

# Introduction à la SEO (optimisation de pages Web)

par Guillaume Rossolini ([Tutoriels Web](#))

Date de publication : 13 juin 2006

Dernière mise à jour : 06 juillet 2006

Avoir un site Web est une chose ; permettre que les internautes le trouvent en est une autre.

À ce jour, le positionnement d'un site Web dans les résultats d'une recherche précise est devenu une science à part entière : il convient de longuement l'étudier afin de savoir de quoi il retourne. Cette page est en quelque sorte le sommaire d'une série de tutoriels que j'écrirai sur ce sujet.

Certaines techniques présentées plus bas sont mises en place sur le **Forum Cinéma**.

- I - Problématique
  - I-A - Présentation
  - I-B - Historique
  - I-C - Comprendre les moteurs de recherche
  - I-D - Les techniques
    - Chapeau blanc ("white hat")
    - Chapeau noir ("black hat")
  - I-E - Principes généraux
- II - Glossaire
  - II-A - Acteurs
    - Annuaire (directory)
    - Moteur de recherche (search engine)
  - II-B - Termes
    - Backlink
    - Robot / bot / Web crawler / spider
    - SEO (Search Engine Optimization, optimisation pour les moteurs de recherche)
    - Spamdexing
  - II-C - Algorithmes
    - PageRank (Google)
    - TrustRank (Yahoo! Search)
- III - Liens
  - III-A - Tutoriels
    - III-A-1 - La réécriture de liens
    - III-A-2 - Faire évoluer sa réécriture de liens
    - III-A-3 - Bonnes pratiques de SEO ('white hat techniques')
  - III-D - Liens externes

## I - Problématique

Faire connaître son site Web n'est plus suffisant. Il faut aujourd'hui se focaliser sur chaque thème abordé par un site, chacune des pages ayant son propre contexte.

Le référencement est une bataille de tous les instants dans laquelle ceux qui font attention aux moindres détails prennent de l'avance sur ceux qui laissent leur site vivre sa vie et sur ceux qui, au contraire, s'en occupent de trop.

Il ne s'agit pas de simplement remplir les balises META de chaque page HTML. Il s'agit de techniques bien plus avancées, plus subtiles, qui sont même parfois hors de notre contrôle.

### I-A - Présentation

À l'origine, il y avait des sites. Ils restaient isolés et n'étaient pas prévus pour le grand public.

Ensuite, les besoins de notoriété se sont faits sentir. C'est ainsi que sont arrivés les annuaires et les moteurs de recherche.

Il me semble fondamental de situer le contexte avant de rentrer dans le vif du sujet.

### I-B - Historique

Avant 1993, il était difficile de trouver du contenu. Le concept de "spider" n'existait pas (du moins, pas en tant que programme automatisé permettant de construire des index aussi complets que ceux qui existent aujourd'hui). Chacun conservait une liste de ses sites favoris et l'échangeait avec les autres internautes. Ce comportement n'a d'ailleurs pas totalement disparu.

Grâce à ce système basé sur la confiance, chacun pouvait se reposer en ce qu'un site affirmait contenir au moyen de la fameuse balise META.

En 1995 sont apparus deux majeurs du référencement que nous connaissons encore aujourd'hui : **Yahoo!** (Yet Another Hierarchical Officious Oracle, "*encore un oracle officieux et hiérarchique*") et **AltaVista** ("*vue de haut*", dont les créateurs travaillent maintenant pour Google). Yahoo! organisait les sites selon des thématiques hiérarchisées, tandis qu'AltaVista préférait prendre une approche de recherche.

À cette époque, le référencement était fondé sur les informations que chaque webmestre pouvait renseigner dans ses propres pages (balise META). Le propriétaire avait donc un contrôle total, permettant de donner des indications arbitraires. C'est un héritage de l'ère précédente. Les webmestres commencèrent ainsi à donner de fausses informations de manière à attirer davantage de visiteurs.

**Google** est arrivé en 1998 avec une nouvelle méthode de référencement : le **PageRank**, fondé principalement sur les **backlinks**. L'idée était de trouver un moyen qui permette de déterminer la popularité d'une page Web et qui soit le critère principal pour classer les résultats d'une recherche.

Le résultat est explosif, le succès fulgurant. Aujourd'hui, Google est incontestablement le moteur de recherche le plus utilisé mais dont les prédécesseurs ne se laissent pas abattre.

