

Le PageRank : ordonner le web en tant que tel

Présentation, calcul et propriétés

Guillaume Seguin

26 janvier 2009

Résumé

Ce document présente l'algorithme du PageRank et plusieurs de ses propriétés, issues de [Bianchini *et al.*, 2005]. Nous poserons dans une première partie le contexte d'utilisation du PageRank et nous expliciterons son calcul original tel que décrit dans [Brin et Page, 1998]. Nous nous intéresserons ensuite aux propriétés de cet algorithme, en présentant tout d'abord la notion de *balance d'énergie* et en donnant ses implications dans la visibilité des pages pour le PageRank, puis en donnant deux résultats de calculabilité et complexité : le PageRank peut être calculé sur un graphe évoluant dans le temps, et il peut être calculé par un algorithme en temps $\mathcal{O}(|H| \cdot \log(1/\epsilon))$, avec $|H|$ le nombre de liens sur Internet et ϵ une borne supérieure sur l'erreur commise dans le calcul.

Table des matières

1	Présentation et calcul du PageRank	2
1.1	Présentation	2
1.1.1	Problématique	2
1.1.2	Le PageRank	2
1.2	Calcul	3
1.2.1	Algorithme de base	3
1.2.2	Problème des pages sans liens	3
1.2.3	Importance du facteur d'atténuation	4
2	Propriétés du PageRank	5
2.1	Balance d'énergie et implications	5
2.1.1	Balance d'énergie	5
2.1.2	Conséquences sur la visibilité des pages	6
2.2	Calculabilité et complexité	7
2.2.1	Calcul sur un graphe évoluant dans le temps	7
2.2.2	Complexité temporelle	7
	Conclusion	8

1 Présentation et calcul du PageRank

Dans cette section nous posons le cadre du problème, l'utilité du PageRank avant de nous intéresser au calcul concret du PageRank et à deux détails intervenant dans le calcul, sur les pages ne contenant pas de liens et l'importance du facteur d'atténuation.

1.1 Présentation

1.1.1 Problématique

La structure de l'Internet – un ensemble de milliards de pages reliées par des liens hypertexte – en fait un espace dans lequel il est difficile de trouver ce que l'on cherche. De fait, le besoin d'outils adaptés s'est ressenti très vite, et dès 1994 les premiers moteurs de recherche faisaient leur apparition, avec par exemple le WWW (World Wide Web Worm, [McBryan, 1994]).

Toutefois, ces premiers moteurs n'étaient pas adaptés à la croissance exponentielle de l'Internet et à l'indexation de milliards de pages, et il a fallu trouver des solutions plus efficaces et produisant des résultats de meilleure qualité pour les recherches. La plupart des nouvelles approches, basées sur les techniques classiques de recherche d'information (Information Retrieval (IR) en anglais), ne prenaient en compte que le contenu des pages web et souffaient par conséquent de problèmes de *web spamming*, où certaines pages contiennent beaucoup de mots ou d'expressions populaires placés au bon endroit afin de faire monter le score de la page et tromper ainsi le moteur de recherche, dans des perspectives plus ou moins légales. Il a donc fallu trouver des méthodes d'indexation et de classement plus solides, prenant en compte la topologie du web.

1.1.2 Le PageRank

L'algorithme du PageRank, développé par Larry Page à l'Université de Stanford et présenté pour la première fois dans [Brin et Page, 1998], est un des facteurs intervenant dans la notation des pages web par le moteur de recherche Google.

Il faut noter que la valeur du PageRank fournie par aux utilisateurs Google, comprise entre 0 et 10, correspond à la valeur calculée par l'algorithme, comprise entre 0 et 1, à laquelle on a appliqué un changement d'échelle vers une échelle

logarithmique. De plus, l'algorithme utilisé par Google est une version améliorée (et privée) de celui présenté ici, prenant entre autres en compte la pertinence des liens entrant et sortant d'une page vis à vis d'une recherche pour noter la pertinence de cette page vis à vis de cette recherche.

Cet algorithme, inspiré des algorithmes se basant sur le nombre de citations pointant vers un article pour estimer son importance et sa qualité, se base uniquement sur la structure topologique de l'internet, c'est à dire sur les liens entre les différentes pages plutôt que sur les pages elle-mêmes : le PageRank ne s'intéresse pas à la pertinence des pages, mais à leur *autorité*.

