Robots.txt, crawl et indexation - 2ème partie : la perte de PageRank

Retour au sommaire de la lettre

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

Le fichier robots.txt est un grand classique du Web et du SEO. Pourtant, êtes-vous sûr de l'utiliser à bon escient et de bien comprendre son fonctionnement ? Nous avons présenté le mois dernier un état de l'art de la meilleure façon d'utiliser le fichier robots.txt pour mieux contrôler la vision de votre site qu'auront les moteurs de recherche. Ce mois-ci, nous vous présentons les dangers, en termes de PageRank, de l'utilisation trop assidue de ce fichier. Ou quand les "dandling pages" compromettent le référencement de votre site au lieu de l'aider...

Dans la première partie de cet article, le mois dernier, nous avons insisté sur les erreurs les plus courantes à propos de la manière de configurer et d'utiliser un fichier robots.txt. Mais il en restait un à aborder: les fuites de pagerank engendrées par une mauvaise utilisation de ce fichier... Le mécanisme à l'origine du problème est souvent méconnu, et un peu difficile à comprendre. Mais il est important de prendre conscience des conséquences funestes que ce phénomène peut avoir sur le référencement de vos sites.

Or, c'est une erreur assez commune, et que l'on trouve paradoxalement surtout sur des sites optimisés par des référenceurs : le recours au blocage par une directive du fichier robots.txt est souvent fait à des fins d'optimisation (en général pour faire du "PR sculpting"). En réalité, le remède peut s'avérer plus grave que le mal.

Mais pour mieux comprendre pourquoi un robots.txt peut générer un véritable "trou noir à pagerank", rappelons d'abord quelques mécanismes fondamentaux à propos de cet algorithme de Google.

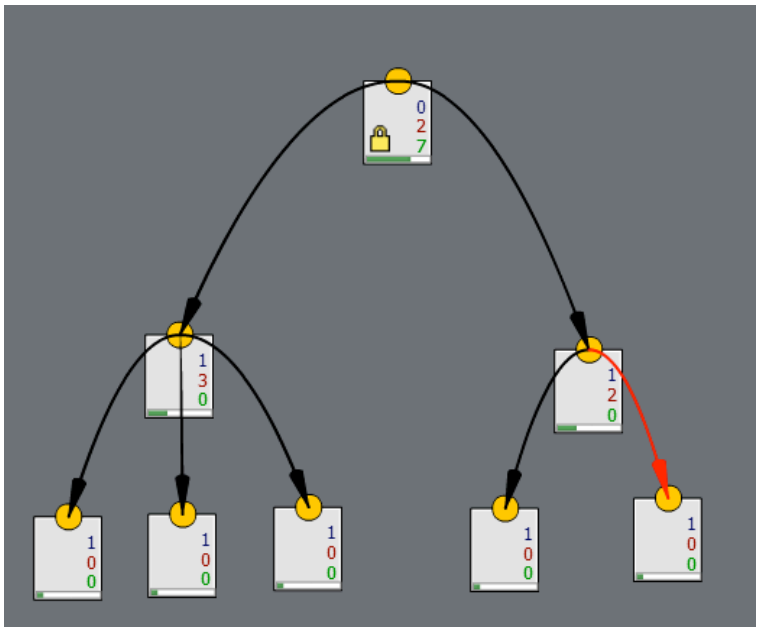
Le pagerank augmente grâce aux liens entrants, et diminue via les liens sortants

Le pagerank est un "signal" (parmi de nombreux autres signaux utilisés par Google) mesurant la popularité d'une page sur le net, calculé à partir des informations tirées des liens hypertexte reliant les pages web entre elles.

Si l'on s'intéresse à la formule du pagerank d'une page, on s'aperçoit vite que le PR d'une page dépend de la somme des PR transmis à cette page par les liens entrants : plus une page a de backlinks, et plus ces backlinks transmettent un PR élevé, plus la page reçoit de PR.