En 2001, Google atteint une popularité sans précédent, juste après que les moteurs de recherche abandonnent les balises META comme critères (cet abandon a lieu au début du siècle).

En 2004 sont lancés les deux concurrents majeurs actuels de Google Search : Yahoo! Search et MSN Search (Microsoft).

## I-C - Comprendre les moteurs de recherche

### **Le référencement est constitué de trois étapes :**

- 1 Web crawling ("parcours du Web")
- 2 Indexing ("mise à l'index du contenu")
- 3 Searching ("recherche")

Je vous propose, afin de parfaitement comprendre comment fonctionne un spider (la première étape), de faire le nôtre. N'ayons pas peur, ce n'est pas très complexe.

Notre robot se contentera de lire une page Web et d'en donner la liste des images, du contenu et des liens ; il suivra quelques-uns de ces liens afin de nous donner des statistiques sur les pages liées, et ce sur quelques niveaux. Un véritable spider devra aller bien plus loin (reproduire cette opération à l'infini) mais nous ne disposons pas d'une puissance de calcul phénoménale...

# parcourtPage

- lit la page
- compte les caractères du texte
- compte les mots du texte
- compte la taille du code source
- compte les images
- compte les liens

présence de liens & niveau d'imbrication maximum non atteint

vrai


faux


- retourne



trouve un lien

### Squelette du crawler

 Télécharger le script (écrit en PHP) : [ <ftp://ftp-developpez.com/g-rossolini/tutoriels/seo/fichiers/spider.zip> ] ou [ <http://g-rossolini.developpez.com/tutoriels/seo/fichiers/spider.zip> ]

 Je ne mets pas ce script en démonstration car il est très gourmand en bande passante.

Cet exemple a plusieurs objectifs : d'une part, vous démontrer que le spider ne peut pas analyser autre chose que du texte (adieu les images et les animations en Flash) ; d'autre part, vous sensibiliser aux éléments qui ont de l'importance dans une page (attributs "alt" des images et "title" des liens, par exemple : je les ai mis en gras quand ils sont disponibles).

L'étape de mise à l'index comprend une analyse complète du contenu, des liens, du code, etc. C'est ici que l'algorithme entre en jeu et que la plus grosse partie des calculs sont effectués.

La recherche est une étape relativement simple : il s'agit simplement de trouver (dans l'index) les pages qui correspondent aux termes recherchés, puis de les classer.

## I-D - Les techniques

### Chapeau blanc ("white hat")

Ce sont les méthodes honnêtes.

Il s'agit des méthodes de SEO permettant simplement de suivre les conseils des moteurs de recherche. Cela vise simplement à construire des sites au contenu utile et correctement mis en valeur.

#### Exemples :

- Sélectionner les mots clefs avec grand soin
- Ne pas trop diversifier les thèmes traités par un même site Web
- Utiliser du code HTML correct
- *etc.*

Je ne souhaite pas donner davantage de détails ici puisque cela fera l'objet de divers tutoriels séparés.

### Chapeau noir ("black hat")

Ce sont les méthodes manipulatrices de résultats. Je vous les déconseille car elles ne sont pas pérennes ; de plus, elles sont éthiquement incorrectes.

Il s'agit des méthodes permettant de manipuler les résultats de moteurs de recherche en utilisant des failles dans les algorithmes des moteurs. Ces techniques peuvent fonctionner mais les moteurs de recherche les combattent activement, ce qui laisse penser qu'elles deviennent inefficaces (voire pénalisantes) avec le temps.

En février 2006, Google supprimait de son index les sites de BMW Allemagne et de Ricoh Allemagne pour avoir utilisé ces techniques. Les sites en question ont évidemment remédié à la situation dans des délais très brefs.

## Exemples :

- Spamdexing : parvenir à tromper l'algorithme du moteur de recherche pour que le site reçoive davantage d'audience qu'il le mérite
- Cloaking : fournir au moteur de recherche une version différente du site par rapport à ce que voient les visiteurs
- Link farms : construire un réseau de sites qui s'échangent des liens, de manière à augmenter leur quantité de backlinks
- *etc.*

Je ne souhaite pas donner davantage de détails ici puisque je n'adhère pas à ces techniques.

