Comment utiliser le nombre de liens entrant pour estimer un ordre de grandeur du PageRank ?

Retour au sommaire de la lettre

Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Le PageRank et la façon dont il est calculé par Google (et les autres moteurs) reste souvent une énigme insondable pour de nombreux référenceurs. Pourtant, plusieurs chercheurs ont tenté de comprendre la façon dont il était établi, en utilisant des méthodes statistiques et probabilistes. Il semblerait notamment qu'il soit possible d'établir une relation d'ordre de grandeur entre le PageRank et le nombre de liens entrants d'une page. Cette piste d'exploration peut notamment nous aider à mieux répartir les liens internes d'un site pour obtenir une meilleure optimisation. Explications...

Les scores de type "PageRank" sont largement utilisés dans les moteurs de recherche, essentiellement parce qu'ils sont de bons indicateurs de la popularité d'une page, mais aussi parce qu'ils sont "relativement" faciles à calculer. Mais il faut bien comprendre ce que signifie ce "relativement"... Historiquement, la formule mathématique qui sert à calculer ce score s'est avérée plus "calculable" que d'autres méthodes, mais elle fait appel à des calculs sur des matrices carrées dont le nombre de lignes et de colonnes se chiffre en ... milliards.

Dans les premières années d'existence du moteur Google, le calcul du PageRank prenait plusieurs jours. Plusieurs astuces ont permis d'accélérer le calcul, comme le recours à l'extrapolation quadratique, et le recours à des PR approchés en calculant un PR pour des nouveaux liens à partir des valeurs existantes et sur une matrice réduite aux liens les plus proches (ex : stratégie du BlockRank, voir articles précédents à ce sujet dans cette lettre). Mais calculer les PageRanks pour tous les noeuds d'un graphe de liens aussi vaste que le World Wide Web reste un défi mathématique et informatique.

Les propriétés mathématiques du PageRank ont été (et sont toujours) largement étudiées. Au cours de ces recherches, plusieurs équipes scientfiques ont été intriguées par un résultat statistique : les valeurs des PR des pages et le nombre de liens entrants sont distribués d'une manière très proche. Nous allons voir dans cet article ce que signifie ce résultat et pourquoi il est surprenant. Et nous allons en tirer un indicateur très opérationnel pour le SEO, notamment pour le PageRank Modeling.

Le nombre de liens entrants : un médiocre indicateur de la popularité d'un site ?

Si l'on souhaite calculer un score de popularité pour une page web, il vient tout de suite à l'idée de prendre pour indicateur le nombre de liens entrants : plus une page reçoit de liens, plus elle est populaire.

Mais cet indicateur a un inconvénient majeur : il ne tient pas compte de l'importance relative des pages, car en comptant les liens entrants, on part du principe que tous les liens "se valent", ce qui ne colle pas à la réalité. Un lien en provenance de la page d'accueil du New York Times ne doit pas être équivalent au lien reçu grâce au billet publié sur le blog de votre belle-soeur !

Ce biais est censé être éliminé dans l'algorithme du PageRank : le PR transmis dépend de la popularité de la page qui fait le lien. Evidemment, à chaque étape du calcul de popularité, le score de PR est réévalué, ainsi que les PR transmis aux pages liées : l'algorithme est itératif, et au bout d'une vingtaine ou une trentaine de calculs successifs (pour la formule initiale du PageRank publiée par L. Page et S. Brin) on aboutit à une valeur qui ne change pratiquement plus.

Bref on a théoriquement grâce au PageRank, un indicateur beaucoup plus fiable de la popularité d'une page sur le web que le nombre de liens entrants.