Mais le calcul du PR est un processus itératif : car comme la note d'une page dépend des notes des pages qui contiennent un lien pointant vers elle et que chaque calcul modifie les notes des pages en cascade, il faut faire "tourner" l'algorithme plusieurs fois pour faire converger les notes vers une valeur stable.

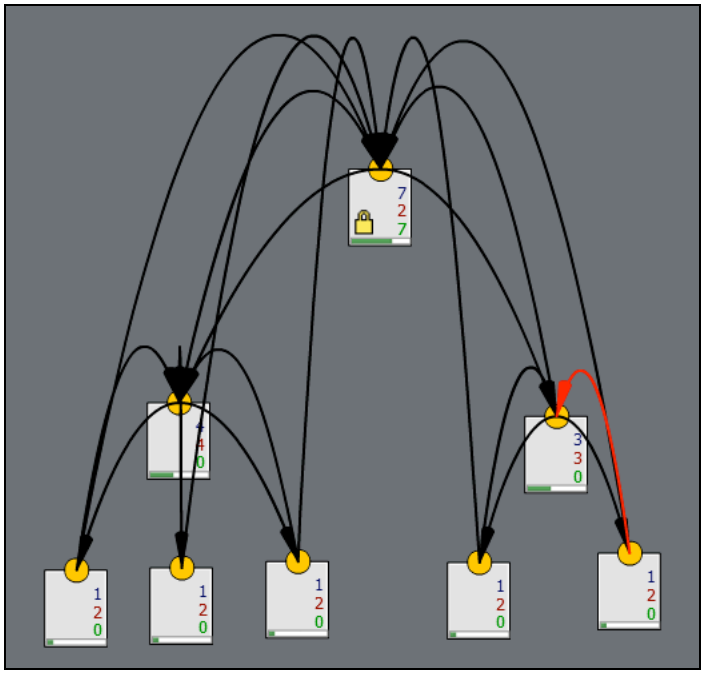
Ce calcul itératif conduit à constater que plus une page fait de liens vers l'extérieur, et surtout, plus le ratio liens sortants sur liens entrants est important, plus la page a tendance à "perdre" du pagerank.



Un exemple calculé avec la formule du pagerank donné dans l'article originel de Page et Brin décrivant le Pagerank : les rectangles représentent des pages web, et les flèches les liens reliant ces pages. Dans cet exemple, si la page d'accueil a une note de 7, les deux pages intermédiaires ont finalement un PR de 3,12.

Les "fuites" de pagerank : un mécanisme plus complexe et moins intuitif qu'il n'y parait au niveau local

On pourrait se dire : "si je veux maximiser le pagerank d'une page, j'ai donc intérêt à avoir le moins possible de liens sortants sur ma page". Mais comme le pagerank est calculé de manière itérative, le "jus" diffusé vers l'extérieur peut très bien revenir grâce aux liens remontant vers les pages qui font des liens vers la page considérée. Et le moins que l'on puisse dire, c'est que les raisonnements intuitifs au niveau d'une seule page sont souvent battus en brêche dès lors que l'on considère le phénomène de diffusion/concentration du pagerank au niveau d'un réseau de pages plus large.



Il suffit de rajouter des liens sortants sur les pages pour que la note de PR augmente : ici, les pages font toutes un lien vers la page d'accueil, et les pages filles vers les pages parentes. Résultat : la note des pages intermédiaires passe de 3,12 à 4,54 et 4,28 (elles ont « concentré » du pagerank). Nota bene : ces valeurs données par la formule du PR n'ont rien à voir avec les valeurs données par la barre d'outils de Google.

Par contre, si on raisonne en termes de "potentiel de pagerank", c'est à dire en terme de "somme de tous les pageranks des pages considérées", alors on retrouve une notion de perte plus intuitive...

