Le PageRank en 2009 : mythe ou réalité ?

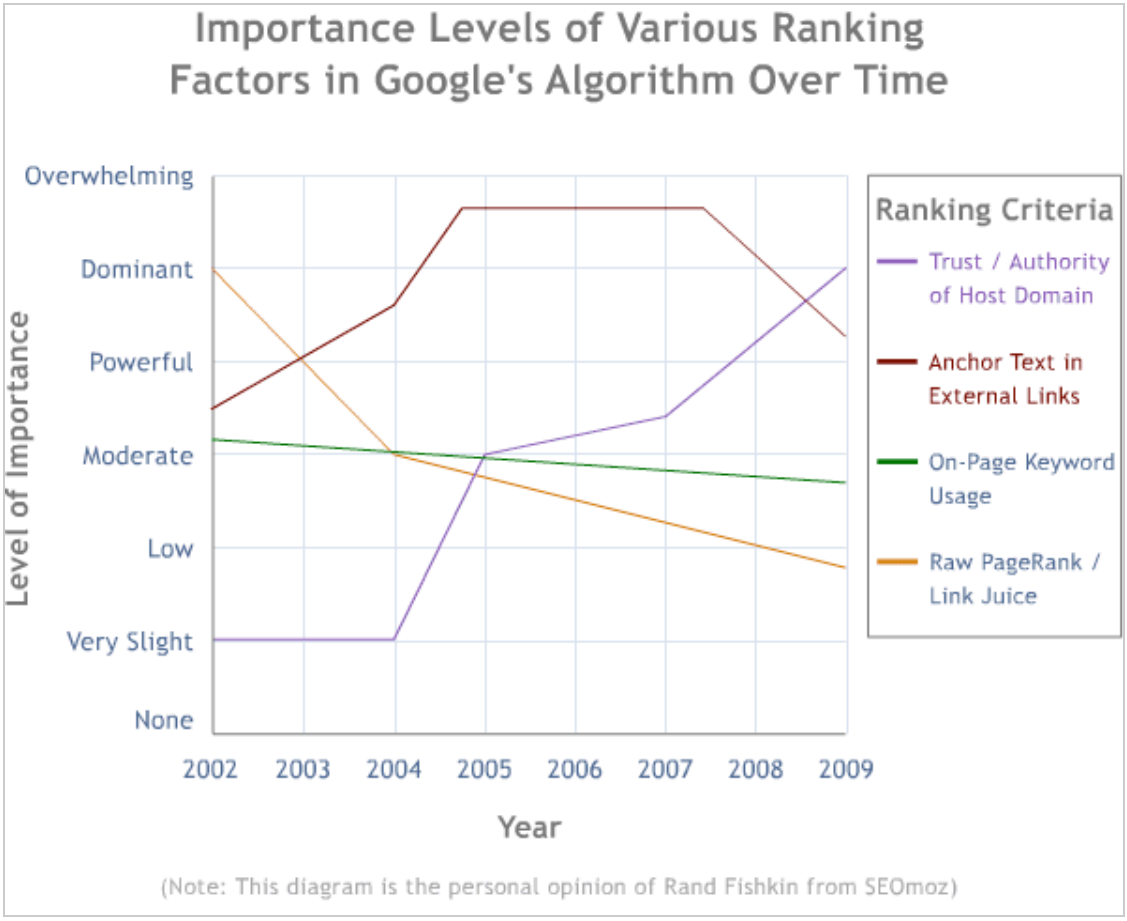
Retour au sommaire de la lettre

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

Depuis l'avénement de Google, à la fin des années 90, on entend parler de PageRank, indice de popularité calculé à partir de l'analyse quantitative et qualitative des liens pointant sur une page. Certains, aujourd'hui, pensent cependant que ce critère est désuet. D'autres estiment, en revanche, qu'il est toujours très important dans l'algorithme de pertinence du moteur de recherche leader. Cet article a pour but de faire le point sur ce facteur et de décrire les différentes avancées et évolutions du PageRank depuis dix ans. Et Dieu sait s'il y a à dire sur ce sujet parfois polémique mais toujours passionnant...

La plupart des spécialistes des outils de recherche s'accordent sur un point : le PageRank (PR) est un excellent outil pour produire une mesure universelle et facile à calculer de l'importance d'une page sur le web. Mais ce consensus autour du PR n'est pas universellement partagé, notamment par les webmasters et les référenceurs. Son rôle exact et son importance réelle dans l'algorithme est le sujet de débats interminables et passionnés, et l'algorithme a été, depuis son apparition il y a plus de 10 ans, tour à tour idolâtré puis voué aux gémonies.

Cette tendance lourde a été illustrée récemment par un billet du très influent Rand Fishkin de SEOMoz (http://www.seomoz.org/blog/how-googles-rankings-algorithm-has-changed-over-time-), décrivant son évaluation personnelle de l'évolution du poids des facteurs dans l'algorithme de Google. La courbe en orange qui ne cesse de décliner est celle du PageRank et de ce qu'il appelle le "jus de lien".



Mais qu'en est-il exactement ? Le PageRank est il toujours utilisé par Google ? La petite barre verte n'est-elle devenue qu'un leurre pour amuser les webmasters, un simple argument marketing ? Voici quelques éléments d'information, qui, nous l'espérons, vous permettront de vous faire une idée sur ces questions.

C'est quoi exactement le PageRank ?