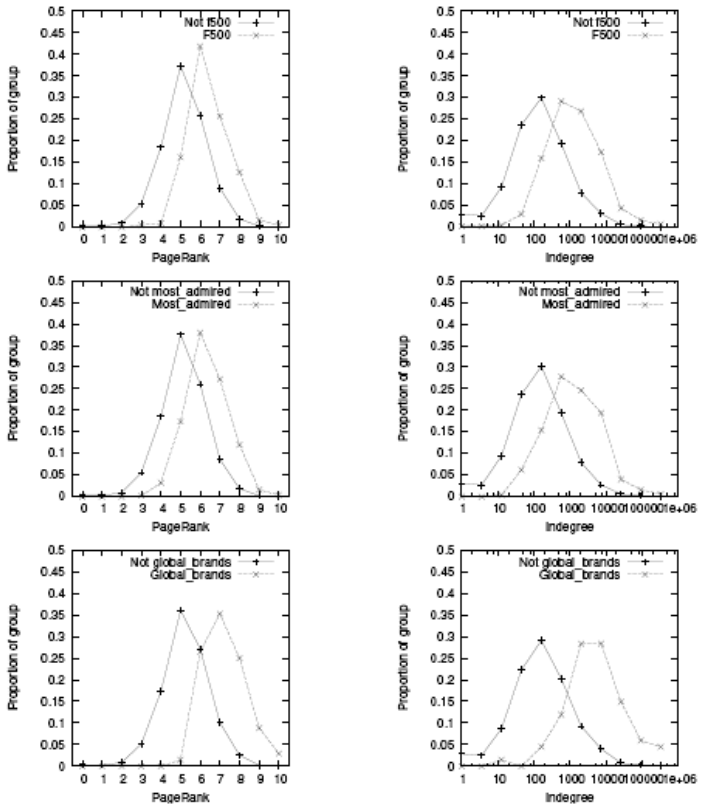
Sauf que certains ont cherché à vérifier, avec des méthodes statistiques, l'efficacité des deux indicateurs en tant qu'estimateurs de la popularité d'une page. Et les résultats ont été franchement contre-intuitifs !

Nombre de liens entrants et PageRank : des indicateurs équivalents pour les sites des entreprises du classement Fortune 500 !

Trois chercheurs australiens (Trystan Upstill, Nick Craswell et David Hawking, de l'Unversité Nationale Australienne de Canberra) ont fait une expérience simple : comparer les PageRanks des pages d'accueil de sites du classement Fortune 500, avec le logarithme (à base 10) du nombre de liens entrants sur ces pages.

Le PageRank observé ici est le PageRank indiqué par la Toolbar (qui est, rappelons-le, la seule indication "officielle" fournie par Google de valeur du PR d'une page). On sait que les valeurs de la toolbar sont approximativement logarithmiques, même si personne ne connait la base utilisée.

Les conclusions de cette étude sont sans appel : pour ces sites, la corrélation est extrêmement élevée, et on aurait pu substituer à la valeur du PR le logarithme du nombre de liens entrants et obtenir la même hiérarchie de valeurs (mais pas exactement les mêmes valeurs) !



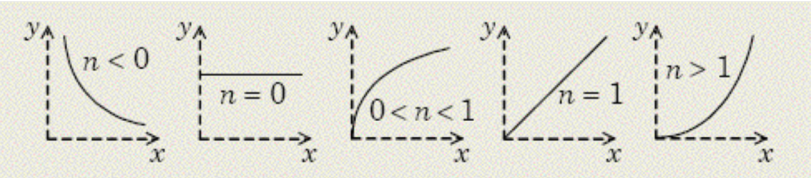
Comparaison entre les valeurs du PR et du nombre de liens entrant (Indegree) pour les sites des entreprises du classement Fortune 500 : les courbes, très similaires, semblent indiquer une corrélation forte entre les deux critères.

La distribution des PageRank et le nombre de liens entrants obéissent à une loi de puissance de même exposant

Cette similitude est elle limitée à cet échantillon de sites un peu particuliers ? Non, pas du tout, elle peut être généralisée.

D'autres études statistiques plus poussées ont démontré que les valeurs des PR obéissent à des lois statistiques communes, dès lors que l'on s'intéresse à des pages ayant un PR élevé. Il a été prouvé que la distribution des nombres de liens entrants suit une loi de puissance. En fait, le nombre de liens entrants n'est autre que le degré entrant du graphe des liens, et on peut donc réutiliser tous les résultats mathématiques de la théorie des graphes à propos du degré entrant.