1.2 Calcul

1.2.1 Algorithme de base

L'idée générale du calcul du PageRank est de mesurer l'autorité d'une page comme fonction de l'autorité des pages pointant vers cette page, et du nombre de lien sortant de chacune de ces pages.

Formellement, en notant x_p le PageRank de la page p , h_p le nombre de liens sortant de cette page, I_p l'ensemble des pages pointant vers p et d un facteur d'atténuation, on a :

$$x_p = d \sum_{q \in I_p} \frac{x_q}{h_q} + (1 - d) \quad (1)$$

Plus généralement, on introduit le vecteur \vec{x} des x_p , la matrice de transition $W = (w_{i,j})$ avec $w_{i,j} = 1/h_j$ si il existe un lien de la page j à la page i et $w_{i,j} = 0$ sinon et $\vec{\mathbb{1}}$ le vecteur dont toutes les composantes valent 1. Alors :

$$\vec{x} = dW\vec{x} + (1 - d)\vec{\mathbb{1}} \quad (2)$$

L'algorithme itératif suivant, qui n'est autre que l'algorithme de Jacobi, proposé dans [Brin et Page, 1998], permet de calculer le PageRank :

$$\vec{x}(t) = dW\vec{x}(t-1) + (1 - d)\vec{\mathbb{1}} \quad (3)$$

Pour $0 \leq d < 1$, ce système converge vers le vecteur \vec{x} défini précédemment.

1.2.2 Problème des pages sans liens

La matrice de transition W est presque une matrice stochastique (une matrice à coefficients positifs où la somme de chaque colonne est égale à 1), à part pour les colonnes liées aux pages ne contenant aucun lien, dites *dangling pages*.

En effet, pour une telle page j , il n'y a pas de i tel qu'il existe un lien de j à i , donc tous les $w_{i,j}$ sont nuls, et la somme de la colonne est 0 ; au contraire, si j n'est pas une telle page, pour les h_j pages pointées par la page j , il y a un coefficient de W égal à $1/h_j$, donc la somme de la colonne j vaut 1.

Les résultats de la deuxième partie s'appuyant sur le fait que la matrice W soit une matrice stochastique, il faut donc transformer W . Pour cela, deux techniques sont en général utilisées.

La première est d'introduire une nouvelle page, pointant sur elle-même et sur laquelle pointent toutes les *dangling pages*. Ainsi, plus aucune colonne de la matrice de transition \widetilde{W} de ce graphe étendu ne contient de 0, et la matrice est stochastique.

La seconde, proposée dans [Brin *et al.*, 1999], consiste à considérer les *dangling pages* comme des pages pointant sur toutes les autres, technique justifiée par certaines propriétés des processus stochastiques.

Dans les deux cas, on ramène aisément le vecteur \widetilde{x} calculé par la méthode de calcul du PageRank avec la matrice \widetilde{W} obtenue au vecteur \overline{x} initial, dans le premier cas en effectuant quelques opérations algébriques sur \widetilde{x} , dans le second cas en modifiant l'équation de base.

1.2.3 Importance du facteur d'atténuation

La valeur du facteur d'atténuation joue un rôle non négligeable dans le calcul du PageRank. Tout d'abord, un facteur d'atténuation nul entraînera que tous les PageRanks seront égaux à 1. D'autre part, un facteur d'atténuation égal à 1, le calcul peut ne pas converger, tout en restant borné ; et on note que dans ce cas de nombreuses pages se retrouvent avec un PageRank nul. Cette constatation peut s'étendre plus généralement pour des valeurs de d proches de 1.

On introduit la notion de pages *essentielles*. Ces pages sont celles qui appartiennent à des sous graphes dont on ne peut sortir, tandis qu'une page inessentielle est une page pour laquelle il existe un chemin permettant de partir de la page sans pouvoir y revenir (à partir de la page p , il existe une page p' telle qu'on puisse aller de p à p' mais pas de p' à p).

La théorie des chaînes de Markov nous donne alors que le PageRank des pages inessentielles tend vers 0 quand d tend vers 1. En pratique, cela donne que les communautés ayant pas ou peu de liens vers l'extérieur seront favorisées.