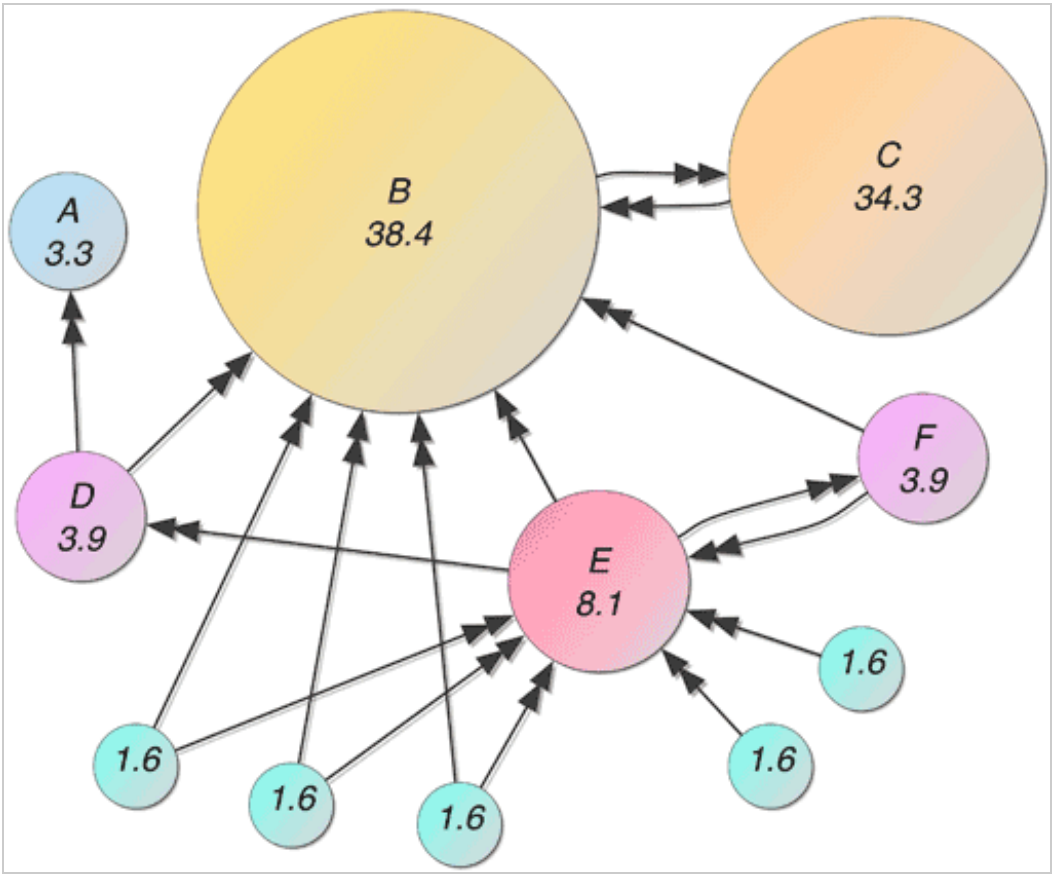
Le PageRank est un algorithme calculant une valeur pour une page web en effectuant des calculs sur les graphes de liens hypertextes reliant les pages entre elles sur internet. La dénomination "PageRank" provient non pas du mot "page", mais du nom de son fondateur : Larry Page. Larry Page a développé cet algorithme dans le cadre de ses travaux de recherche commencés dès 1995 au sein de l'Université de Stanford. Il a été rejoint rapidement par Sergey Brin, avec lequel il a pu développer un prototype de moteur de recherche, devenu Google en 1998. Cette origine explique pourquoi le brevet du PageRank est au nom de l'Université de Stanford, Google disposant simplement du droit exclusif d'exploiter ce brevet. Depuis lors, Google continue d'avoir des relations privilégiées avec les équipes de recherche de Stanford.

> Voici comment Google présente le PageRank :
> *Le PageRank est un champion de la démocratie : il profite des innombrables liens du Web pour évaluer le contenu des pages Web - et leur pertinence vis-à-vis des requêtes exprimées. Le principe du PageRank est simple : tout lien pointant de la page A vers la page B est considéré comme un vote de la page A en faveur de la page B. Toutefois, Google ne limite pas son évaluation au nombre de "votes" (liens) reçus par la page ; il procède également à une analyse de la page qui contient le lien. Les liens présents dans des pages jugées importantes par Google ont plus de "poids", et contribuent ainsi à "élire" d'autres pages.*

En d'autres termes, le PageRank d'une page dépend du nombre de liens pointant vers cette page, mais aussi de la valeur de ces liens. La valeur d'un lien correspond au PageRank de la page qui contient ce lien, mais divisé par le nombre de liens contenus dans cette page. Le PageRank est donc une quantité définie récursivement et qui demande, pour être évaluée, d'effectuer un calcul itératif (en boucle).

Le PageRank peut également être interprété comme la probabilité qu'un surfeur aléatoire arrive sur une page donnée. (Brin et Page proposaient comme image des singes munis d'une souris cliquant n'importe où sur un écran affichant des pages web). Comme de temps en temps, le surfeur se lasse, il finit par ne plus cliquer mais par se téléporter dans une autre zone du web en entrant une autre url dans la barre d'adresse de son navigateur. Ces "téléportations" sont représentées dans l'algorithme par un coefficient d'atténuation (*damping factor*).

Globalement, la formule du PR a été conçue pour représenter des probabilités, et être convergente (l'algorithme converge vers une valeur fixe au bout d'un certain nombre d'itérations).



Exemple de calcul de PageRank entre un groupe de pages

Euh ! Un PageRank de 38,4 ? Commment c'est possible ?

Google ne donne pas de moyens pour connaître la **vraie** valeur du PageRank d'une page. La seule valeur connue est une indication donnée par la barre d'outils de Google, sous la forme d'une échelle de valeurs entières allant de 0 à 10. On sait par ailleurs que cette échelle est logarithmique, ce qui veut dire qu'entre un PR 5 et un PR 6, il y a peut-être dix fois plus de différence qu'entre un PR 4 et un PR 5 (on ne connait pas la base de cette échelle logarithmique). Les vraies valeurs du PR ne sont pas des valeurs entières, ne se situent pas entre 0 et 10, et peuvent donc ressembler à celles indiquées sur le graphe.

Le TBPR (le "*toolbar PageRank*") est donc très grossier, et ne permet pas de comparer les valeurs du PR de deux pages ayant le même TBPR, y compris dans des cas où la différence de valeur de PR est grande.

