Comment les moteurs de recherche détectent-ils le spam dans leurs résultats ?

Retour au sommaire de la lettre

Il est souvent intéressant de se mettre à la place d'un moteur de recherche pour mieux comprendre ses contraintes et pouvoir adapter ses méthodes de travail afin d'obtenir une optimisation de son site qui corresponde aux recommandations de ces outils. Antoine Mussard, de la société VRDCI, explore dans cet article les différentes formes de spamdexing (fraude aux moteurs) utilisées à l'heure actuelle par certains webmasters, avec un focus sur le cloaking, et explique comment les moteurs les combattent et pourquoi il peut être très dangereux de cotoyer ces rivages peu hospitaliers...

Une attitude intéressante à avoir, lorsqu'on désire améliorer son propre référencement, est parfois de se mettre à la place des moteurs de recherche afin de comprendre leur fonctionnement, ce qui permet de plus facilement s'adapter à leurs contraintes. Cela est certainement tout à fait vrai en ce qui concerne la détection du spam (ou spamdexing : fraude sur l'index des moteurs).

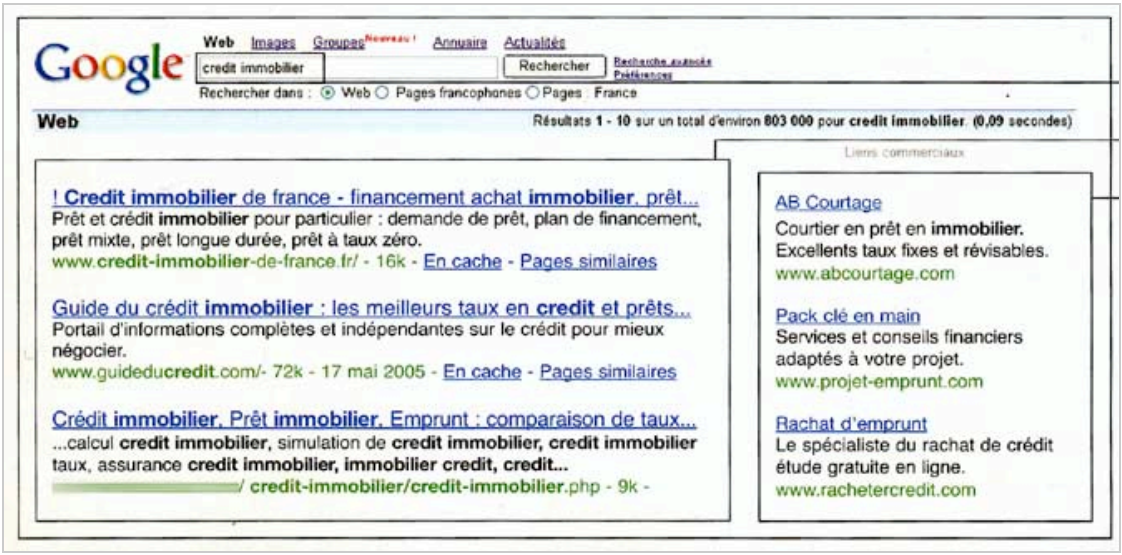
En effet, les moteurs de recherche font actuellement face à des centaines de milliers de webmasters de par le monde qui optimisent, voire sur-optimisent leurs pages de façon plus ou moins « border-line ». Sachant qu'en fait toute optimisation peut être assimilable à du spam, selon le degré d'optimisation mis en place au regard des règles internes de chaque moteur, parfois bien difficiles à appréhender, il faut bien le dire...

Quelques pistes de réflexion

Voici donc quelques pistes que les moteurs de recherche pourraient explorer, ou explorent déjà, afin de détecter les sites sur-optimisés et d'améliorer la qualité de leurs données. Il est important de noter que, bien entendu, tous les moteurs font face sans exceptions à ce "fléau" du spamdexing.

Ces pistes à l'étude par les moteurs ont été vérifiées, pour les plus complexes, avec notre outil de mesure d'audience SeeLog (http://www.seelog.com) grâce à deux de ses fonctions :
- La gestion des robots et des visiteurs autres que les navigateurs.
- La détection des visiteurs. Exemple (simplifié) : Vous saisissez « BNP » ou « Google » dans le back-office client Seelog et vous aurez toutes les connexions de la « BNP » ou de « Google » sur votre site web.