2 Propriétés du PageRank

Dans cette section, nous nous intéressons à plusieurs propriétés du PageRank : nous introduisons d'abord les notions d'énergie et de balance d'énergie et étudions les communautés et leurs interactions, puis nous nous intéressons à deux résultats de calculabilité et de complexité sur le PageRank.

2.1 Balance d'énergie et implications

2.1.1 Balance d'énergie

Afin de quantifier l'autorité d'une communauté donnée (un site internet, un ensemble de pages de chercheurs sur un sujet commun...), dont l'ensemble des pages est noté I , on introduit la notion d'énergie :

Définition 1. L'énergie de la communauté I , notée E_I , est égale à la somme des PageRanks des pages de la communauté. Formellement :

$$E_I = \sum_{p \in I} x_p$$

On note respectivement $out(I)$, $in(I)$ et $dp(I)$ les ensemble des pages de I pointant vers l'extérieur, l'ensemble des pages pointant vers des pages de I et l'ensemble des *dangling pages* de I . On définit intuitivement E_I^{out} , E_I^{in} et E_I^{dp} comme les énergies respectivement allant de I vers l'extérieur, de l'extérieur vers I et perdue depuis I . On note enfin $f_p(I)$ le ratio du nombre de pages pointées par p qui sont dans I sur le nombre total de pages pointées par p .

Définition 2. Balance d'énergie [Bianchini *et al.*, 2005] donne les trois expressions suivantes pour ces énergies, expressions qualifiées d'*équations de balance d'énergie* :

$$\begin{aligned} E_I^{in} &= \frac{d}{1-d} \sum_{p \in in(I)} f_p(I) x_p \\ E_I^{out} &= \frac{d}{1-d} \sum_{p \in out(I)} (1 - f_p(I)) x_p \\ E_I^{dp} &= \frac{d}{1-d} \sum_{p \in dp(I)} x_p \end{aligned}$$

En notant $|I|$ le nombre de pages de I , on a le résultat suivant reliant les différentes énergies liées à I :

Théorème 3.

$$E_I = |I| + E_I^{in} - E_I^{out} - E_I^{dp}$$

On note en particulier que l'énergie d'une communauté a pour valeur de base le nombre I de pages de la communauté.

2.1.2 Conséquences sur la visibilité des pages

D'un point de vue pratique, nous pouvons tirer quelques conclusions basées sur cette quantification de l'autorité d'une communauté.

- Étant donné que le facteur $|I|$ est l'énergie de base d'une communauté, il vaut mieux répartir le contenu sur plusieurs pages que de tout rassembler en une seule.
- Cependant, on a l'inégalité $E_I \leq |I| + E_I^{in}$: les petites communautés vers lesquelles il y a peu de liens ne pourront pas avoir de hauts PageRanks.
- Les *dangling pages* sont à l'origine d'une perte d'énergie pour la communauté, qui est d'autant plus faible que les pages pointant vers ces *dangling pages* ont un score peu élevé et/ou pointent vers beaucoup d'autres pages de la communauté.
- Les liens vers l'extérieur sont eux aussi à l'origine d'une perte d'énergie pour la communauté, cette perte est d'autant plus grande que ces pages ont un PageRank élevé, mais elle est atténuée si ces pages pointent également vers un grand nombre de pages de la communauté, et plus généralement si il y a peu de pages de la communauté pointant vers l'extérieur, par rapport au nombre de liens dans la communauté.
- S'il est possible de tirer parti des liens pointant de plusieurs communautés vers une communauté cible que l'on veut promouvoir, les modifications réalisées par le PageRank dépendent très fortement de l'énergie de ces communautés originie : des petites communautés à faible énergie ne pourront pas modifier grandement le PageRank. Cette propriété fait du PageRank un algorithme robuste face au *spamming*¹.