Qui plus est, le TBPR est depuis toujours stocké dans une base à part. A l'époque des *Google Dances*, c'est-à-dire avant que le calcul du PageRank soit effectué beaucoup plus fréquemment, cette base contenait une copie des PR calculés au cours du mois précédent. La base contenait déjà des erreurs, notamment des pages sans PR alors que le PR avait été calculé pour ces pages. Depuis l'apparition des "*rolling updates*" au cours de l'été 2003, un nouveau type de problème est apparu : la base des TBPR contient des valeurs qui correspondent à une photographie d'une époque révolue (il est arrivé qu'aucune mise à jour ne soit effectuée pendant plusieurs mois).

Le TBPR ne donne pas une valeur suffisamment précise, ni exacte (car le plus souvent obsolète) de la valeur d'une page. Malgré ces défauts, c'est le seul moyen que Google donne aux webmasters pour se faire une idée sur le PR de leurs pages...

A propos de l'origine du PageRank

Comme beaucoup d'inventions, le PageRank est avant tout le produit d'un contexte de découvertes scientifiques. A la fin des années 90, de nombreux chercheurs ont remarqué les analogies profondes qui existaient entre le problème de l'évaluation des articles scientifiques et l'évaluation des pages sur le web. En effet, on évalue les articles scientifiques par le nombre de citations qu'ils reçoivent dans les bibliographies d'autres articles, chaque citation pouvant elle même être pondérée en fonction de l'importance de l'article. Cela ne vous rappelle rien ?

Il s'avère que les algorithmes permettant de faire ces calculs avaient déjà identifiés dès les années 30 (certains chercheurs de Google reconnaissent d'ailleurs que le calcul du PageRank n'aurait pas été possible sans les travaux d'un Polytechnicien français : Gaston Julia. Ce qui a valu d'ailleurs à cet illustre mathématicien un "Google Doodle" à son effigie sur la home page de Google en 2004, *voir ci-contre*), puis avaient été fortement perfectionnés dans les années 60 avec l'avènement de l'informatique. Les premières méthodes opérationnelles automatiques de "notation" des articles scientifiques sont apparues dans les années 70.

Larry Page et Sergey Brin citent clairement dans leurs articles sur le PageRank les travaux datant des années 50 d'Eugene Garfield sur l'analyse des citations à l'Université de Pennsylvanie.

Un autre algorithme, inventé par Jon Kleinberg à la même période (1996/97 contre 1997/98 pour Google) tire son origine également d'une analogie avec une méthode inventée pour la bibliométrie (application de techniques statistiques ou mathématiques permettant d'analyser les citations dans des ensembles de références bibliographiques ; lorsqu'il s'agit de citations d'articles scientifiques, on parle aussi de "scientométrie") : "*Hyperlink-Induced Topic Search*" (HITS).

Les avantages du PageRank

Le PageRank : un critère facile à calculer ?

La facilité de calcul du PageRank est réelle, mais toute relative. L'algorithme est robuste, simple, et facile à programmer (en tout cas pour un développeur de métier). Calculer un PR sur quelques dizaines de millions de pages prend quelques minutes sur une machine de base actuelle, avec une implémentation naïve de l'algorithme, c'est-à-dire sans optimisations.

Dans la pratique, il s'agit de faire des calculs sur une matrice comportant autant de lignes et de colonnes que de pages dans l'index, soit quelques dizaines de milliards si l'on en croit les dires de Google sur la taille de leur index (et même mille milliards de liens! : http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html). Là, il est clair que c'est un peu plus compliqué :
"We've known it for a long time: the web is big. The first Google index in 1998 already had 26 million pages, and by 2000 the Google index reached the one billion mark. Over the last eight years, we've seen a lot of big numbers about how much content is really out there. Recently, even our search engineers stopped in awe about just ***how*** *big the web is these days -- when our systems that process links on the web to find new content hit a milestone: 1 trillion (as in 1,000,000,000,000) unique URLs on the web at once!"*

Google déclare ainsi aujourd'hui être en mesure de calculer la matrice des PageRanks sur mille milliards de liens plusieurs fois par jour, là où il lui fallait plusieurs jours dans ses débuts lorsque l'index, en 2000, ne comportait qu'1 milliard de pages ! Cette prouesse a été rendue possible par trois évolutions majeures.

Evolution n°1 : le développement de la capacité de calcul des machines

D'abord la puissance des machines a connu des progrès très importants. Il faut rappeler que Google a, dès l'origine, fait le choix de systèmes distribués sur un grand nombre de machines bon marché, plutôt que de passer par des stations de travail IBM ou Cray. Il s'avère que les architectures hautement distribuées et hautement parallèles sont particulièrement adaptées aux calculs matriciels du type de ceux nécessaires pour le calcul du PageRank.

La capacité de calcul de Google suit donc la même évolution que le PC de base acheté par l'internaute lambda. Les processeurs sont plus puissants, la capacité des disques durs augmente et les temps d'accès en lecture s'améliore, et surtout, la mémoire vive disponible de base sur les machines augmente, améliorant d'autant les capacités de calcul des *datacenters* de Google. On peut d'ailleurs prédire que le remplacement progressif des disques durs par des mémoires SSD de grande capacité va révolutionner le monde des outils de recherche : certains types de calculs et d'approches, éliminés jusqu'ici parce qu'impossibles à calculer à la volée, vont pouvoir être utilisés. Il est déjà certain que certains types de prétraitement de requêtes que l'on observe dans le fonctionnement actuel de Google n'auraient pas pu être réalisés il y a encore deux ou trois ans.

Evolution n°2 : l'amélioration des algorithmes

L'enjeu autour du calcul du PageRank a incité un certain nombre de chercheurs à travailler sur l'algorithme pour simplifier son calcul. De réels progrès ont été faits dans ce domaine à partir de 2003. Cela s'est traduit par exemple par l'introduction de méthodes de type "*interpolation quadratique*" appliquées au calcul du PR ou la méthode baptisée "*power extrapolation*" qui, selon ses auteurs, divise par quatre les temps de calcul.