Une loi de puissance est une relation mathématique entre deux quantités, représentée par une équation du genre : **y = a*x^k** L'exposant k est appelé parfois "degré" de la loi de puissance.

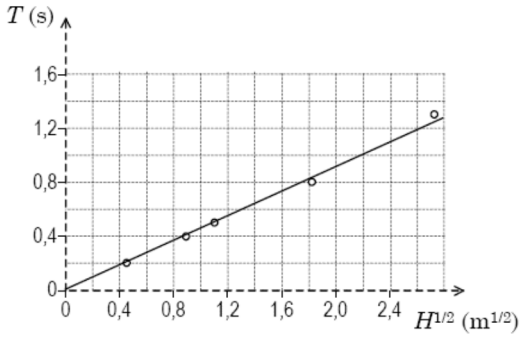


Différentes formes de courbes en fonction de l'exposant de la loi de puissance. Dans le cas étudié, n est > 1 et la courbe ressemble à une exponentielle comme la dernière courbe à droite...

Pour déterminer l'exposant de la loi de puissance de manière empirique, il est d'usage de travailler sur les logarithmes des valeurs y et x, ce qui transforme la loi de puissance en une droite affine.

$$\log(y) = k \log(x) + \log(a)$$

Si l'on souhaite détermine l'exposant et la valeur de a, il suffit de reporter sur un graphe les valeurs expérimentales trouvées, et de déterminer la droite affine la plus proche des points de ce graphe.



Que signifie cette loi de puissance, dans le cas des liens entrants ?

Dire que la distribution des nombres de liens entrants obéit à une loi de puissance de degré k signifie, pour être exact, que la probabilité pour que le nombre de liens entrant soit supérieur à un nombre x est proportionnelle à x^{-k} .

Le résultat de Gopal Pandurangan : PageRank et Indegree suivent une loi de puissance de degré quasi identique !

En 2002, un chercheur tente d'évaluer l'exposant de la loi de puissance décrivant la distribution de l'*outdegree* (nombre de liens sortants), de l'*indegree* (nombre de liens sortants) et du PageRank calculé d'après la formule initiale de Page et Brin. Gopal Panduragan trouve un exposant de 2,7 pour les liens sortants, et le même coefficient pour l'indegree et le PR : 2.1 Il remarque que pour les 20% des pages ayant le PR le plus élevé, les courbes sont tellement parfaitement identiques que l'hypothèse d'une corrélation et même relation mathématique entre les deux grandeurs semble plausible.

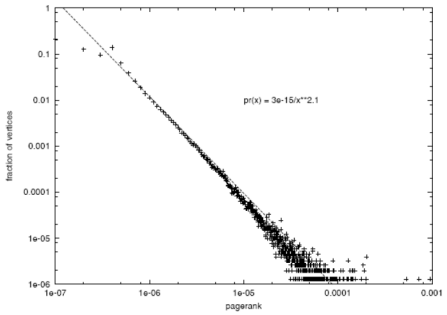


Figure 4. Log-log plot of the PageRank distribution of the WT10g corpus. The slope is close to 2.1. Note that the plot looks much sharper than the corresponding plot for the Brown web. Also, the tapering at the top is much less pronounced.

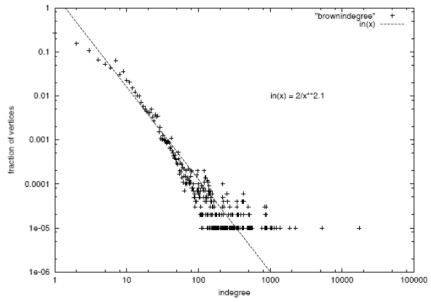


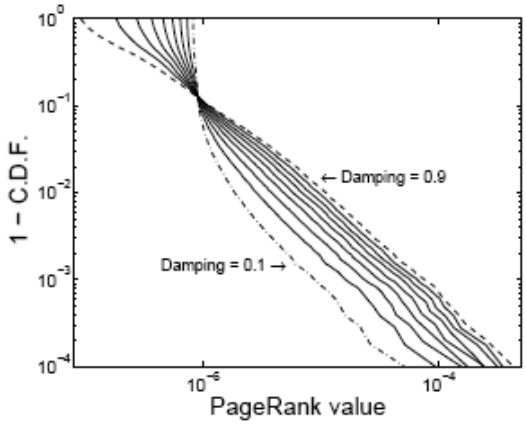
Figure 1. Log-log plot of the in-degree distribution of the Brown domain (*.brown.edu). The in-degree distribution follows a power law with exponent close to 2.1.