La formule du potentiel de pagerank

Trois chercheurs italiens ont, au cours de l'année 2004, réalisé une étude assez complète des propriétés mathématiques de la formule du pagerank (cette étude a fait l'objet d'une publication scientifique sous le nom "Inside Pagerank", voir reférence dans la bibliographie). Au passage, ils se sont intéressés à la notion de potentiel de pagerank d'un site, et ce qui permettait de "maximiser" globalement (et non localement) le potentiel de pagerank d'un site.
Ils sont parvenus à la conclusion que l'on pouvait décomposer le potentiel de PR d'un site de la manière suivante :

$$P = |I| + P_{in} - P_{out} - P_{Dp}$$

Où :
- **|i|** est le potentiel dit "intrinsèque" du site, c'est à dire le PR qui provient uniquement des pages du site à l'exclusion de tout lien externe. Rappelons au passage que :
  * une page qui reçoit zéro lien a, selon la formule du pagerank, un pagerank non nul ;
  * dans la formule originelle, il n'y a pas de différence entre pages "externes" ou "internes" dans le calcul du pagerank (un backlink interne compte exactement comme un backlink externe).

Bref, une page peut très bien avoir un pagerank élevé en raison uniquement des liens qui proviennent d'un même site. Il y'a donc une véritable prime aux sites avec un fort volume de pages, au sein duquel il est possible de "booster" des pages individuelles à l'aide du maillage interne, même si ces pages ne reçoivent aucun backlink depuis d'autres sites.

Il semble d'ailleurs que la mise à jour Mayday ait eu pour objectif d'atténuer en partie ce phénomène pour les cas les plus flagrants.

- **PRin** mesure la somme des PR transmis depuis des backlink externes.

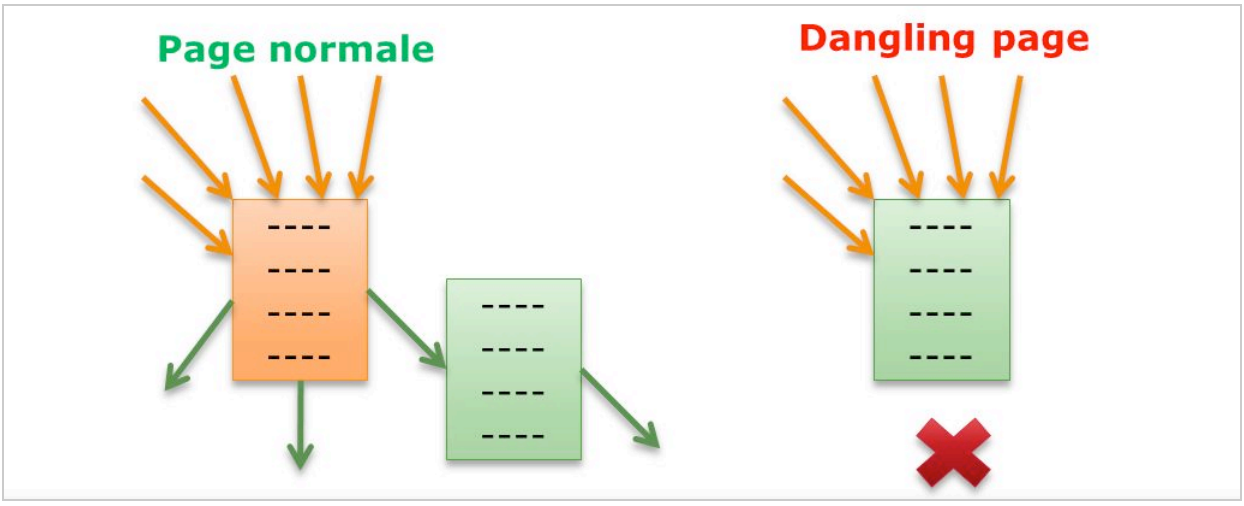
- **PRout** mesure la somme des PR transmis à des pages externes.

- **PRdp** : le PR transmis à des "dangling pages", et nous allons expliquer ce que sont ces pages en détail ci-après.

L'impact funeste des dangling pages pour le référencement

Les dangling pages (les pages "pendantes" en traduction littérale) sont des pages web un peu particulières qui ont pour caractéristiques de recevoir des liens en provenance d'autres pages, mais qui en même temps, ne comportent aucun lien pointant vers une autre page...

