

# Moteurs de Recherche, Cours Master 2, 2011

Michel Habib

habib@liafa.jussieu.fr

<http://www.liafa.jussieu.fr/~habib>

Mars 2011

# Plan

Google aurait-il changé son algorithme ?

Le brevet du classement Google

L'indice Freshrank de Google

Comment optimiser le référencement de votre site

Google a lancé une nouvelle offensive contre les sites "à la demande" ou "fermes à contenu" générés pour répondre à l'intérêt des internautes mais aussi parfois des seuls publicitaires, en annonçant que son moteur de recherche rétrograderait les sites de "mauvaise qualité". "Nous avons lancé une amélioration assez importante de l'algorithme de nos classements - un changement qui touche 11,8% des recherches", ont indiqué jeudi des ingénieurs de Google, Amith Singhal et Matt Cutts, sur le blog officiel du groupe. "Cette actualisation vise à rétrograder les sites de mauvaise qualité - les sites qui apportent peu de valeur pour les internautes, qui copient des contenus d'autres sites, ou qui ne sont tout simplement pas très utiles", ont-ils ajouté. "En même temps, cela fournira un meilleur classement pour les sites de qualité, ceux qui ont des contenus originaux, et des informations comme de la recherche, des études détaillées, des analyses réfléchies".

Cette offensive répond aux critiques d'internautes ayant décelé une dégradation de la qualité des résultats de recherches sur Google, et prolonge une initiative prise le mois dernier. Le groupe californien avait alors annoncé un premier changement de ses calculs pour purger son moteur de recherche des sites internet plus ou moins bidons, dits "webspam". Il avait aussi annoncé son intention d'identifier les "content farms", un terme péjoratif qu'il n'a pas utilisé cette semaine et visant des sites copiant des informations éparées répondant à des questions précises et qui, selon leurs détracteurs, n'ont pour seule raison d'exister que d'attirer de la publicité. Google n'a pas précisé quels sites étaient visés.

A la bourse de New York, l'action du groupe Demand Media, une société qui se plaint d'être traitée de "content farm", a perdu jusqu'à 4% avant de se redresser et de finir en hausse de 1,59% (à 22,96 dollars), après avoir assuré ne pas être concernée. "Il est impossible de spéculer sur l'impact qu'auront les changements", a assuré sur le blog de Demand Media un des dirigeants de la société, Larry Fitzgibbon. "Mais pour le moment nous n'avons pas constaté de gros impact".

Le fonctionnement de Demand Media, comme de certaines filiales des groupes internet Yahoo! et AOL, repose sur l'utilisation d'armées de pigistes, sollicités pour produire des pages sur des sujets recherchés par les internautes, mais peu présents sur le web et à fort potentiel publicitaire : c'est le retour sur investissement publicitaire qui détermine les sujets sur lesquels on écrit ou on publie des vidéos. A terme Google, dont l'initiative a été globalement saluée par les médias, prend le risque de se poser en arbitre de la qualité.

"Les gens n'aiment pas que Google ait autant de pouvoir et de contrôle sur internet", remarque Greg Sterling, un des responsables du site SearchEngineLand, interrogé par l'AFP. "Les contributeurs (des sites 'content farms') ont l'impression d'être dévalorisés". Sur le site Webmasterworld, plusieurs webmestres se sont plaint d'une subite chute de trafic, comme l'internaute "rowtc2". "Cela fait plus de quatre ans que je gère un site, beaucoup d'heures de travail, j'ajoute du contenu et de la valeur, j'obtiens des liens ... et maintenant ... une chute de 29% du trafic (...) c'est plutôt démoralisant!", y écrit-il.

"Google est dans une situation difficile", résume M. Sterling. "Son succès a engendré tout une économie avec des pigistes qui produisent des articles conçus pour susciter de la publicité et bien figurer dans les résultat de recherche". En même temps, "Google est critiqué de toutes parts parce qu'il a trop de spam, et se rend compte que s'il ne résout pas ce problème, c'est son existence qui est menacée".

Essayer la requête :

La nièce de Dior

## Le brevet de 2007

Le brevet décrit les critères qui permettent de classer les pages Web

- ▶ La date du document (en fait celle du premier référencement Google)
- ▶ La fréquence des modifications du contenu \*
- ▶ L'analyse des requêtes et des clics sur les résultats \*
- ▶ La vitesse d'apparition de nouveaux liens pointant sur une page \*
- ▶ Le texte des balises (ou ancres). L'ancienneté du texte est gage de pertinence. (Analyse sémantique du contexte autour de la balise).

## Mais aussi ...

- ▶ Le trafic sur la page
- ▶ Le comportement des visiteurs sur la page (temps passé ...)
- ▶ Le nom de domaine
- ▶ Les classements précédents
- ▶ être ou ne pas être dans des bookmarks
- ▶ Le liens non pertinents (indicateur de spam)
- ▶ Le sujet du document

## Commentaires

- ▶ \* signifie : pas du tout c'est mauvais, un peu c'est bien, trop c'est louche (spam)
- ▶ Une équipe d'une centaine d'ingénieurs qui pondère les paramètres décrits ci-dessus continuellement.
- ▶ Le brevet ne doit donc pas être trop strict (car les autres moteurs font pareil)

## Le brevet Google de 2008

Permet de définir quand une page est devenue obsolète ou une page de référence (par ex. le texte de la déclaration des droits de l'homme de 1789).

- ▶ La date du document (en fait celle du premier référencement Google)
- ▶ La fréquence des modifications du contenu \*
- ▶ L'analyse des requêtes et des clics sur les résultats \*
- ▶ Un indice récursif de **fraîcheur** qui se transmet par les liens
- ▶ Un indice récursif de **confiance** qui se transmet par les liens

## Mais aussi ...

- ▶ Le texte des ancres
- ▶ Le trafic sur la page
- ▶ Le comportement des visiteurs sur la page (temps passé ...)
- ▶ Le nom de domaine
- ▶ Les classements précédents
- ▶ être ou ne pas être dans des bookmarks
- ▶ Mots uniques dans les ancres, apparition de même ancres dans plusieurs pages (indice de spam)
- ▶ Le liens non pertinents (indicateur de spam)
- ▶ Le sujet du document

## Le sujet

- ▶ Catégorisation
- ▶ Analyse des URLs
- ▶ Analyse du contenu
- ▶ Clustering
- ▶ Création d'un sommaire
- ▶ Présence de mots-clés uniques propres au domaine
- ▶ Si le sujet change reclasser le document. Trop de sujets dans la même page peut indiquer du spam.

## Quelques règles simples

- ▶ Vaut mieux garder un vieux nom de domaine bien connu
- ▶ Le titre est important
- ▶ Ne pas oublier de mot-clé du domaine
- ▶ Vérifier sur Google ce que donne ces mots-clés.
- ▶ Ne pas placer de barrières aux robots de référencement
- ▶ Soigner le libellé des balises
- ▶ Ne pas oublier les méta-tags
- ▶ Ne laisser pas de liens morts
- ▶ ... En anglais on parle de **SEO : search engine optimization**