## I-E - Principes généraux

Pour optimiser ses pages Web, il suffit d'être le plus honnête possible.

### Voici quelques éléments :

- Construisez des pages au contenu conséquent : ayez du volume sans pour autant faire dans la longueur
- Organisez votre contenu : structure du site, arborescence
- Rédigez correctement : orthographe, grammaire, etc.
- Mettez en forme : titre, gras, italique, etc.
- Pensez à mettre une balise **<h1></h1>** dans chaque page, à renseigner la balise **<title></title>**...

Vous remarquez le point central : **le contenu**. Tout s'applique à mettre en valeur le contenu de votre page. Pourquoi ? Simplement parce que c'est ce que le visiteur cherche dans le moteur de recherche. Il veut une réponse à une question, donc du contenu. C'est à cela qu'il faut penser quand vous optimisez : sélectionnez les mots clefs que le visiteur devra pouvoir trouver dans vos pages et brodez à partir de cela. Vos liens doivent contenir un texte d'ancrage faisant référence à ces mots ou à des synonymes ; les mots eux-mêmes doivent être situés à des endroits stratégiques (titres, en début de page, mis en forme, etc.)...

## II - Glossaire

### II-A - Acteurs

Nous ne nous intéresserons pas tant aux annuaires qu'aux moteurs de recherche. Voici néanmoins une courte présentation des deux acteurs du référencement sur le Web :


#### Annuaire (directory)


L'organisation des sites s'y fait **au moyen d'une hiérarchie thématique**.

L'inscription d'un site auprès d'un annuaire est faite par quelqu'un (le webmestre, un utilisateur ou bien un programme automatisé). Certains annuaires se spécialisent dans un thème précis (la recherche, la cuisine, etc.) tandis que d'autres préfèrent gérer toutes les situations.

#### Voici quelques exemples d'annuaires :

- <http://www.webdirectory.com/>
- <http://search.yahoo.com/dir>
- <http://directory.google.com/>

 *Ces annuaires disposent d'un moteur de recherche interne par mots clefs mais cela ne permet pas de les considérer comme des moteurs de recherche (cf. définition suivante).*

 *Globalement, la plupart des sites communautaires (dont les fournisseurs d'accès à Internet font partie au travers des pages perso de leurs abonnés) proposent une organisation hiérarchique sur ce principe.*

#### Moteur de recherche (search engine)

L'organisation des sites s'y fait **au moyen de mots clefs**.

L'inscription d'un site auprès d'un moteur de recherche est automatiquement faite par le robot du moteur en question.

#### Voici quelques exemples de moteurs de recherche :

- <http://www.altavista.com/>
- <http://search.yahoo.com/web>
- <http://www.google.com/>
- <http://search.msn.com/>

### II-B - Termes

#### Backlink

C'est la fondation du PageRank de Google. Un backlink est un lien pointant vers la page que nous souhaitons optimiser. L'idée générale est que chaque lien vers une page est considéré comme un vote en sa faveur. Bien

entendu, c'est plus complexe que cela en réalité : Google utilise un grand nombre de paramètres pour déterminer la valeur (le poids) de chaque lien.

## Robot / bot / Web crawler / spider

Tous ces termes sont équivalents.

Ils désignent des programmes ou des scripts exécutés par les serveurs des moteurs de recherche. Les spiders (c'est mon terme favori) lisent le code source des pages Web (il s'agit généralement de HTML mais ce n'est pas toujours le cas), en analysent le contenu, en extraient les hyperliens et recommencent l'opération (sur chacun de ces hyperliens) afin de parcourir la Toile peu à peu.

### Une note de sémantique :

- "web" signifie "toile" (notion de réseau comme pour les toiles d'araignées)
- "spider" signifie "araignée"
- "to crawl" signifie "ramper"

## SEO (Search Engine Optimization, optimisation pour les moteurs de recherche)

Ce sont toutes les techniques permettant d'améliorer le positionnement d'une page Web dans les résultats d'une recherche. Attention, il ne s'agit que d'optimisation (amélioration) : les techniques artificielles (factices) sont dénommées *spamdexing*.

## Spamdexing

C'est de la SEO qui a été poussée trop loin : l'optimisation est devenue mensonge.

