2} \simeq 3,6$.

– La stratégie en rayons assure pour v_0 un PageRank au moins égal au PageRank moyen $\frac{1}{|V|}$ même si P_{ee} est nul.

– Pour $1 \ll n \leq |V|$ (un site générant automatiquement des pages en rayon), le rapport $\frac{P(v_0)}{P_{moyen}}$ est de l'ordre de $\frac{d}{1+d}n$.

5 Conclusion

Nous venons de définir un découpage du flot de PageRank suivant une notion de site. Cette méthode ouvre la voie à tout un éventail de techniques de calcul en local du PageRank. Par exemple, une variante de l'algorithme semi-distribué de calcul de PageRank beaucoup moins coûteuse en ressources que l'algorithme proposé dans cet article consisterait à simplifier l'étape d'inversion de matrice. Il suffirait de condenser le PageRank externe en un unique PageRank par site que l'on pourrait facilement calculer en redirigeant tous les liens venant de l'extérieur sur une unique page du site (la page d'accueil par exemple). L'inversion de $(1-A_s)^{-1}$ devient alors facile à calculer, et A_e est de taille bien inférieure à A puisque équivalente à une matrice stochastique sur le graphe quotient par site.

La décomposition du flot a permis également d'analyser finement les stratégies que pourrait employer un concepteur de site de manière à maximiser le PageRank de ses pages. Une certaine versatilité du PageRank a été mise en évidence au passage, qui semble indiquer que le PageRank tel qu'il est défini dans les articles de recherche est peu résistant à des stratégies non coopératives. Il nous semble en revanche que P_{ee} fournit des informations plus fiables, à la condition d'être capable de trouver une véritable partition en sites, et non de se contenter d'une partition en serveur.

Références

[Abi01] S. Abiteboul. Page rank incremental, novembre 2001. personnal communication with L. Viennot.

[AK98] G. Allaire and S. M. Kaber. *Algèbre linéaire numérique*. Ellipses, 1998.

[BP98] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7) :107–117, 1998.

[Goo98] Google. http ://www.google.com/, 1998.

[Hav99] T. Haveliwala. Efficient computation of PageRank. Technical report, Computer Science Department, Stanford University, 1999.

[PBMW98] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking : Bringing Order to the Web. Technical report, Computer Science Department, Stanford University, 1998.

[SC96] L. Saloff-Coste. Lectures on finite Markov chains. In G.R. Grimmet E. Giné and L. Saloff-Coste, editors, *Lecture Notes on Probability Theory and Statistics*, number 1665 in LNM, pages 301–413. Springer Verlag, 1996.

[§] Les algorithmes de type PageRank éliminent systématiquement les boucles, ce qui fait qu'une page seule ne peut pas s'auto-amplifier.

[¶] La situation est souvent rencontrée dans le cas de sites avec *frames*.