Evolution n°3 : l'amélioration des méthodes de calcul

Parmi les chercheurs les plus en pointe en 2002/2003 sur l'amélioration de l'algorithme, on trouve trois chercheurs de Stanford fondateur d'une startup, Kaltix, vite rachetée par Google pour récupérer à la fois les cerveaux et les brevets, notamment celui du "*Topic Sensitive PageRank*" dont nous parlerons plus loin. Parmi les idées des "*Kaltix boys*", on trouve notamment une méthode pour calculer très facilement et en temps quasi réel un PR estimé pour de nouvelles pages, sans avoir à recalculer la matrice entière (méthode de calcul par blocs ou "*blockrank*"). Curieusement, quelques mois plus tard, Google abandonnait les *Google Dances* pour passer à une mise à jour du PR en continu...

Un critère universel, immédiatement disponible, et qui donne un bon reflet de l'importance d'une page sur le web

Les critères de mesure de popularité par les liens ont le mérite de réutiliser des données qui sont présentes dans les moteurs de recherche depuis leur création. En effet, un crawler constitue pour faire son travail une base de données de liens, qui est utilisée par l'ordonnanceur (ou "*scheduler*", programme chargé d'organiser le travail du crawler : il distribue les urls à crawler entre les machines, et choisit l'ordre de crawl et le moment où chaque page est appelée) pour organiser son travail. Calculer un PR ne demande pas de réunir d'autres données que celles préexistantes dans un moteur de recherche basique.

Le critère ainsi constitué est universel : il permet de noter tout type de page, quelle que soit la thématique de ces pages, la langue du contenu. Le critère marche aussi bien sur des sites canadiens, chinois ou européens.

Par ailleurs, le critère du PageRank ajouté dans l'algorithme d'un moteur de recherche améliore sérieusement la pertinence des résultats. C'est grâce à cet atout que les internautes ont massivement basculé d'Altavista vers ce nouveau moteur (Google) capable de renvoyer directement un résultat plutôt pertinent en cliquant sur le bouton "*I feel lucky*" (le bouton "*J'ai de la chance*" renvoie directement vers la page apparaissant en tête de résultat sur la requête. Les premiers utilisateurs de Google jouaient beaucoup avec cette fonctionnalité, mais il semble qu'ils aient perdu l'habitude d'utiliser le bouton au point de ne plus le remarquer dans l'interface !).

Ce qui est curieux, on le verra plus loin, c'est que personne n'est parvenu à démontrer pourquoi cela marchait aussi bien (il faut noter que la même absence de fondement théorique explicatif pèse sur l'autre pilier de l'*information retrieval* : "*les espaces vectoriels de Salton*"... Ca marche, les résultats sont pertinents, mais on ne sait pas expliquer vraiment pourquoi un outil qui mesure en fait autre chose produit une note en rapport avec la pertinence !). Au contraire, de nombreuses voix se sont élevées pour expliquer pourquoi ce critère était fondamentalement biaisé.

Un algorithme peu sensible au spam ?

La plupart des chercheurs spécialistes des outils de recherche reconnaissent en général que le PageRank est assez robuste face au *linkspam*. Il est notamment beaucoup moins sensible à des manipulations que ne le sont les algorithmes issus de HITS, même si Teoma (moteur de recherche américain qui utilisait un algorithme de type HITS amélioré, et en particulier durci contre le spam ; Teoma a été racheté par AskJeeves, puis abandonné) en son temps avait démontré que l'on pouvait construire un moteur viable avec cette technologie.

Dans la pratique, il est en fait malgré tout très vulnérable aux attaques de webmasters imaginatifs et dotés de moyens suffisants. Google a trouvé assez vite la parade contre les "fermes de liens", mais pas totalement contre d'autres formes de manipulation, à bases de galaxie de sites ou de sous domaines, ou la création de "fermes de pages" comportant des milliers, voire des millions de pages de contenu.

La période 2003 à 2006 a connu un explosion des recherches sur le *linkspam*. Des avancées spectaculaires ont été réalisées dans la détection de structures destinées à manipuler le PageRank. L'une des méthodes les plus efficaces, inventée par l'équipe de Ricardo Baeza Yates, consiste à faire varier la valeur du coefficient d'atténuation en fonction du nombre de clics qui sépare une page de la page à évaluer. Si le PR change beaucoup en faisant varier le "*damping factor*", c'est que la page est entourée d'une ferme de pages. Le problème c'est que toutes ces méthodes ont tendance aussi à engendrer des "faux positifs", c'est-à-dire à voir du spam là où il n'y en a pas.

Une autre forme de manipulation s'est nourrie également de l'avantage concurrentiel que permettait le fait de disposer d'un plus fort PageRank que son voisin. L'"achat de PageRank" est une technique qui s'est fortement développée jusqu'en 2007. En avril 2007, Matt Cutts a annoncé la fin de la récréation en annonçant qu'il allait, avec son équipe, sanctionner les sites qui manipulaient le PR en achetant des liens. Des sanctions sont réellement apparues quelques semaines plus tard (à l'automne 2007 en France) sous la forme d'une diminution de deux ou trois unités du PR de la barre d'outils de Google. Il faut préciser ici qu'il ne faut pas s'arrêter au discours de l'équipe qualité : Google sanctionne les manipulations du PageRank, quelles qu'elles soient, et pas uniquement l'achat ou la vente de PageRank. Google n'est capable de voir si les liens sont vendus que dans des cas "patents" où le webmaster a indiqué par une mention de type "liens sponsorisés" qu'il s'agissait de liens commerciaux. La pratique de l'achat de PageRank n'a donc pas totalement disparu en 2009, loin de là...

Le PageRank est-il toujours calculé comme expliqué dans l'article initial ?