De telles pages ont des propriétés étonnantes si l'on considère l'algorithme du PR : logiquement, ces pages reçoivent du PR transmis par d'autres pages, mais ne transmettent pas ce PR à d'autres pages. Par conséquent, si une page est environnée par un grand nombre de "dangling pages", elle verra son PR diminuer !



Une page web « normale » reçoit des liens en provenance d'autre pages, et contient des liens pointant vers d'autres pages. Une « dangling page » au contraire ne contient aucun lien hypertexte.

Les dangling pages sont moins rares qu'on ne le croit

Evidemment, si on cherche ce genre de pages sur le web, on se rend compte qu'il est assez rare de trouver des pages qui ne contiennent aucun lien sortant. De plus, un simple lien sur un site vers l'accueil suffit à faire perdre à une page le statut de "dangling page" : le PR reçu par la page est transmis à la home page et "recyclé" en quelque sorte dans l'arborescence du site.

Il est en fait assez rare que les webmasters créent volontairement ce type de pages. Il s'agit souvent :
- soit de pages générées par des erreurs techniques,
- soit des "pop up" qui ne contiennent aucun lien, mais un formulaire par exemple, ou juste un message ou *disclaimer*.

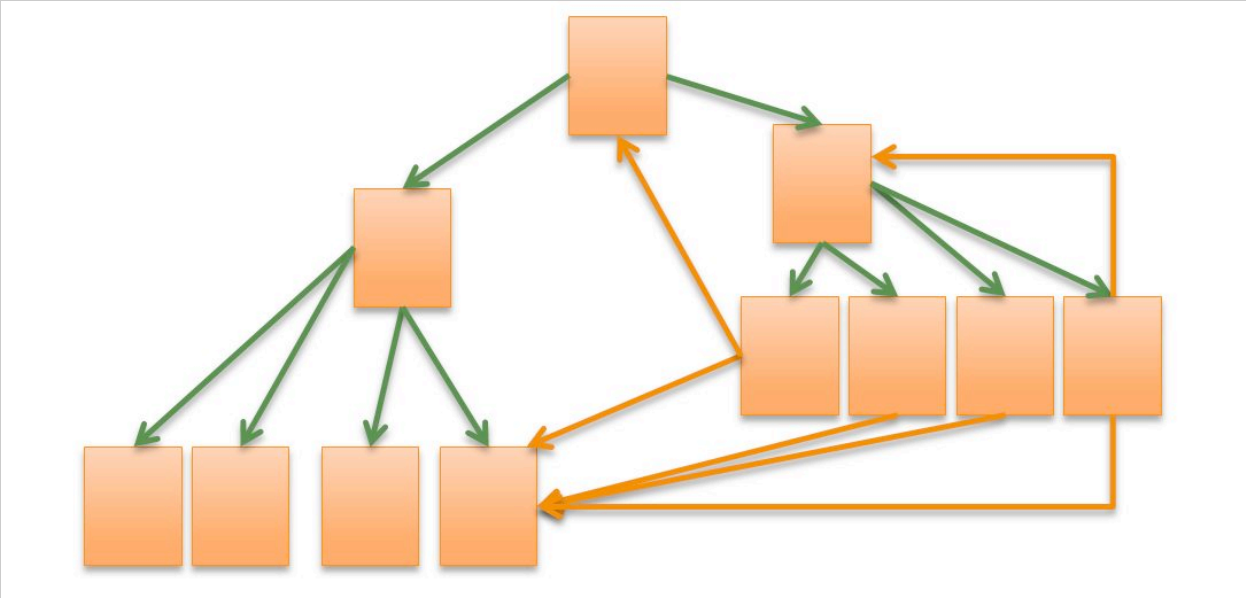
Dans la plupart des cas, ces "dangling pages" accidentelles génèrent des pertes de pagerank assez anecdoctiques à l'échelle d'un site (de l'ordre de quelques %).