Il est toujours étonnant de voir des sites sur-optimisés qui sont en première page des moteurs de recherche sur des requêtes ultra-concurrentielles alors que le spam est visible très rapidement. Voici par exemple une publicité papier récente de Google laissant apparaitre un troisième résultat naturel qui ressemble fort à du spam (nous avons masqué son URL) :



Etonnant de voir ce type de lien dans une publicité officielle de Google, publicité que nous reproduisons ici pour information :

Publi-information

Les liens commerciaux AdWords
Trouver des clients avec Internet

Google

Le problème avec la publicité, c'est que l'on sait toujours combien ça coûte, mais jamais combien ça rapporte. Heureusement, Google a créé pour vous les liens commerciaux AdWords. Désormais, votre publicité ne vous coûte que si elle vous rapporte. AdWords : ce sont encore nos clients qui en parlent le mieux...

Avec AdWords, mon portefeuille clients se développe !

Ari Bitton
Directeur Associé
www.abcourtage.com
Financement immobilier
Paris

Pourquoi utilisez-vous les liens commerciaux AdWords ?

C'est le moyen le plus efficace de toucher à moindre coût des prospects à fort potentiel. L'internaute qui tape "crédit immobilier" sur Google est aussitôt exposé à notre message qui apparaît automatiquement à côté des résultats de sa recherche. D'un clic, il peut aller sur notre site pour effectuer des simulations de crédit. Autre avantage : à tout moment, je peux gratuitement changer le texte de mon annonce pour coller à l'actualité. Quel autre support permet une telle réactivité ?

Quel budget accordez-vous à la publicité sur Google ?

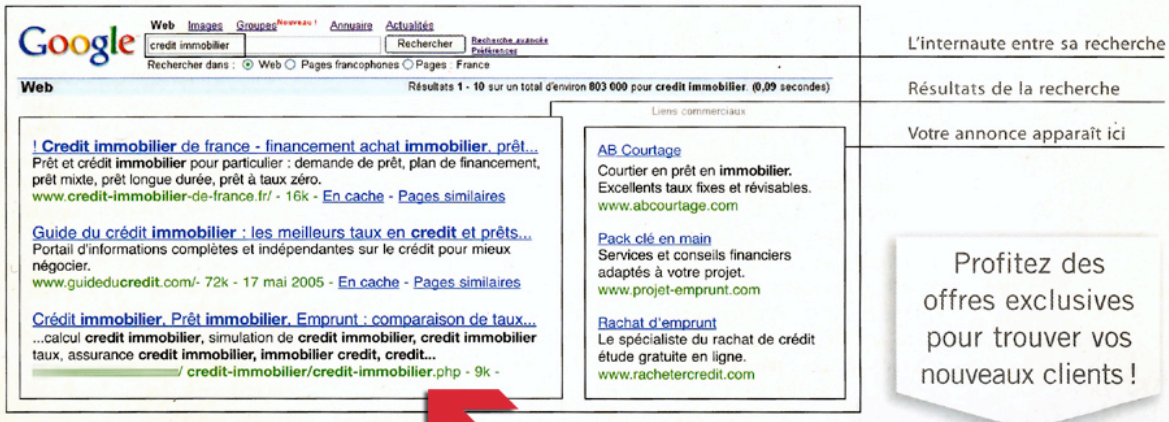
La publicité sur Google ne représente que 7 % de notre budget communication. C'est donc très abordable et j'ajouterais : très équitable. C'est le seul média que je connaisse où l'on ne paye qu'en fonction des résultats, c'est-à-dire quand le prospect clique effectivement sur notre annonce.

Quels résultats obtenez-vous ?

Excellents : 60 % des visites sur notre site sont générées exclusivement par Google. De plus, il s'agit d'une audience très qualifiée puisqu'elle génère 15 % de notre chiffre d'affaires. En clair, un contrat est signé chaque jour grâce à Google AdWords !

Quels conseils donneriez-vous à d'autres utilisateurs des liens commerciaux ?

Je leur dirais de bien choisir les mots-clés sur lesquels faire apparaître leur annonce. Exemple : le mot "crédit" est plus recherché par les internautes que le mot "emprunt". En affinant ainsi, on peut doubler les résultats d'une campagne.



Profitez des offres exclusives pour trouver vos nouveaux clients !

OFFERT SANS ENGAGEMENT

Le guide des meilleures pratiques pour gagner des clients avec Internet.

