Amplification de PageRank : du surfeur aléatoire à l'optimisation du maillage

Par Guillaume et Sylvain Peyronnet

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

Le mois dernier, nous avons expliqué ici le concept du PageRank thématique. Mais savez-vous qu'une gestion intelligente du « surfeur aléatoire », qui est à la base même de la définition du PageRank, peut vous faire gagner énormément en référencement naturel au travers d'un netlinking optimisé et d'une structure de liens bien pensée ? Voici comment faire...

Le mois dernier, nous avons évoqué dans ces mêmes colonnes le PageRank thématique. Mais en parcourant les blogs, nous avons pu nous rendre compte que le concept même de PageRank classique était souvent mal compris. Cette mécompréhension est largement alimentée par la communication fumeuse de Google, et dans cet article nous espérons clarifier certains faits scientifiques concernant ce critère de pertinence.

Le PageRank, qu'est ce que c'est ?

Si vous avez lu et compris l'article du mois dernier sur le PageRank thématique, vous pouvez probablement sauter ce qui suit. Sinon, n'hésitez pas à lire le prochain paragraphe :-)

Le PageRank a été créé en 1998 par Sergey Brin et Larry Page (voir [1]). Il s'agit d'abord d'une mesure de popularité qui utilise la modélisation d'un internaute virtuel (le **surfeur aléatoire**) pour quantifier l'importance des pages web les unes par rapport aux autres. Le surfeur aléatoire est un internaute qui parcourt le web en cliquant aléatoirement sur les liens sortants de la page sur laquelle il est, à un moment donné. Parfois, le surfeur aléatoire peut décider de changer totalement de voisinage, ce qui est implicitement modélisé par une probabilité de continuer, ou non, à suivre les liens. Lorsqu'il décide de ne plus suivre les liens, il recommence son parcours à partir d'une page tirée au sort aléatoirement sur le Web. On parle de téléportation pour désigner cet évènement.

La définition du PageRank est intimement liée au concept de surfeur aléatoire. En effet, la probabilité que le surfeur aléatoire se trouve sur une page donnée à un moment précis est exactement le PageRank, par définition. Même s'il existe de belles formules pour calculer la valeur du PageRank d'une page, il n'y a nul besoin de les connaître pour agir sur ses sites web en vue d'augmenter leur popularité. En effet, avec quelques raisonnements intuitifs sur le surfeur aléatoire, on peut facilement comprendre ce qu'il faut faire en matière de linking.

Les idées de base

Nous allons tout d'abord passer en revue quelques faits très simples qu'il faut garder à l'esprit.

1. On peut gagner du PageRank avec un lien sortant

C'est sans doute ce qui est le moins connu de la plupart des SEO. Ce qui fait la popularité d'une page, ce n'est pas son nombre de liens entrants, mais bien le fait que le surfeur aléatoire y passe très souvent. En faisant un lien sortant de manière habile, on peut créer des cycles qui vont faire revenir souvent le surfeur aléatoire, et qui ont donc pour effet d'augmenter le PageRank (nous verrons plus loin jusqu'à quel point on peut augmenter son PageRank ainsi).

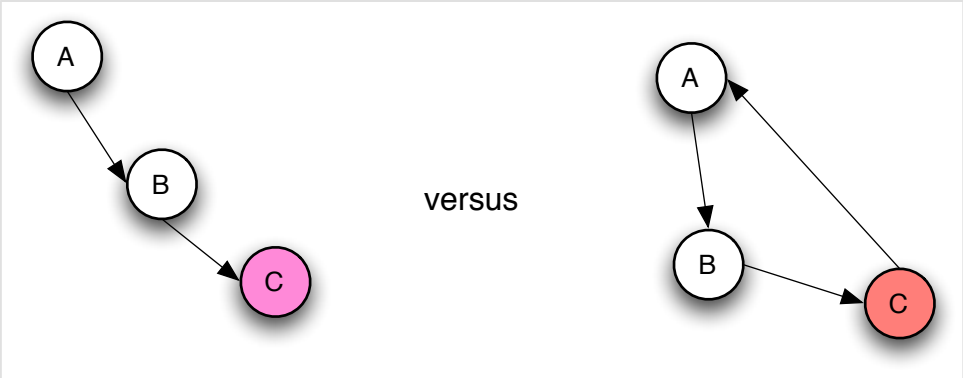

Fig. 1. Deux types de linking standard.