Mais il existe un cas où cette perte de pagerank peut théoriquement monter à des niveaux préoccupants : jusqu'à 30 ou 40% du potentiel de PR total du site. Ce cas de figure correspond à des cas où un grand nombre de pages du site sont bloquées par le fichier robots.txt

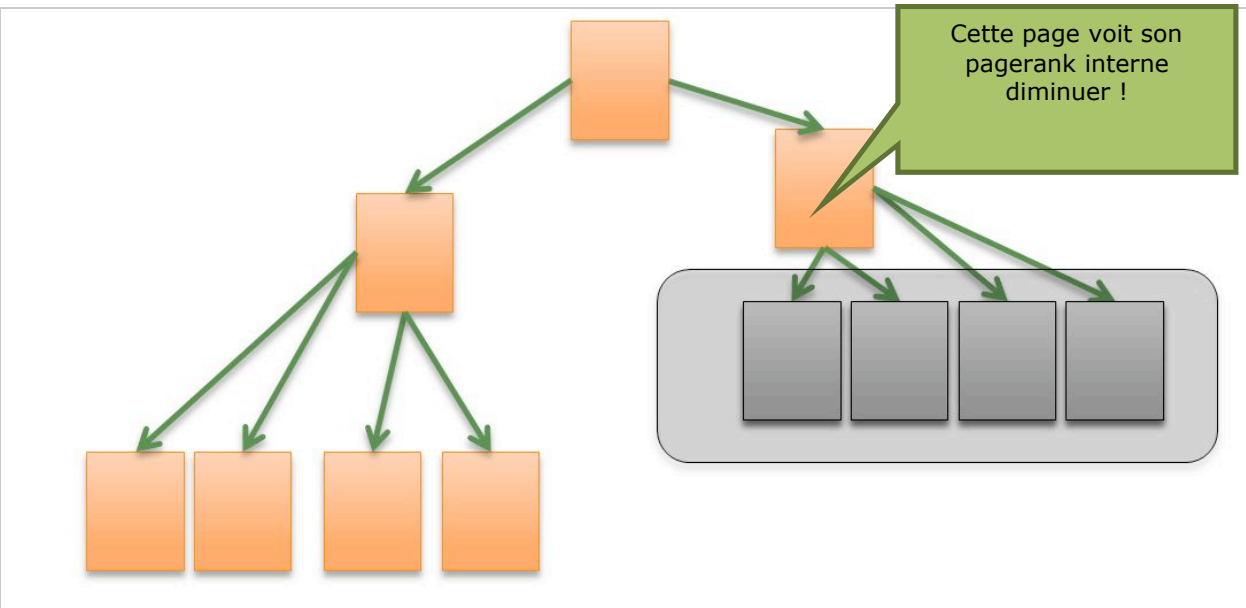
Pourquoi une page bloquée par un fichier robots.txt est-elle une "dangling page" ?

Comme nous l'avons rappelé dans l'article du mois dernier, un moteur de recherche peut connaître l'URL d'une page bloquée par un robots.txt, et même l'indexer. Dès lors qu'il existe des liens entre une page connue de Google et une page dont l'URL est bloquée : ces liens transmettent du PR à l'URL bloquée ! (ce point a été confirmé par Matt Cutts dans une interview réalisée par Eric Enge en 2007) et comme la page ne peut pas être téléchargée, les liens présents sur la page dont l'URL est bloquée ne sont pas découverts par Google, cette URL ne comporte donc aucun lien sortant connu du moteur, et donc deviennent des dangling pages !

Evidemment, si le ratio "nombre de pages bloquées / nombre de pages total" du site est faible, une fois de plus la perte de potentiel de PR induite est faible. Par contre, si, comme on le constate parfois, le nombre d'URL bloqués est égal ou supérieur aux URL crawlables, alors on arrive à des impacts réellement importants.



La situation avant : les pages situées dans la partie droite de l'arborescence reçoivent des liens (et donc du linkjuice depuis la home et la page rubrique. Ce pagerank se diffuse normalement, et est renvoyé en partie vers la home page, la page de rubrique ainsi que les autres parties du site.