OFFRE D'ESSAI ADWORDS
50€ OFFERTS*
pour démarrer votre première campagne

Pour bénéficier de ces offres, rendez-vous sur : www.google.fr/resultat3

Pour un moteur de recherche, il n'y a pas pléthore de méthodes permettant de détecter un tricheur. La première chose à vérifier est le cache. En effet, un site web qui refusera d'afficher son cache pourra d'emblée être considéré comme suspect (ce qui n'est pas synonyme de coupable, notez-le bien) car il ne désire pas que les internautes voient la version de ses pages qu'il a fourni aux moteurs de recherche.

Les différentes méthodes de spam pourraient être regroupées en 4 familles : répétition abusive de mots, texte caché, développement artificiels de liens, cloaking. Nous nous attarderons surtout sur ce dernier point tout en présentant succinctement les autres.

Détection des abus de densité de mots clés

Il n'est pas naturel pour un site web de répéter énormément de fois un même mot clé dans un document, ne serait-ce que par souci rédactionnel. Analyser une page d'accueil est la plupart du temps suffisant pour détecter un site trop optimisé.

La première étape consiste à découper dans un tableau les zones contigües de texte pour obtenir tous les blocs de textes. En analysant les zones suffisamment denses, y trouver une répétition anormale de mots clés sera assez aisé et parfois assimilable à du spam. Les moteurs se basent actuellement sur l'indice de densité (nombre d'occurrences du mot dans la page divisé par le nombre total de mots) ce qui n'est pas à notre sens suffisant car il est très facile de rédiger du texte qui ne sera pas lu, rempli de mots clés. Allez consulter par exemple le cache des pages Google sur des mots clés ultras concurrentiels pour en être convaincu... Répéter plus de dix fois une expression-clé dans une page "classique" n'est en règle générale pas naturel.

Voici un exemple de méthode pour détecter les mots clés trop denses :
- Ne pas prendre en compte la balise Meta Keyword (comme Google le fait déjà).
- Détecter les titres non naturels (ceux qui se présentent sous la forme d'une succession de mots clés).
- Détecter les mots clés ayant plus de dix occurrences dans la même page ou un indice de densité trop important (par exemple, supérieur à 5%).
- Prévoir *in fine* une intervention humaine afin de vérifier la page douteuse et prendre une éventuelle décision de sanction.

Détection des textes cachés

Pour les textes cachés, la tâche est plus compliquée car si ces derniers ne font pas l'objet d'abus de densité, il faudra alors déterminer s'ils sont ou non du spam, ce qui n'est pas aisé de façon automatique.

S'il est facile de détecter du texte ayant la même couleur ou une couleur proche de la couleur de fond d'un site web (méthode assez préhistorique, il faut bien le dire), il n'en est pas de même pour les autres façons de cacher du texte dans une page HTML.

En effet, du texte peut être caché pour des raisons tout à fait valables notamment dans des *layers*. C'est par exemple le cas dans de nombreux sites qui se prévalent du label "Web 2.0"... De même, les fonctionnalités W3C d'accessibilité peuvent permettre de cacher du texte.

Le moyen ultime, mais très lourd en ressources, serait de se baser sur le texte réellement affiché sur le navigateur en lançant un script automatique qui va récupérer, par l'intermédiaire d'un navigateur comme Firefox, le texte affiché de la page scannée et de le comparer avec le texte "aspiré" par le robot du moteur. S'il diffère beaucoup, une intervention humaine visuelle sera nécessaire pour comprendre pourquoi et vérifier qu'il s'agit ou non de spam.

Le moteur de recherche devra donc disposer de ressources importantes pour vérifier visuellement ces textes cachés en utilisant diverses IP qui ne lui sont pas liées.

Détection des barres de liens et des faux liens

Dans le cas où, comme pour Google, l'algorithme donnerait un poids important à l'analyse des liens internes et externes, la tentation est grande de mettre en place ce type de lien de façon quasi "industrielle".

Solution relativement simple : détecter des enchaînements consécutifs de plus de 3 liens avec un contenu textuel contenant des mots clés (sans texte de présentation). Ce sera une bonne façon de détecter d'éventuels liens "non naturels".

Exemples de liens non naturels :

référencement, achat appartement, immobilier var (ils sont LEGION sur l'Internet !)) :

Exemples de liens naturels :

- Lien + texte de description (y compris avec un mot clé pour le lien).
Exemple : Référencer votre site web en lisant Abondance

- Texte de description + Lien + texte de description (y compris avec un mot clé pour le lien).
Exemple : pour aller encore plus loin, contactez un SEO ou un spécialiste du référencement

- Lien avec du texte (ne correspond pas à un mot clé couramment recherché, tous les moteurs disposent de ce genre d'information).
Exemple : Abondance dévoile un secret bien gardé, SCOOP !