On peut remarquer qu'en réalité, Brin et Page ont communiqué trois formules de PR différentes dans leurs premières publications. Une seule de ces versions n'est pas contradictoire avec les caractéristiques décrites dans les textes des articles. Etait-ce une volonté de brouiller les pistes, ou une simple maladresse ? Mystère...

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)}$$

L'une des formules du PageRank selon les articles fondateurs de Page & Brin

Il est probable que beaucoup de choses ont changé, à la fois dans la formule du PageRank, et dans la manière de le calculer. Mais il semble qu'il soit difficile de s'éloigner du modèle initial sans connaître de nombreux problèmes, soit au niveau de la difficulté de calcul, soit au niveau de la qualité de la note ainsi calculée (*voir à ce sujet dans la bibliographie ci-dessous les articles "Inside Pagerank" et "Deeper Inside Pagerank" qui donnent de nombreuses infos sur les problèmes de calcul posés par l'algorithme*). Il est donc probable que le PageRank de 2009 reste assez proche des principes décrits en 1998.

Les améliorations les plus utiles que l'on peut apporter portent sur le facteur d'attenuation (*Damping factor*), qui était au départ une constante fixée arbitrairement à 0,85 pour des raisons empiriques et mathématiques (pour assurer la convergence du calcul itératif). Mais il est possible de substituer à ce coefficient d'attenuation, une fonction d'attenuation qui peut dépendre de multiples critères.

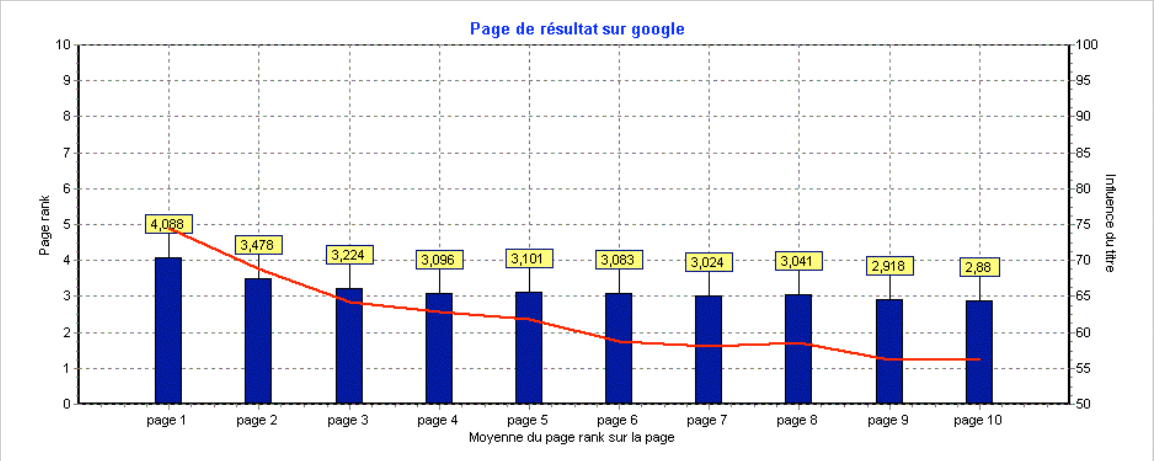
La "souplesse" apportée par ces fonctions d'attenuation "new look" a ouvert la porte à des applications nouvelles du PageRank.

Comment le PageRank est-il utilisé dans l'"algorithme ?

Le PageRank n'a été et n'est toujours que l'un des critères utilisés par Google pour élaborer le classement des pages sur une requête donnée. Créer un algorithme d"un moteur de recherche consiste à élaborer une fonction d'évaluation, qui peut mélanger des critères précalculés (comme le PageRank de la page) et d'autres calculés à la volée (le moins possible pour des raisons de performance).

Le PageRank sert tout d'abord à départager des pages dont le contenu est proche (sémantiquement) de la requête tapée, et qui présentent des scores proches sur l'ensemble des autres critères "*on page*". Dans cette situation, il est intéressant de faire remonter les pages qui peuvent apparaître comme "importantes" aux yeux des internautes, car ces pages sont en règle générale jugées comme des résultats plus pertinents que des pages ayant un faible PageRank.

La pondération de ce critère doit donc être ajustée à une valeur précise pour que, sur une requête, le critère de proximité sémantique reste toujours vérifié. Cela signifie en pratique qu'il est tout à fait normal que sur une requête donnée, une page puisse arriver en première position sans avoir un PR élevé (pourvu que son contenu en fasse un résultat pertinent). Sur des requêtes concurrentielles, par contre, il est difficile d"être présent en première page des résultats sans avoir un PR élevé. Le jeu des autres critères, dont la pondération peut être supérieure à celle du PageRank, fait que même pour des requêtes concurrentielles, l'ordre des pages n'est pas fixé par le PageRank.



Le graphe ci-dessus est tiré d"une étude de Yooda sur la relation entre PR et classement. La courbe montre le PR moyen en fonction des pages de résultat de Google, comparé au poids

d'autres critères comme le titre de la page. On voit que le PR moyen est plus fort en page 1, décroit sur les pages suivantes mais que le critère finit par devenir quasi constant à partir de la page 4 (31e position et plus). On voit aussi (courbe rouge) que le critère "title" semble bien plus discriminant (http://www.yooda.com/info/article/PageRank_positionnement/)

Une autre utilisation du PageRank est de pré-identifier les pages importantes pour faciliter la constitution d'un index optimisé. En effet, il est important dans un moteur de recherche de disposer de beaucoup de "signaux" (les critères utilisés par Google) pour départager les pages qui sortent en tête des résultats. Par contre, le classement des résultats présentés en page 10, voire 50, n'a pas vraiment d'importance. Dans la pratique, on crée donc des index dont le contenu est riche pour des pages importantes (qui ont donc des chances de sortir en tête des résultats) et moins riches pour les autres. Le PageRank peut servir à préordonner les informations dans l'index, améliorant ainsi de manière importante les performances du moteur de recherche. Cette utilisation a été évoquée dans plusieurs brevets déposés par Google, mais aussi par d'autres outils de recherche.