La situation après le blocage des 4 pages grisées par un robots.txt : les urls de ces pages ne peuvent plus être téléchargées par le moteur, leur contenu est inconnu, et donc en particulier les liens figurant sur les page colorées en gris. Pour le moteur de recherche, ces pages sont de fait devenus des dangling pages : elles sont reliées au reste du site par la page de rubrique, mais le linkjuice qui leur est transmis n'est pas reçyclé ailleurs dans l'arborescence et se retrouve donc « perdu ». La page de rubrique voit son PR diminuer sensiblement, et le potentiel de PR du site diminue également.

Comment corriger le problème ?

Selon la motivation qui a présidé au "blocage" des URL, on procèdera différemment :

- soit il faut obfusquer tous les liens pointant vers ces pages (faisable, mais pas forcément durable et pas toujours conforme aux TOS de Google).

- soit on laisse Google crawler ces pages (on retire le disallow) et on place éventuellement une valeur d'attribut à "noindex" sur ces pages *via* la balise meta robots ou la directive x-robots-tag, si ces pages sont de mauvaises qualité selon les critères de Google.

- si les URL ont été bloquées pour empêcher leur indexation : une meta robots avec la balise "noindex" est toujours plus appropriée (cf l'article du mois dernier).

- si les URL sont bloquées parce qu'on ne veut pas que Google les crawle : dans ce cas, une directive *disallow* dans le robots.txt reste pertinente, mais il convient dans ce cas :
    * soit de supprimer les liens pointant vers ces pages.
    * soit d'empêcher google de découvrir ces liens, en les obfusquant, ou en les appelant en ajax, ou en remplaçant le lien par un formulaire.

Important :
Ces remarques sur les fuites de pagerank n'ont de sens que dans le contexte de sites comportant un grand nombre d'URL, c'est-à-dire des sites pour lequel le terme "|i|" est très important comparé aux termes PRin et PRout dans la formule du pagerank.
Sur de petits sites, le PR interne dépend beaucoup plus des backlinks externes que du maillage, et on peut négliger ces "optimisations" dont l'impact sera probablement faible voire nul.

Conclusion : apprenez à utiliser votre robots.txt à bon escient

Au terme de ces deux articles sur le robots.txt, nous avons évoqué tous les cas où le fichier robots.txt est détourné de sa fonction première à des fins de SEO ou d'optimisation. Le problème est que par méconnaissance des mécanismes mis en œuvre et de bonnes pratiques dans le contrôle du crawl et de l'indexation, le remède s'avère souvent pire que le mal.

Le webmaster attentif devra donc s'interroger à chaque fois que son robots.txt "bloque" un grand nombre d'URL qui par ailleurs sont normalement visibles par les internautes et ne font pas partie des répertoires que l'on souhaite "cacher", soit aux internautes (backoffice) soit aux moteurs.

En général, ces blocages d'URL peuvent se révéler inadaptés, voire franchement contreproductifs (mais il existe bien sûr des exceptions).

Cela prouve une fois de plus qu'avant de mettre en application des méthodes de référencement dites "sophistiquées" sur son site, il faut déjà maîtriser les fondamentaux du webmastering... Et qu'à jouer les apprentis sorciers, parfois, on se brûle !

Bibliographie

Formule du potentiel de pagerank :
"Inside Pagerank"
Monica Bianchini, Marco Gori, Franco Scarselli
Université de Sienne (publié en 2005)
http://dl.acm.org/citation.cfm?id=1052938&dl=ACM&coll=DL&CFID=155803582&CFTOKEN=96040722

Interview de Matt Cutts par Eric Enge, 8 octobre 2007
http://www.stonetemple.com/articles/interview-matt-cutts.shtml
Lire les développements sur le robots.txt, et le fait que les pages bloquées reçoivent du pagerank

Deux articles en anglais sur le même sujet :
http://www.bpwrap.com/2009/06/hanging-dangling-web-pages-can-be-pagerank-black-holes/
http://www.seomoz.org/blog/headsmacking-tip-13-dont-accidentally-block-link-juice-with-robotstxt

Philippe YONNET , ***Directeur de l'agence Search-Foresight / Groupe MyMedia. Président de l'association SEO Camp (http://www.seo-camp.org/)***