### Il peut s'agir de :

- propriétés CSS permettant de cacher de grandes quantités de texte et/ou de le positionner hors de l'écran (le visiteur ne voit pas ce texte mais le spider le repère et l'analyse, ce qui permet d'ajouter sournoisement des mots clefs à une page)
- fermes de liens (augmenter le nombre de backlinks)
- etc.

## II-C - Algorithmes

Il s'agit des divers algorithmes permettant de classer les pages Web selon leur fiabilité. Ce classement (propre à chaque moteur) est consulté à chaque fois qu'un internaute effectue une recherche.

Chaque moteur a son algorithme et les détails en sont secrètement gardés.

## PageRank (Google)

Le principe de base du **PageRank** est extrêmement simple : un lien représente, dans la majeure partie des cas, une volonté de son auteur de montrer à ses visiteurs une page qu'il estime intéressante. Plus il y a de liens vers une page précise, plus cela signifie que des gens l'ont trouvée intéressante. Les backlinks sont donc le cœur du PageRank ;

Google calcule ensuite une multitude de paramètres (plus de 200, semble-t-il) afin de déterminer quelle valeur (quel poids) peut avoir un backlink : c'est ce qui est gardé secret.

Tout est automatisé : c'est d'ailleurs le problème du PageRank. Google y tient fermement et lutte activement contre le *spamdexing*.