Un PageRank mis à toutes les sauces ...

Lorsque l'on dispose d'un algorithme robuste, fiable, facile à calculer et capable de définir des notes pour des pages à partir de la structure des liens, il est tout à fait logique de penser à réutiliser ce principe pour calculer d'autres notes pour des critères influencés par les liens.

L'application la plus directe est la lutte contre le spam, et notamment au link spam (celui qui s'attaque justement au critère du PageRank). Dès les années 2000, on a vu apparaître dans la littérature scientifique la notion de "*BadRank*". Le BadRank est un PageRank calculé "à l'envers" où le fait de faire un lien sortant vers un site marqué comme "spammy" transmet une mauvaise note à la page. Mais on a par la suite également imaginé différentes versions de "spamranks" dans lequel le fait de recevoir des liens en provenance de site spammy transmet une mauvaise note au site. Plus récemment est apparue (chez Yahoo) la notion de Trustrank, qui est la face positive du spamrank, dans laquelle on note les pages sur un critère de "confiance" transmis par les liens émanant de sites marqués comme "dignes de confiance".

D'autres ont également découvert qu'en "biaisant" le facteur d'atténuation, en le faisant varier en fonction du nombre de sauts (de clics) séparant une page d'une autre, on peut découvrir des structures comme des fermes de liens, comme nous le disions auparavant, ou des collections de pages chargées de doper le PR des pages cibles.

Parmi les variantes à l'aide de PageRank "biaisés", on trouve également l'algorithme Hilltop, qui avait été évoqué comme explication aux changements de classement spectaculaires intervenus fin 2003.

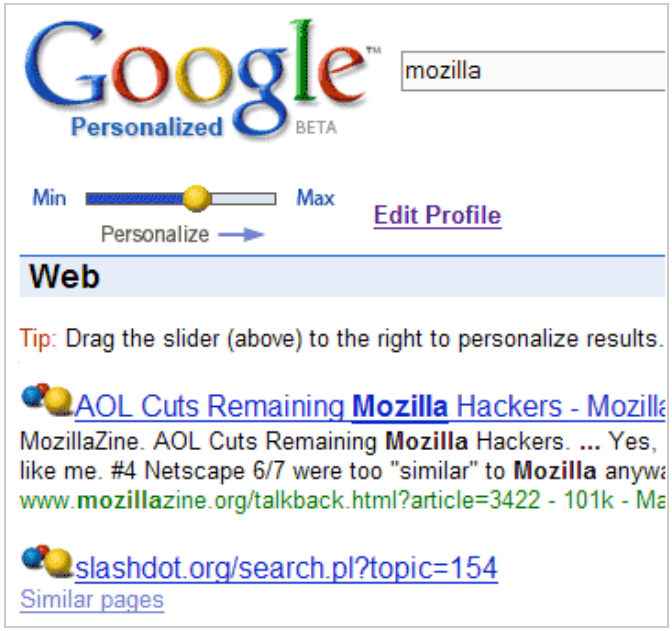
Un seul PageRank, ou plusieurs PageRanks ?

Google ayant développé de manière considérable sa capacité à calculer des PageRanks, certaines méthodes de classement s'appuyant sur plusieurs notes sont devenues possibles. Nous venons d'évoquer le fait qu'une page pouvait se voir attribuer non seulement un PR, mais aussi un Trustrank ou un Spamrank.

En 2003, Kaltix avait développé un nouveau concept : le "*Topic Sensitive PageRank*" (TSPR). Le TSPR s'appuie sur une série de PageRank biaisés : on identifie un certain nombre de sites appartenant à une grande thématique donnée (l'une des seize catégories principales de l'annuaire DMOZ par exemple), et on calcule un PageRank à partir des liens sortants de ces pages. Plus le site est "hors thématique", moins il a de chances de recevoir des liens de ces sites, et plus son TSPR est faible. C'est l'inverse si le site est parfaitement dans la thématique de la catégorie DMOZ. On peut calculer autant de TSPR que de thématiques définies (mais un nombre limité de TSPR suffit pour calculer un spectre continu de notes).

Une fois le jeu de TSPR calculés, il est possible soit d'utiliser le TSPR pour améliorer le classement après avoir identifié à quelle thématique se rattache une requête, soit de personnaliser les résultats en fonction des thématiques préférées des internautes. Cette

dernière application possible est apparue dans Google Labs quelques mois après le rachat de la technologie de Kaltix, pour évoluer vers le système de personnalisation des résultats actuel de Google.



La version "Google labs" de la recherche personnalisée. Le fonctionnement de la personnalisation par l'utilisateur montrait une filiation directe avec le système à base de TSPR inventé par Kaltix.

Bref, il est probable que les index des principaux moteurs ne calculent plus un seul PR, mais plusieurs "ranks" calculés avec des objectifs divers.