Sur la figure 1, on voit à gauche un schéma de linking standard, avec la page C qui reçoit un lien de la page B. A droite, on a les mêmes pages, la différence étant juste que C fait un lien sortant vers A. En faisant cela, une boucle est créée, et le surfeur aléatoire va « cycler » : il va visiter la page C plus souvent dans le schéma de droite que dans celui de gauche (comme il n'y a pas de lien sortant, il ne peut que se téléporter ailleurs sur le web). Le PageRank de la page C a donc augmenté grâce à ce lien sortant.

Plus une page sera au cœur d'un grand nombre de boucles, et plus ces boucles seront courtes, plus son PageRank sera grand. Il faut donc toujours penser à bien intégrer ses pages dans des cycles lorsque l'on met en place son linking (interne comme externe).

2. On peut gagner du PageRank grâce à un lien entre deux sites différents de celui que l'on veut pousser

Voici encore une phrase qui peut sembler étonnante pour ceux qui ne connaissent le PageRank que *via* la formule mathématique, mais qui est pourtant tout à fait vraie et naturelle. La figure 2 explique ce principe de manière très claire.

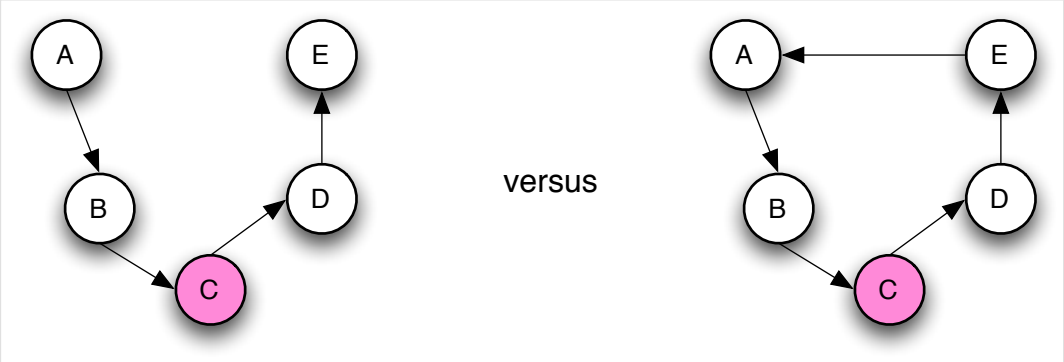

Fig. 2. Deux types de chaînes de pages.

A gauche on a une chaîne de pages qui sont liées entre elles, et à droite une légère modification : la même chaine avec en plus un lien entre la page E et la page A. Ce lien supplémentaire permet au surfeur aléatoire de revenir plus souvent sur la page C, ce qui améliore son PageRank.

Ce type d'amplification du PageRank est particulièrement intéressante de par sa discrétion. Qui pourrait penser que la page C visait à augmenter son PageRank via un lien entre deux pages sans rapport avec elle ? En revanche, l'amplication obtenue reste mesurée, et profite à toutes les pages de la boucle qui a été créée.

Après avoir vu ces deux idées très simples, conséquences directes de la définition même du PageRank comme probabilité de passage du surfeur aléatoire, nous allons voir comment on peut améliorer de manière optimale son PageRank.

Structure optimale pour amplifier le PageRank d'une page

Bianchini, Gori et Scarselli ont étudié en profondeur le PageRank et ses mécanismes internes (voir l'article [3]). Cette étude leur a permis de mettre en évidence la structure optimale qui permet de maximiser la valeur du PageRank d'une page. Sans surprise, cette structure vise à maintenir le surfeur aléatoire le plus proche possible de la page dont on souhaite maximiser le PageRank, dans le but de le faire passer plus souvent sur cette page.