Il convient de ne pas prendre en compte les liens non naturels mais bien entendu, également de ne pas pénaliser les sites cibles. De plus, Un moteur devra prendre en compte tout de même les pages liens présentant des URL avec des descriptions. Tout un art...

Détection de méthodes « avancées » : détection de Cloaking

Le cloaking consiste à détecter le type de visiteur qui se connecte sur un site web et à afficher du texte différent pour les visiteurs "humains" et pour les robots des moteurs de recherche. Il peut être diaboliquement efficace mais reste cependant assez facile à détecter pour un moteur. Il est aujourd'hui totalement interdit (blacklistage quasi assuré après détection). On peut être d'accord ou non avec cette vision du cloaking (qui serait pourtant un remède intéressant pour le référencement de sites web "à problèmes" comme le Flash), mais les moteurs de recherche ont aujourd'hui tranché, certainement au vu de certaines pratiques détectées sur certains sites : l'utilisation du cloaking est considéré comme du spamdexing, point barre... La communication, notamment de la part de Google et de Yahoo!, est très claire sur ce point depuis de nombreux mois.

Il existe trois types de cloaking (en fait quatre, comme nous le verrons plus tard) : le cloaking par agent ("Googlebot" pour Google, "Slurp" pour Yahoo!...), par adresse IP et par hôte. Il existe sur le Web d'innombrables sources d'informations fournissant moults données de ce type sur les spiders des moteurs de recherche, des plus connus aux plus obscurs...

Exemples :
Quelques adresses IP de Googlebot, le robot de Google :
http://www.robots.darkseoteam.com/adresses-ip-googlebot.php
Quelques user-agents de GoogleBot :
http://www.robots.darkseoteam.com/user-agent-googlebot.php
Quelques hôtes de Googlebot :
http://www.robots.darkseoteam.com/hotes-googlebot.php

Le cloaking nécessite un développement avec un langage dynamique tel que PHP/ASP/JSP/PERL. Il consiste pour l'éditeur d'un site, à analyser en temps réel qui se connecte sur le serveur, *via* l'agent, son adresse ip ou son hôte et ainsi détecter un éventuel moteur de recherche et lui afficher un texte optimisé/spammé.

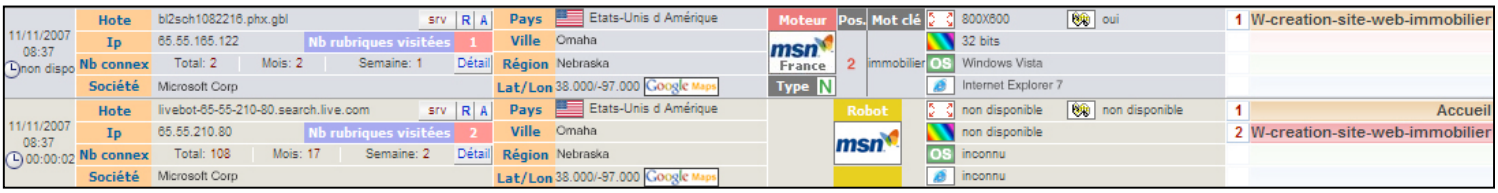
Tous les moteurs de recherches se sont fait "berner" par le passé par ce type d'action. Le premier à réagir a été Google il y a 5 ans en commençant à blacklister de façon manuelle après vérification

les sites tricheurs, le plus souvent suite à une dénonciation. De très nombreux sites français ont alors été blacklistés dont un constructeur automobile très connu (comme BMW en Allemagne).

Une méthode simple et donc infaillible à mettre en place pour les moteurs est donc d'utiliser pour leurs spiders une adresse IP n'étant pas reliée aux dits moteurs et qui va simuler de façon automatique une connexion naturelle (du cloaking inversé en quelque sorte !) à un site web en parallèle d'une connexion émanant "officiellement" du moteur de recherche. La comparaison des deux résultats indiquera ensuite si une procédure de cloaking a été mise en place par le webmaster.