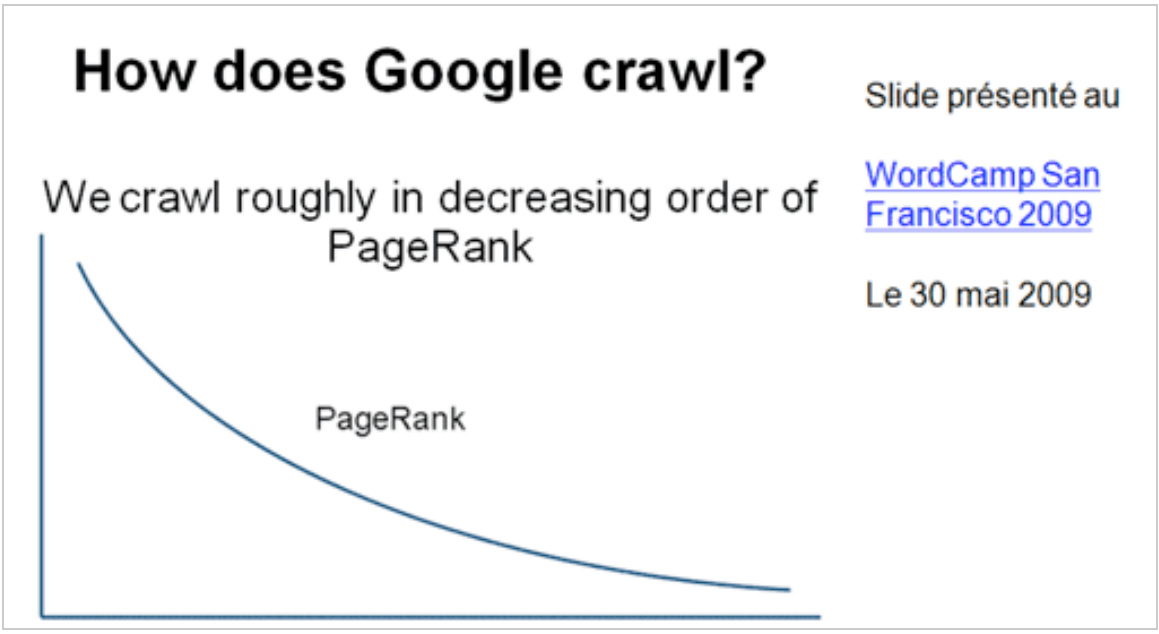
Est-ce que le PR sculpting permet d'améliorer ses positions ?

Un certain nombre de spécialistes du référencement utilisent sur les sites qu'ils optimisent la technique dite du "PR sculpting". La méthode repose sur des changements apportés à la structure des liens reliant les pages d'un site entre elles, destinés à changer le niveau de PR obtenu par chaque page du site. Le but est d'augmenter le PR des pages importantes du site, et de diminuer le PR obtenu par des pages moins importantes.

Pour obtenir ce résultat, soit on change à la fois la structure des liens visibles par les internautes et par les moteurs, soit on crée une structure visible uniquement par le moteur, en s'aidant de l'attribut *rel=nofollow*, ou en cryptant les liens à l'aide de scripts Javascripts complexes ou encore en programmant les liens dans les tréfonds d'un fichier flash. La deuxième alternative a tendance à créer une structure éloignée de la réalité vue par l'internaute, ce qui peut constituer une violation des conditions générales de Google si on va trop loin dans ce domaine.

Est-ce que le PR peut avoir une influence sur la manière dont Google crawle mon site ?

Matt Cutts a rappelé récemment quelque chose (*slide présentée au WordCampSan Francisco le 30 mai 2009*) que les webmasters qui suivent les visites de Googlebot sur leur site savent bien : le PageRank d'une page influence la fréquence avec laquelle Googlebot visitera cette page, et Google a tendance à ne pas crawler du tout des pages qui n'ont pas de PageRank.



Les failles du PageRank et les améliorations apportées au modèle

Le PageRank présente de nombreuses failles théoriques, qui concernent en particulier les axiomes sur lesquels il repose. Voici un florilège des remarques qui ont pu être faites par les chercheurs spécialistes des outils de recherche :

- Tous les liens ne se valent pas : les internautes ne choisissent pas "au hasard" les liens sur les pages qu'ils visitent. Certaines pages sont plus importantes que d'autres et les gens ne lisent que rarement les pages "*mentions légales*" ou "*qui sommes nous*". Le modèle du "surfeur aléatoire" est donc totalement déconnecté de la réalité.

- Un surfeur qui se lasse ne va pas non plus visiter certaines pages. Le modèle théorique du PageRank prévoit la possibilité d'une "téléportation" de temps à autre, c'est-à-dire que le surfeur aléatoire se lasse de cliquer sur des liens, rentre une url au hasard dans la barre d'adresse de son navigateur, et se retrouve téléporté ailleurs sur le web sans avoir cliqué de lien. Là aussi, le modèle ne colle pas à la réalité : la probabilité que le nouveau site visité soit la page d'accueil de Google ou d'un grand portail est bien plus grande qu'une obscure page d'un site personnel.

- Les pages changent et perdent de la valeur (ou en gagnent) à des rythmes très différents, et cette "valeur" aboutit à une déconnexion entre "l'importance" de la page mesurée par PR, et la vraie valeur de la page sur des critères de pertinence pure. Le critère du PR ignore que certaines pages sont achetées, que leur contenu peut être modifié pour des raisons commerciales, politiques, stratégiques, ou que leur contenu se déprécie avec le temps.

- Les "astuces" pratiques pour calculer le PageRank ont des effets de bord : notamment, les raccourcis du type "blockrank", où on agrège les liens au niveau d'un bloc (domaine, sous domaine, ou machine hôte), finissent par attribuer des PR inexacts à certaines pages.

Du surfeur aléatoire au surfeur intelligent

Dès 2002, Richardson et Domingo ont proposé un algorithme modifié. Cet algorithme part du principe que le surf de l'internaute ne clique pas sur n'importe quel lien sur une page, mais, dans un contexte de recherche, sur une page qui est en rapport avec la requête. Le *Query Directed PageRank* ainsi créé est construit pour mesurer la probabilité à chaque saut que la page soit une réponse pertinente à une requête donnée. Il se base donc à la fois sur la

structure des liens, mais aussi sur l'analyse du contenu de la page. Par opposition au modèle du surfeur aléatoire du PageRank, Richardson et Domingo ont appelé leur modèle le "surfeur intelligent".

Leur algorithme repose à la fois sur des valeurs précalculées pour un nombre important de requêtes et sur un calcul fait à la volée à partir de ces valeurs précalculées. L'ensemble est beaucoup plus lourd à utiliser que le PR dans un outil de recherche, mais l'implémentation de ce genre d'algorithmes est devenue beaucoup plus simple avec les progrès obtenus dans la puissance et la capacité des machines.