Le travail de Bianchini et ses collègues a ainsi abouti à la mise en place d'une structure optimale pour la maximisation du PageRank de chaque page d'un groupe de pages données, mais aussi pour la maximisation d'une page en particulier.

Ces structures optimales sont présentées dans la figure 3, qui différencie le cas où l'on peut faire un lien « bouclant » de celui où cela n'est pas possible.

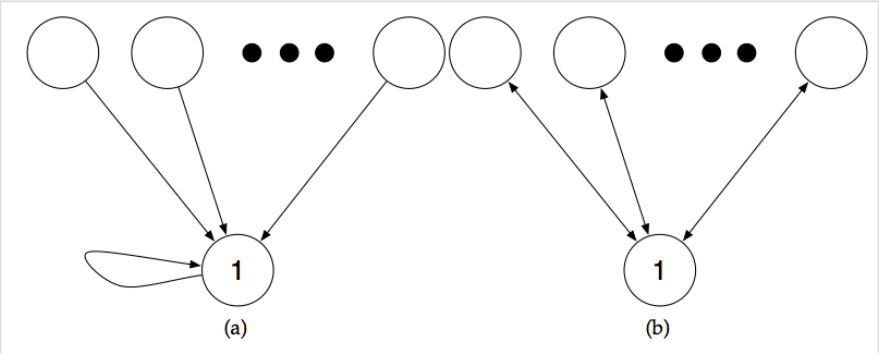

Fig. 3. Deux structures de liens.

A gauche, on peut voir la structure avec un lien bouclant, et à droite celle avec des liens réciproques entre la page cible et les pages qui fournissent du PageRank. Ces deux structures donnent une amplification optimale (on ne peut pas faire mieux) du PageRank, mais ne sont pas satisfaisantes en matière de référencement : celle de gauche est sans doute caduque par défaut au niveau du moteur (qui ne va pas prendre en compte le lien de la page sur elle-même), et celle de droite n'est pas réaliste car la page cible doit faire des liens vers toutes les pages « sources ».

Heureusement, Zoltan Gyöngyi (qui a fait sa thèse sous la direction de Hector Garcia-Molina, tout comme Sergey Brin), a montré dans l'article [4] qu'on pouvait créer un schéma optimal à l'aide d'une page de boost. Ce schéma optimal est visible dans la figure 4.

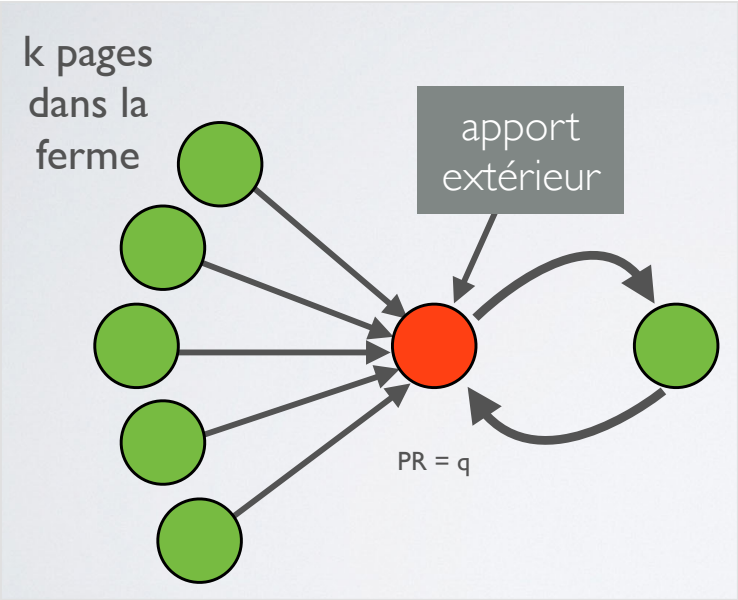

Fig. 4. Exemple de ferme de liens.

Une telle structure s'appelle une ferme de liens, et la page de *boost* est ici la page en vert, à droite de la page à optimiser, qui est en rouge.