A gauche, la distribution des PageRanks, à droite la distribution des indegrees : les deux droites ont la même pente : 2,1, révélant une corrélation possible.

Des résultats confirmés par des études plus récentes

Cette étude statistique a été réalisée à nouveau depuis par différentes équipes de chercheurs et sur des échantillons différent, notamment par Debora Donato en 2004 (de l'Université de Rome), et par Santo Fortunato en 2005 (de l'Université de l'Indiana).

Becchetti et Castillo ont cherché ensuite à savoir si cette corrélation apparente était liée au facteur d'atténuation dans la formule du PageRank (fixé à 0,85 dans la formule initiale). Leurs conclusions ont été que le "damping factor" a bel et bien une influence, et que la corrélation spectaculaire observée provient du fait que le facteur utilisé dans la pratique est compris entre 0,85 et 0,90. Toutefois, les résultats du "top tail" restent toujours identiques, comme si en réalité la distribution était indépendante de la formule utilisée pour calculer le PageRank !



Distribution des valeurs du PageRank sur le WWW
pour différentes valeurs du facteur d'atténuation (damping factor)

Nelly Litvak a ensuite cherché en 2007 à expliquer l'origine de ces similitudes statistiques. L'*indegree* étant réutilisé dans la formule du PageRank, il n'était pas surprenant de trouver une corrélation, toutefois il restait à expliquer pourquoi les autres facteurs de la formule devenaient sans effet lorsque le PR devenait suffisamment grand. Mlle Litvak est parvenue à une démonstration mathématique et non empirique de la convergence des probabilités pour les deux indicateurs.

Quelles conséquences pour le SEO ?

Maintenant que l'on sait que PageRank et nombre de liens entrants obéissent à la même loi de puissance pour des valeurs suffisamment grandes de ces indicateurs, quelles informations peut-on en tirer pour le SEO ?

D'abord, il faut comprendre que ce résultat ne signifie pas que les deux indicateurs sont équivalents : jusque que la hiérarchie des pages créées en partant des valeurs de ces deux indicateurs sont similaires, et que c'est d'autant plus vrai que l'on considère des valeurs élevées de ces indicateurs.

Néanmoins, cela signifie qu'il est possible d'estimer grossièrement le PageRank d'une page à partir du nombre de liens entrants. Or le plus souvent, ce que l'on a besoin d'évaluer à des fins de SEO, ce n'est pas une valeur précise du PageRank, mais une hiérarchie d'ordre de grandeurs ! Donc cette estimation peut se révéler fort utile dans la pratique.

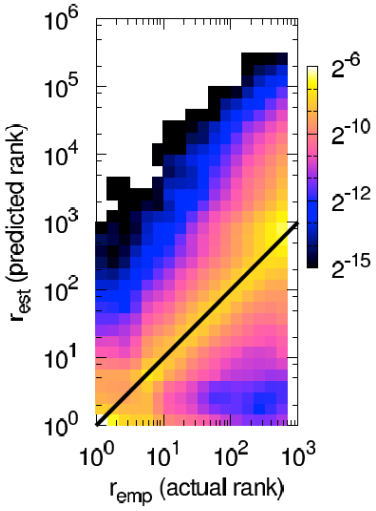
Dans ce cas, Santo Fortunato a établi que l'on pouvait obtenir une valeur assez précise de l'ordre de grandeur du PageRank en utilisant cette formule, calculée à partir des nombres de liens entrants seulement :