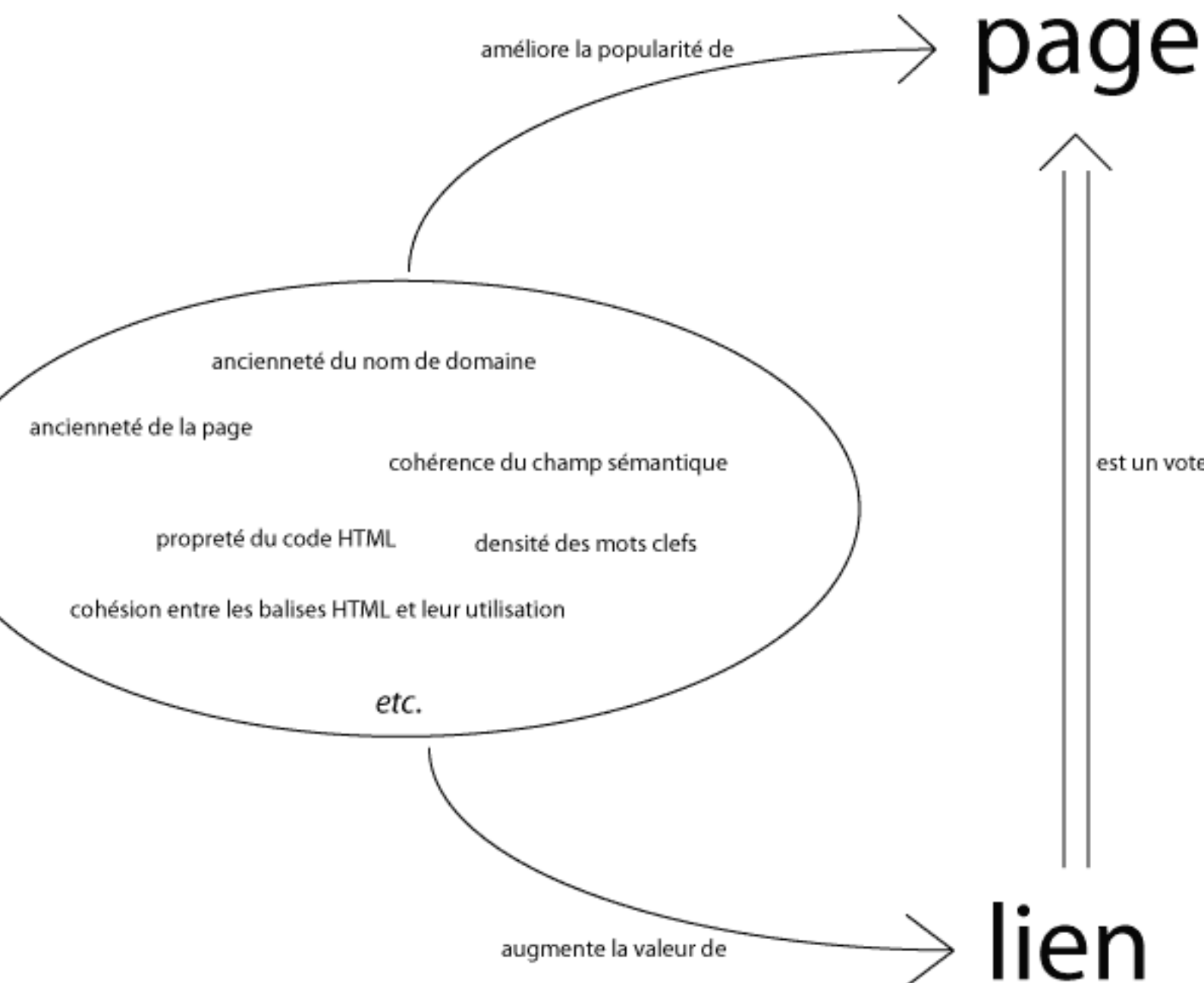



Schéma simplifié du PageRank de Google

Ce schéma illustre deux choses : d'une part, l'objectif du PageRank est de donner une sorte de note à une page Web ; d'autre part, cette note dépend de paramètres propres à la page ainsi que de la note attribuée aux backlinks de la page. Cela signifie qu'un grand nombre de backlinks permet plus facilement d'obtenir un classement correct.

La valeur du vote d'un backlink pour sa page de destination dépend directement du classement de la page hôte de ce backlink ainsi que de la manière dont il est formé.

## TrustRank (Yahoo! Search)

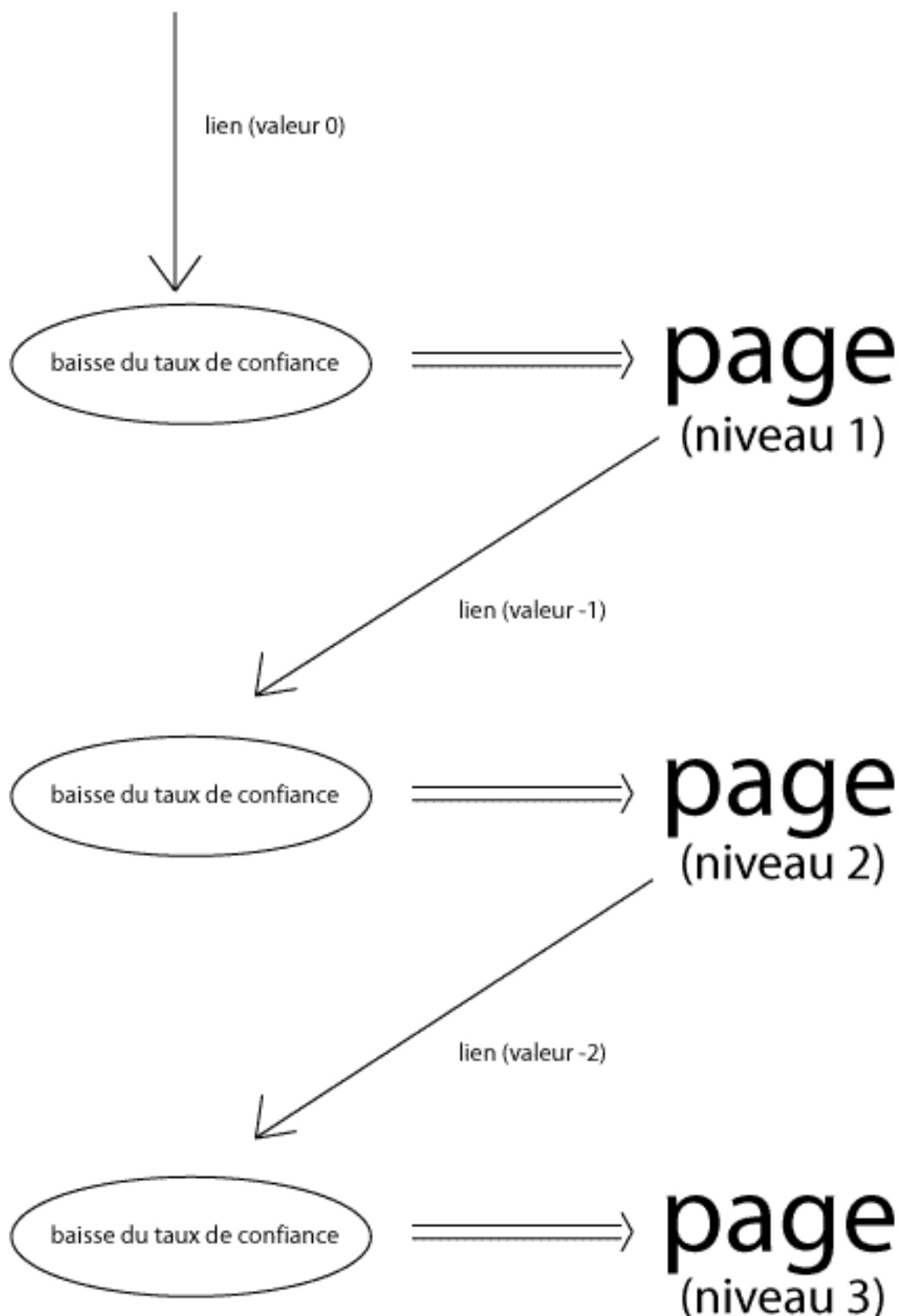
Le cas du **TrustRank** est différent : il s'agit également de compter les backlinks mais leur valeur dépend principalement de l'éloignement à une page source. Le postulat est le suivant : nous examinons attentivement une page à la main ; toute la suite repose sur la confiance que nous accorderons à cette page ; les liens provenant d'une page digne de confiance sont, *a priori*, dignes d'une confiance similaire ; au fur et à mesure que les liens nous éloignent de la page d'origine, nous ne pouvons plus leur accorder autant de confiance.