Zoltan Gyöngyi a pu montrer qu'une telle structure permettait de multiplier le PageRank de la page cible (en rouge) par 3,6. Il faut bien se rendre compte que cela signifie que si on a une page web avec déjà 1 millions de liens entrants, on obtient le même effet en

rajoutant 2,6 millions de liens entrants qu'en faisant une seule boucle avec une page de boost. Voici une information qui devrait remettre en perspective un certain nombre d'opérations de netlinking...

Ce type de structure est optimal lorsque l'on est capable de les mettre en place. Ce qui est loin d'être évident, et qui est en plus facilement repérable par les principaux moteurs de recherche, Google en tête. Dès lors, il est important de savoir ce qui peut être fait pour augmenter le PageRank d'une ou plusieurs pages cibles sans pour autant en créer des milliers d'autres. Il est par exemple intéressant de partir du principe que l'on ne peut que manipuler la topologie du graphe du Web autour des pages qui nous intéressent, sans jamais rajouter ou enlever de pages supplémentaires. Dans un tel contexte, le propriétaire d'une page web peut manipuler les liens sortants de ses pages. Il n'a en revanche aucun moyen de travailler les liens provenant de l'extérieur et dirigés vers ses pages. En procédant ainsi, on obtient une amplification sous-optimale, mais difficile à tracer pour le moteur, qui se cantonne généralement à pénaliser les anomalies sur les liens entrants (les backlinks).

Structure de lien optimale pour une page seule

Il s'agit ici plutôt d'un cas d'école tant on est sur un contexte improbable. Avrachenkov et Litvak [5] ont étudié les effets de l'ajout de liens dans le graphe du Web, sur la valeur du PageRank des pages. Ils ont ainsi pu donner la stratégie de création de liens optimale pour la maximisation du PageRank d'une page seule.

Quand on a compris le concept de surfeur aléatoire, le résultat n'est pas surprenant (et même plutôt décevant) : la stratégie optimale pour augmenter le PageRank d'une page cible seule est de faire un seul lien sortant depuis cette page, et que ce lien pointe vers la page du Web qui va permettre de revenir le plus vite possible vers la page cible. Lorsqu'il y a une page de type boost, ce sera cette dernière qui aura le temps de retour le plus faible, mais de manière générale cela peut être n'importe quelle page de l'entourage de la page cible (car le temps de retour dépend de tous les chemins de liens qui passent par les pages étudiées).

Au final, ce résultat théorique n'a pas un impact pratique très important. En effet, il est souvent très peu pratique de ne faire qu'un seul lien, par exemple car cela ne permet pas de mettre en place un système de navigation réaliste au sein du site. De plus, il est en pratique extrêmement difficile d'estimer la valeur du temps de retour pour toutes les pages de l'entourage. Il est donc plus raisonnable de pousser sa chance en créant soi-même une page de boost, ce qui permet d'obtenir l'amplification optimale facilement.

Structure de liens optimale pour un ensemble de pages

Un ensemble de pages est typiquement un site web. On va donc voir maintenant comment faire pour obtenir une popularité maximale pour un site complet. La popularité d'un site n'a pas été définie plus haut, mais il s'agit d'un concept très simple : c'est la probabilité que le surfeur aléatoire soit à un moment donné sur une des pages du site, c'est donc la somme des PageRank des pages du site.

De Kerchove, Ninove et Van Dooren (voir l'article [6]) se sont appuyés sur les travaux d'Avrachenkov et Litvak [5] pour obtenir une généralisation à un ensemble de pages. Comme dans les travaux précédents, ils vont partir du principe qu'il n'est pas possible de générer des pages à volonté. Ainsi, un certain nombre d'hypothèses « raisonnables » vont être demandées afin de fournir une structure optimale pour la somme des PageRank d'un ensemble de pages prédéterminées. La principale hypothèse est le fait que chaque page doit avoir un accès au reste du web. L'hypothèse est naturelle : une structure de site qui ne permet pas d'accéder à d'autres sites web est largement inhabituelle. Il faut cependant bien noter qu'on exige un accès vers le reste du web, mais pas un lien direct. Si une seule page du site permet d'accéder à un site tiers, c'est suffisant si la page en question est accessible depuis toutes les pages du site.

La structure optimale obtenue est celle présentée dans la figure 5.

