

Révisions Moteurs de Recherche, Cours Master 2, 2011

Michel Habib

habib@liafa.jussieu.fr

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

Mars 2011

Plan

Rappel sur les principes

Fermes de contenus

Le brevet du classement Google

Grosso modo voila le principe :

1. A chaque page on associe un ensemble de mots-clés (lors de l'analyse des pages) on en fait un fichier inversé adressable par mots clés.
2. Dans un autre temps on calcule le PageRank de toutes les pages
3. Vous faites une question Google :
par exemple : la nièce de Dior
4. Google y repère deux mots clés : nièce et Dior
5. il cherche toutes les pages contenant ces deux mots clés (à l'aide du fichier inversé)
6. il affiche les résultats ordonnés selon les valeurs de PageRank (ou de tout autre algorithme de classement)

Fermes de contenus

Automates de l'information

<http://www.monde-diplomatique.fr/2011/03/RAMONET/20221>

Ignazio Ramonet du monde diplomatique s'inquiète de la généralisation des pigistes de l'information et son étape ultime les algorithmes de l'information.

Commentaire de Décembre 2010

Google a reconnu - de façon inhabituelle et publiquement - qu'elle allait modifier son moteur de recherche afin de lui permettre de différencier les bons des mauvais commerces et ajuster en conséquence leur classement dans les résultats.

Le géant de la recherche en ligne a décidé de mettre en oeuvre ces mesures après la publication d'un article dans le New York Times détaillant la tactique d'un opticien en ligne. Le commerçant explique, en se vantant, comment son classement dans le moteur de recherche de Google a grimpé grâce aux nombreuses plaintes postées par des clients mécontents. «J'ai exploité cette possibilité parce que cela fonctionne, » a déclaré au journal Vitaly Borker, le fondateur et propriétaire de DecorMyEyes. « Quel que soit le site où les clients ont publié leurs commentaires négatifs, ils ont contribué à mon retour sur investissement. Alors, autant utiliser ces critiques négatives à mon avantage, » a-t-il ajouté.

La solution simple pour éliminer les fermes de contenus

Dégrader le classement des pages qui affichent trop de pub

Remarque : ce paramètre n'intervient dans le brevet de classement.

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)

La solution simple pour éliminer les fermes de contenus

Dégrader le classement des pages qui affichent trop de pub

Remarque : ce paramètre n'intervient dans le brevet de classement.
Il faut maintenant trouver une façon de formaliser : ne pas contenir trop de pub

Merci pour votre attention !!