 Selon René Descartes, toute chose provient d'une chose au moins aussi parfaite. On peut remonter l'arbre généalogique de quelque chose et ainsi découvrir des choses de plus en plus parfaites.

*Le TrustRank fonctionne sur ce principe : prendre une chose parfaite comme point de départ et s'en éloigner peu à peu. Chaque génération perd en perfection, en crédibilité.*

# page de référence

(taux de confiance évalué par un humain)



*Schéma simplifié du TrustRank de Yahoo!*

Le TrustRank accorde de moins en moins de valeur aux backlinks au fur et à mesure que leur site hôte est éloigné du site de référence.

Bien entendu, une même page peut avoir plusieurs backlinks : cela lui permet d'améliorer sa popularité.

## III - Liens

### III-A - Tutoriels

#### III-A-1 - La réécriture de liens

Pour diverses raisons (optimisation de site, faciliter la mémorisation des liens, cloaking, etc.), il peut être souhaitable de modifier la forme que prennent les liens d'un site Internet, sans pour autant changer toute la structure des pages physiques.

C'est ce que permet la réécriture dynamique de liens, alias URL Rewriting. J'ai mis en place cette technique sur le Forum cinéma RNZ.

 Aller au  [tutoriel d'URL Rewriting \(réécriture de liens\)](#)

#### III-A-2 - Faire évoluer sa réécriture de liens

Ce tutoriel explique comment il est possible de changer d'URL Rewriting (ou bien simplement le mettre en place) sans perdre le référencement auprès des moteurs de recherche.

Attention, il n'est pas question ici d'expliquer comment changer de méthode de réécriture de liens. Ce tutoriel se contentera de vous sensibiliser aux problèmes posés par un tel changement et vous apportera les solutions qui conviennent.

 Aller au  [tutoriel de modification d'URL Rewriting](#)






#### III-A-3 - Bonnes pratiques de SEO ('white hat techniques')

Le référencement est une sorte de science complexe. Les moteurs de recherche ne donnent pas toutes les précisions sur le fonctionnement de leurs algorithmes, ce qui laisse aux spécialistes du référencement la charge de deviner la plupart des paramètres.

Les techniques approuvées par les moteurs de recherche portent généralement le nom de "white hat" (soit "chapeau blanc", une référence à l'honnêteté du webmestre en question).

 Aller au  [tutoriel de bonnes pratiques de référencement](#)

## III-D - Liens externes

-  [Wikipedia : Search Engine Optimization](#)
-  [Wikipedia : List of search engines](#)
-  [Wikipedia : Optimisation pour les moteurs de recherche](#) (largement incomplet pour le moment)
-  [Wikipedia : Liste des moteurs de recherche](#)
-  [Google : Centre d'aide Administrateur Web](#)