$$\overline{p}(\mathbf{k}) = \frac{q}{N} + \frac{1-q}{N}\frac{k_{in}}{\langle k_{in}\rangle},$$

dans laquelle :
q = le facteur d'atténuation utilisé dans le PageRank (classiquement 0,85).
N = le nombre de pages dans l'index : disons 30 milliards de page après caffeine (cela n'influe que sur la valeur absolue des PR, pas sur la hiérarchie des PR).
$<k_{in}>$ = la valeur moyenne du nombre de liens entrants sur une page du Web. Il n'existe pas de consensus sur ce nombre, qui change selon l'échantillon étudié, mais l'ordre de grandeur est entre 8 et 25, et surement pas 50 ou 100 ou 1000 !

L'intérêt de cette formule est qu'elle permet de faire des évaluations assez justes dans trois cas :

- soit parce que l'on considère un nombre suffisant de pages (plus de 200) pour que les erreurs, aussi soient grandes soient-elles, soient statistiquement acceptables, et permettent des comparaisons justes entre groupes de pages par exemple ;
- soit parce que les k_{in} sont suffisamment élevés pour que la formule donne une estimation juste (k_{in} > 1000) ;
- soit parce que l'on compare des kin d'ordres de grandeur différents : $k_{in}1$ = 1000 et $k_{in}2$=100.

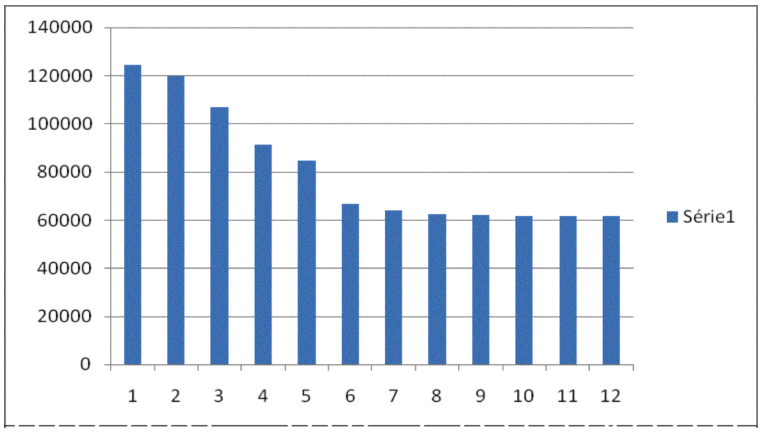


"thermogramme" de l'exactitude des résutats obtenus avec la formule : la plupart des résultats estimés sont proches de la valeur effective du PageRank, même si ces certaines valeurs s'en éloignent fortement. La ligne noire indique les cas où le PageRank estimé est égal au PageRank réel.

Enfin, cette formule est intéressante pour tester l'efficacité des structures hypertextes pour un site web donné, à des fins de PR modeling par exemple. Dans ce cas, on fait abstraction des

liens externes. Par ailleurs, la distribution des PR entre les pages internes est souvent très uniforme, et le PageRank interne d'une page est caractéristique d'un template et ou d'une catégorie/univers (ou d'un croisement des deux). Dès que l'on travaille sur des sites comportant des centaines de milliers ou des millions de pages, le nombre de liens internes entrants devient un indicateur très fiable et très utile, extrêmement corrélé avec le PageRank interne du site.

Pour évaluer la hiérarchie des PageRanks internes au sein d'un site, et la manière dont le "jus" est distribué, il suffit généralement de jeter un oeil au tableau des liens internes dans GWT pour s'en faire une idée juste !



Répartition de l'indegree donné par GWT sur une douzaine de pages.
Les pages numérotées 8 à 12 ont sensiblement le même nombre de liens entrant, tandis qu'une hiérarchie claire se dégage entre les pages numérotées 1 à 6.

Euh ? Et le Mozrank alors ?

Le Mozrank (valeur donnée par l'outil Open Site Explorer de SEOMoz) est une valeur calculée à partir des données sur la matrice des liens du World Wide Web, en utilisant une formule de type PageRank simplifiée et une pincée de données issues de la toolbar de Google pour "calibrer" et "redresser" les résultats.