¹Toutefois, le PageRank reste vulnérable aux opérations de spamming de masse, ou de nombreuses pages de nombreuses communautés différentes se mettent à pointer vers une page en incluant des mots clefs communs. De telles opérations sont qualifiées, dans le cadre du moteur de recherche Google, de Google bombing

2.2 Calculabilité et complexité

2.2.1 Calcul sur un graphe évoluant dans le temps

L'algorithme du PageRank est initialement conçu pour les systèmes statiques, où W ne change pas dans le temps. Toutefois, le résultat suivant permet d'étendre le PageRank aux systèmes dynamiques :

Théorème 4. *Le PageRank peut être calculé en utilisant une matrice de transition $W = W(t)$ évoluant dans le temps, par exemple mise à jour au fur et à mesure de l'exploration du web par les robots indexeurs.*

Pour remédier au fait que le nombre de pages (et donc la taille du vecteur \vec{x} , on prend le vecteur \vec{x} dans \mathbb{R}^∞ , et on initialise le PageRank des pages non découvertes à 1 on suppose alors que ces pages ont un seul lien vers elles même et aucun lien venant de l'extérieur). Le reste de l'algorithme est le même que dans la version statique.

2.2.2 Complexité temporelle

Étant donné que l'Internet est maintenant composé de plusieurs milliards de pages, l'optimisation des algorithmes utilisés pour l'indexation du web est essentielle. Ainsi, utiliser des méthodes telles que le pivot de Gauss pour la résolution du système linéaire (2) dont dépend \vec{x} se fait en $\mathcal{O}(N^3)$ opérations (en virgule flottante, ou *flops*) où N est le nombre de pages, ce qui est plus qu'indésirable avec des N à l'échelle de l'Internet.

L'algorithme (3), au contraire, s'exécute en $\mathcal{O}(m \cdot |H|)$ flops, avec $|H|$ le nombre de liens et m le nombre d'itérations. De plus, les résultats expérimentaux montrant que $m \ll N$, et que par nature du web $|H| \ll N^2$ (chaque page ne pointe que vers un petit nombre d'autres pages, à l'échelle de l'Internet), d'où $m \cdot |H| \ll N^3$: cet algorithme est bien plus efficace que le pivot de Gauss. En particulier, si on note ϵ l'erreur maximale commise en valeur absolue dans le calcul du PageRank, on a $m < \log(1/\epsilon)$. D'où :

Théorème 5. *Le PageRank peut être calculé par un algorithme en temps*

$$O(|H| \cdot \log(1/\epsilon))$$

Il faut enfin noter que d'autres algorithmes de résolution de systèmes linéaires peuvent être utilisés ici, certains proposant des complexités théoriques encore plus faibles. Cependant, une grosse partie de l'implémentation de l'algorithme dépendra

des structures de données, et en particulier de celles utilisées pour représenter W , matrice carrée de plusieurs milliards de cases de large. Ainsi, dans la pratique, l'algorithme de Jacobi (3) est parfois préféré, car il permet d'utiliser directement des représentations compactes de chacune des colonnes (telles que celles produites par les robots d'indexation, qui donnent la liste des pages pointées par une page donnée), alors que certains autres algorithmes se servent directement des lignes de W et non pas de ses colonnes.

Conclusion

Nous avons présenté l'algorithme du PageRank et explicité son calcul, puis nous avons présenté plusieurs résultats importants aussi bien en théorie qu'en pratique sur les scores qu'il produit et sur son calcul. Ainsi, les notions d'*énergie* et de *balance d'énergie* permettent de comprendre les interactions entre les communautés et d'optimiser dans la pratique la structure des sites webs et des communautés. De plus, les résultats de calculabilité et de complexité montrent que cet algorithme est adapté à l'indexation de l'Internet, puisqu'il peut d'une part être lancé au fur et à mesure de l'indexation et pas uniquement après récupération intégrale des données et construction de la matrice W entière, et qu'on peut d'autre part le calculer en un temps raisonnable, même à l'échelle de l'Internet.

Références

- [Bianchini *et al.*, 2005] Monica BIANCHINI, Marco GORI et Franco SCARSELLI (2005). Inside PageRank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128.
- [Brin et Page, 1998] Sergey BRIN et Lawrence PAGE (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 1(1–7):107–117.
- [Brin *et al.*, 1999] Sergey BRIN, Lawrence PAGE, Rajeev MOTWANI et Terry WINOGRAD (1999). The PageRank Citation Ranking : Bringing Order to the Web.
- [McBryan, 1994] Oliver A. MCBRYAN (1994). GENVL and WWW : Tools for Taming the Web. *In First International Conference on the World Wide Web*.