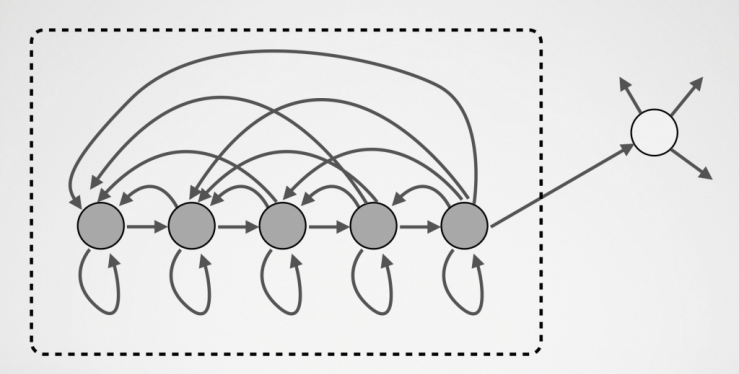

Fig. 5. Interconnexion d'un ensemble de 5 pages.

Il s'agit d'un exemple pour un site de 5 pages, et on voit clairement ici que le schéma est conçu pour maximiser localement et en permanence la présence du surfeur aléatoire sur des pages loin de la page qui permet d'accéder au reste du web. En procédant ainsi, on crée ce que l'on appelle un « piège à surfeur », ce qui est généralement la marque permettant aux moteurs de recherche de repérer les manipulations de PageRank. L'empreinte d'un tel schéma est caractéristique, d'autant plus qu'un tel linking se rencontre rarement au sein des sites web naturels.

Conclusion

Si on résume les bonnes pratiques, on se rend compte qu'elles sont peu nombreuses, et très simples. Et oui, faire du netlinking intelligent, c'est facile, en suivant quelques règles :

• On a un bon schéma d'amplification quand on a des liens réciproques, ou bien quand on a des liens plus classiques et une page de boost.
• La meilleure façon de faire du linking est de faire un seul lien sortant par page. Et comme ce n'est pas du tout réalisable « en vrai », quand on travaille sur des sites, notamment à cause des éléments de navigation, on ne peux pas faire mieux que de faire le moins possible de liens sortants inutiles (= ceux qui ne permettent pas de faire revenir le surfeur aléatoire).
• Quand on a un site web, on évite de faire des liens vers l'extérieur sur toutes les pages (on peut fédérer un certaines nombres de liens sur des pages dédiées).

Voilà, c'est maintenant à vous de jouer chers lecteurs, en espérant que ces quelques réflexions autour du PageRank auront inspirer chez vous de nouvelles méthodes de linking.

Références

[1] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the Web.
http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf

[2] Largillier, T., & Peyronnet, S. (2014). Algorithmique du web: autour du PageRank. Informatique Mathématique Une photographie en 2014. Presses Universitaires de Perpignan.

[3] Bianchini, M., Gori, M., & Scarselli, F. (2005). Inside PageRank. ACM Transactions on Internet Technology (TOIT), 5(1), 92-128.
http://www.di.ens.fr/~vergnaud/algo0910/PageRank.pdf

[4] Gyöngyi, Z., & Garcia-Molina, H. (2005, August). Link spam alliances. In Proceedings of the 31st international conference on Very large data bases (pp. 517-528). VLDB Endowment.
http://ilpubs.stanford.edu:8090/679/1/2005-15.pdf

[5] Avrachenkov, K., & Litvak, N. (2004). Decomposition of the google PageRank and optimal linking strategy.
http://doc.utwente.nl/80247/1/RR-5101.pdf

[6] De Kerchove, C., Ninove, L., & Van Dooren, P. (2008). Maximizing PageRank via outlinks. Linear Algebra and its Applications, 429(5), 1254-1276.
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.331.8108&rep=rep1&type=pdf

Guillaume Peyronnet *est gérant de Nalrem Médias.* **Sylvain Peyronnet** *est co-fondateur et responsable des ix-labs, un laboratoire de recherche privé. Ensemble, ils font des formations, pour en savoir plus : http://www.peyronnet.eu/blog/*