Du surfeur aléatoire au surfeur poursuivant un objectif

En partant de l'idée de Richardson et Domingo, d'autres chercheurs ont imaginé des approches très similaires, et ont donné naissance aux algorithmes basés sur les modèles dits du "surfeur poursuivant un objectif" (*intentional surfer*). Les différentes implémentations de ces algorithmes exploitent toutes l'idée que l'on peut (doit ?) introduire dans l'évaluation de la probabilité qu'un surfeur clique sur une page plutôt qu'une autre, des informations comportementales issues d'outils comme les barres d'outils (et, évidemment, on pense à la Googlebar, l'une des barres les plus répandues et utilisées).

Un moyen pour contourner le problème du nofollow ?

Google a créé un monstre en fournissant aux webmasters un outil pour supprimer des liens pour les spiders mais pas pour les internautes : l'attribut '*rel=nofollow*'. Très vite, de très nombreux webmasters ont mis des nofollows partout, y compris dans des endroits où cela n'avait pas de sens. Dans ce contexte, la matrice des liens du web connue par le moteur a commencé à diverger gravement de la structure des liens "réelle", celle vue et cliquée par les internautes.

On comprend mieux à quoi s'attaque Matt Cutts dans ses propos prononcés au dernier SMX Advanced, le 2 juin 2009 (http://actu.abondance.com/2009/06/google-sur-le-point-de-reconnaitre-les.html) : l'utilisation anarchique des nofollow finit par nuire à la bonne qualité du critère "PageRank". Il faut noter que les algorithmes fondés sur l'"intentional surfer" ne sont pas sensibles à ce problème.

Le PageRank est-il mort, ou faut il toujours compter avec lui ?

Quand on mesure l'énergie considérable dépensée par le service qualité de Google pour défendre les fondements de l'algorithme du PageRank, on peut en conclure deux choses :
- d'une part, Google continue à utiliser un critère ressemblant fortement au PageRank originel, et ce critère constitue toujours l'un des éléments clés qui lui sert à classer les pages.
- Google a sans doute sophistiqué l'algorithme au fil du temps, mais pas assez pour le rendre insensible au spam. Il a donc besoin parfois de le défendre avec d'autres moyens que la force brute de calcul ou les mathématiques : le rappel de bonnes pratiques, et des sanctions exemplaires contre les personnes qui attaquent l'algorithme.

Maintenant, le critère a-t-il été déprécié au fil du temps ? C'est de toute façon difficile à dire, tant le PageRank a toujours été, aujourd'hui comme hier, un critère parmi de nombreux autres qui servent à classer les pages, sans être **LE** critère.

D'autres critères (Google les appels des "signaux") se sont ajoutés dans l'algorithme, et l'impact de ces critères a tendance de plus en plus à masquer le poids du PageRank dans la note finale de la page. Donc, d'une certaine façon, l'appréciation subjective du poids du PR selon Rand Fishkin cité au début de cet article correspond à ce que l'on constate de manière générale.

Par contre, il est clair qu'aucun algorithme de moteur de recherche actuel ne peut se passer dans son algorithme des informations fournies par la structure des liens qui relient les pages du web entre elles. Aujourd'hui encore, les pages d'un site ne peuvent pas en général se

classer en tête des résultats sans bénéficier d'un environnement de linking transmettant au site des votes d'importance, de qualité, de légitimité, et de confiance... Autant dire que le PageRank, ou ses avatars ou successeurs, ont encore quelques belles années devant eux...

Bibliographie

Articles généraux

The Anatomy of a Large-Scale Hypertextual Web Search Engine
Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia

The PageRank citation ranking: Bringing order to the web
L Page, S Brin, R Motwani, T Winograd - 1998 - cpe.ku.ac.th

PageRank as a function of the damping factor
P Boldi, M Santini, S Vigna - Proceedings of the 14th international conference on World ..., 2005 - portal.acm.org

Inside PageRank
M Bianchini, M Gori, F Scarselli - ACM Transactions on Internet Technology (TOIT), 2005 - portal.acm.org

Deeper inside PageRank
AN Langville, CD Meyer - Internet Mathematics, 2004 - projecteuclid.org

Link spam detection based on mass estimation
Z Gyongyi, P Berkhin, H Garcia-Molina, J Pedersen - Proceedings of the 32nd international conference on Very ..., 2006 - portal.acm.org

The cost of attack of PageRank
A Clausen - Proceedings of The International Conference on Agents, Web ..., 2004

The intelligent surfer: Probabilistic combination of link and content information in PageRank
M Richardson, P Domingos - wortschatz.uni-leipzig.de

Articles des fondateurs de Kaltix

Topic-Sensitive PageRank : A Context-Sensitive Ranking Algorithm for Web Search - Taher H. Haveliwala / 10 février 2002 revu le 17 mai 2003

Extrapolation Methods for Accelerating PageRank Computations - Sepandar D. Kamvar Taher H. Haveliwala Christopher D. Manning Gene H. Golub / 28 Février 2003

Exploiting the Block Structure of theWeb for Computing PageRank - Sepandar D. Kamvar Taher H. Haveliwala Christopher D. Manning Gene H. Golub / 4 Mars 2003

The Second Eigenvalue of the Google Matrix - Taher H. Haveliwala and Sepandar D. Kamvar / 11 Mars 2003

Adaptive Methods for the Computation of PageRank - Kamvar, Sepandar ; Haveliwala, Taher ; Golub, Gene / 28 avril 2003

An Analytical Comparison of Approaches to Personalizing PageRank - Taher Haveliwala, Sepandar Kamvar and Glen Jeh / 20 juin 2003

The Condition Number of the PageRank Problem - Sepandar D. Kamvar and Taher H. Haveliwala / 20 juin 2003

Computing PageRank using Power Extrapolation - Haveliwala, Taher ; Kamvar, Sepandar ; Klein, Dan ; Manning, Chris ; Golub, Gene / 16 juillet 2003

Philippe Yonnet, Directeur Technique @Position (http://www.aposition.com) et président de l'association SEO Camp (http://www.seo-camp.org/)

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/06/le-pagerank-en-2009-mythe-ou-realite.html