Bien que les liens de la base Linkscape ne soient qu'un sous-ensemble réduit de la vraie matrice des liens du WWW, et que la formule du MozRank soit essentiellement empirique, ces résultats expliquent aussi pourquoi ce genre d'approche fonctionne plutôt bien : dès que l'on parle de pages disposant d'un grand nombre de liens entrants, certaines formules ont tendance à converger et à dessiner la même hiérarchie des pages.

Vers une quantification du linkjuice en SEO ?

Si l'on consulte la plupart des blogs et forums du SEO, on est frappés du caractère souvent purement empirique des approches, qui se résume souvent à essayer différentes méthodes et de ne retenir que celles qui fonctionnent. Evidemment, l'exercice à ses limites. C'est d'autant plus vrai pour l'optimisation des critères liés à la popularité par les liens : en général, les SEO renoncent à chercher une approche quantitative, soit parce qu'ils ne comprennent pas l'algorithme ou la manière dont le score se combine avec d'autres critères, soit parce qu'ils pensent que le PageRank ne peut pas être calculé en dehors de Google (ce qui est bien sûr vrai, mais on peut tout de même essayer d'estimer le potentiel de linkjuice d'une page et en avoir une idée précise), soit, et c'est bien sûr un obstacle, parce qu'ils ne savent pas ou ne peuvent pas le calculer.

L'approximation par le nombre de liens entrant ouvre une voie nouvelle pour une approche quantitative du SEO. Ce n'est pas une voie facile, car elle peut conduire à de nombreuses erreurs pour quelqu'un qui n'a pas de notions de statistiques ou de probabilités, mais elle est clairement prometteuse pour tout ceux qui voudront se donner la peine de l'explorer.

Bibliographie

Google page rank and beyond
Par Amy N. Langville,Carl Dean Meyer
Princeton University Press

Deeper Inside PageRank.
http://meyer.math.ncsu.edu/Meyer/PS_Files/DeeperInsidePR.pdf
Amy N. Langville. † and Carl D. Meyer. 2004

Web Stats - Site de Sylvain Peyronnet
http://sylvain.berbiqui.org/web-statistics-fr/index.htm

Predicting Fame and Fortune: PageRank or Indegree?
Trystan Upstill, Nick Craswell and David Hawking
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.9918&rep=rep1&type=pdf

Using PageRank to Characterize Web Structure
Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal
http://www.internetmathematics.org/volumes/3/1/Pandurangan.pdf

Large scale properties of the Webgraph.
http://www.dis.uniroma1.it/~cosin/publications/masterjournal.pdf
D. Donato, L. Laura, S. Leonardi, and S. Millozi.
Eur. Phys. J., 38:239–243, 2004.

The egalitarian effect of search engines.
http://arxiv.org/pdf/cs/0511005v2
S. Fortunato, A. Flammini, F. Menczer, and A. Vespignani.
Technical Report 0511005, arXiv/cs, 2005.

The distribution of PageRank follows a power-law only for particular values of the damping factor.
http://www2006.org/programme/files/pdf/p126-poster.pdf
L. Becchetti and C. Castillo.
In Proceedings of the 15th international conference on World Wide Web, pages 941–942. ACM Press, New York, 2006.

In-Degree and PageRank of Web pages: Why do they follow similar power laws?
N. Litvak, W.R.W. Scheinhardt and Y. Volkovich
http://wwwhome.math.utwente.nl/~litvakn/IntMath07.pdf

How to make the top ten: Approximating PageRank from in-degree
http://arxiv.org/pdf/cs/0511016v1
Santo Fortunato, Marian Boguna, Alessandro Flammini, Filippo Menczer (2005)

SEOMoz : Linkscape et le MozRank
http://www.opensiteexplorer.org/

Philippe Yonnet, ***Global SEO Strategist, WEB DMUK (Londres) – Easyroommate / Vivastreet***

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2010/10/comment-utiliser-le-nombre-de-liens.html