Le seul moteur actuellement à procéder à ce type de vérification de manière régulière est, à notre connaissance, MsnBot/LiveBot, le robot du moteur de recherche Live Search de Microsoft. Cependant, ce dernier n'est pas très efficace car il utilise une adresse IP de son réseau interne. Les critères de filtrage de notre outil de Statistiques SeeLog nous permettent donc de ressortir toutes les connexions en filtrant les résultats sur la société « Microsoft Corp ».

Voici donc deux connexions du moteur de recherche qui ont été simultanées (et qui se sont reproduites des centaines de fois) :



Cette illustration montre que MSN/Live s'est connecté simultanément sur un de nos sites web avec deux IP différentes, l'une classique du Robot MSN Live et l'autre avec une des IP de leur système d'information interne et ce de façon automatique (car reproduites des centaines de fois via diverses IP). MSN simule donc dans cet exemple une fausse requête sur Live en générant dynamiquement le faux référent suivant :
http://search.live.com/results.aspx?q=immobilier&mrt=en-us&FORM=LIVSOP

De plus il gère la génération d'information Javascript comme un navigateur avec les données suivantes en Agent :
Agent : Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.2; .NET CLR 1.1.4322)

Et supporte même la gestion de la résolution, ce qui signifie un aller-retour d'échange avec leur développement dynamique. Le simulation dynamique en PHP/ASP/JSP/PERL de ce type de comportement est assez simple à développer.

On remarquera que l'analyse de Microsoft est cependant assez fine puisqu'ils arrivent à simuler du JavaScript en dynamique et donc à simuler une résolution de 800*600. En tout cas MSN/Live est le premier à instituer la vérification en temps réel du cloaking, selon nos informations...

Cette détection laisse envisager un 4ème type de cloaking possible : le cloaking par société propriétaire d'IP, qui consisterait à afficher du texte différent dès que l'on détecte une adresse IP appartenant à (i.e. "utilisée par") un moteur (Google, MSN, Voila, etc.).

Peu répandue, assez complexe à mettre en œuvre, cette méthode risque cependant de se développer rapidement. Voici donc une méthode pour contrer tous les "cloakers" en puissance : cette procédure devrait permettre aux moteurs de recherche de détecter tous les cloakers à 100% (et surtout ceux qui désactivent leur cache sur Google pour ne pas se faire remarquer par les internautes et échapper aux dénonciations *via* l'interface "Google Outils pour Webmaster") :

- Disposer d'IP dans Plusieurs pays.
- Ne pas acheter ces IP avec la raison sociale de la société mère :
  - Pour Yahoo : différent de Inktomi Corporation
  - Pour MSN/Live : différent de Microsoft Corp
  - Pour Google : différent de Google Inc.
  - Pour Baidu : différent de Baidu Kabushiki Gaisha
  - Pour Exalead : différent de EXALEAD

Pour Voila : différent de France Telecom
...

- Vérifier dans un court intervalle et un intervalle moyen sur plusieurs jours (au moins 5) que le contenu textuel ne change pas entre les scans du robot et les scan des IP en simulant un navigateur gérant le JavaScript (à programmer avec par exemple fsockopen en PHP). Faire ces scans lors de période d'inactivité du site s'il y en a.

Aller plus loin

Un autre type de spam qui n'est pas géré par les moteurs de recherches : les sites satellites n'ayant pas le même contenu de texte (pour éviter le phénomène de « duplicate content »). Légion sont aujourd'hui les sociétés qui créent de nouveaux sites web en présentant les mêmes activités mais de façon différentes et sur des noms de domaine différents.

Les moteurs pourraient les détecter et pénaliser les sites les plus récents (ce qui n'a rien à voir avec la prétendue "sandbox" de Google). Il arrive souvent qu'un même site soit représenté 4 fois en page d'accueil alors que c'est la même société et les mêmes activités. Etrange, non ?

Voici des pistes pour les détecter : vérifier les numéros de téléphone (comme Google le gère sur Google Maps), analyser les propriétaires des noms de domaine, etc. A ce titre, vous saviez que le fin stratège Google était *Registrar* (possibilité d'avoir accès aux bases de données des noms de domaines) ?! Travailler à détecter les webmasters "border-line" doit être passionnant (n'est-ce pas Matt Cutts !). :-)

Antoine MUSSARD, Dirigeant de VRDCI, "agence web de référencement naturel avec paiement aux résultats"
Site web : http://www.vrdci.com/
E-mail : antoine.mussard@vrdci.com

Réagissez à cet article sur le blog des abonnés d'Abondance :

http://abonnes.abondance.com/blogpro/2007/11/novembre-2007-comment-les-moteurs-de.html