Les requêtes QDF et le nouvel algorithme "Freshness Update" de Google

Retour au sommaire de la lettre

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

Depuis début novembre, Google a annoncé un nouvel algorithme appelé "Freshness Update", qui touche 35% des requêtes saisies par l'internaute et qui met en avant les contenus ayant une date de publication récente pour les requêtes d'actualité. Pourtant, cette vision avait déjà été mise en avant par Google depuis 2006 et le concept des QDF (Query Desserves Freshness). Qu'est-ce que cette nouvelle mouture apporte de plus et pourquoi Google l'a-t-il mise en place ? Voici quelques éléments de réponse et de réflexion...

Le 3 novembre dernier, Google annonçait sur son blog (http://actu.abondance.com/2011/11/encore-un-nouvel-algorithme-google-35.html) une mise à jour de leur algorithme, dont l'objectif était déclaré était de "renvoyer des résultats plus frais". L'impact de cette mise à jour s'est révélé extrêmement important : selon les chiffres donnés par Google, 35% des requêtes ont vu les pages de résultats modifiées par ce changement, et 6 à 10% dans des proportions notables.

Cela signifie que cette énième modification de l'algorithme a un impact d'une ampleur comparable à celle de Panda. Elle s'est avérée néanmoins plus "indolore" pour la plupart des webmasters, car contrairement à Panda, il ne s'agissait pas de déclasser des sites, mais juste de changer le classement de certaines pages sur certains types de requêtes. Plusieurs situations se sont donc présentées :
- certains sites mal référencés sur des requêtes impactées n'ont pas vu d'effet notable sur leur trafic.
- la plupart des sites ont vu certaines de leurs pages monter dans les classements, et d'autres descendre, la résultante pour le trafic pouvant être positive, négative ou neutre.
- et seuls les sites positionnés majoritairement sur les requêtes nouvellement impactées, et ne présentant pas de résultats "frais", ont pu réellement être touchés par cette mise à jour.

Mais pourquoi Google a t'il ressenti le besoin de renforcer la fraîcheur de ses résultats ? Quelles requêtes sont concernées, et pourquoi ce changement n'affecte que certaines requêtes et pas les autres ?

Pour expliquer tout cela, nous allons devoir tout d'abord introduire le concept des requêtes QDF.

Caffeine winners:

	Domain	SEO Visibility 11/06/2011	SEO Visibility 10/30/2011	Percent
Brand	last.fm	1.386.908	1.232.115	12,56
Brand	overstock.com	814.747	774.752	5,16
Brand	lonelyplanet.com	412.446	377.565	9,24
Brand	rockhall.com	62.610	36.387	72,07
Brand	marthastewart.com	170.392	147.258	15,71
Brand	wikiquote.org	315.054	292.471	7,72
Brand	dominos.com	86.572	65.370	32,43
Brand	papajohns.com	101.561	81.411	24,75
Brand	squidoo.com	406.930	386.806	5,20
Brand	soundcloud.com	229.032	211.331	8,38
Brand	playstation.com	176.529	159.381	10,76
Brand	comcast.com	187.415	170.841	9,70
Brand	hotels.com	210.944	195.153	8,09
Brand	vzw.com	54.978	40.428	35,99
Brand	guitarcenter.com	103.490	89.295	15,90

Caffeine losers:

Domain	SEO Visibility 11/06/2011	SEO Visibility 10/30/2011	Percent
comcast.net	283.036	367.672	-23,02
state.ny.us	151.322	204.599	-26,04
univision.com	65.720	117.392	-44,02
usmagazine.com	210.329	254.976	-17,51
newseum.org	58.367	95.132	-38,65
the570.com	23.214	50.775	-54,28
americanexpress.com	109.808	132.096	-16,87
lyricspick.com	29.360	49.287	-40,43
capitalone.com	95.609	112.678	-15,15
peoplestylewatch.com	48.143	63.439	-24,11
realtor.org	30.849	44.230	-30,25
jezebel.com	63.698	76.736	-16,99
plentyoffish.com	30.987	43.898	-29,41
radioshack.com	41.279	54.039	-23,61
gizmag.com	41.630	54.231	-23,24

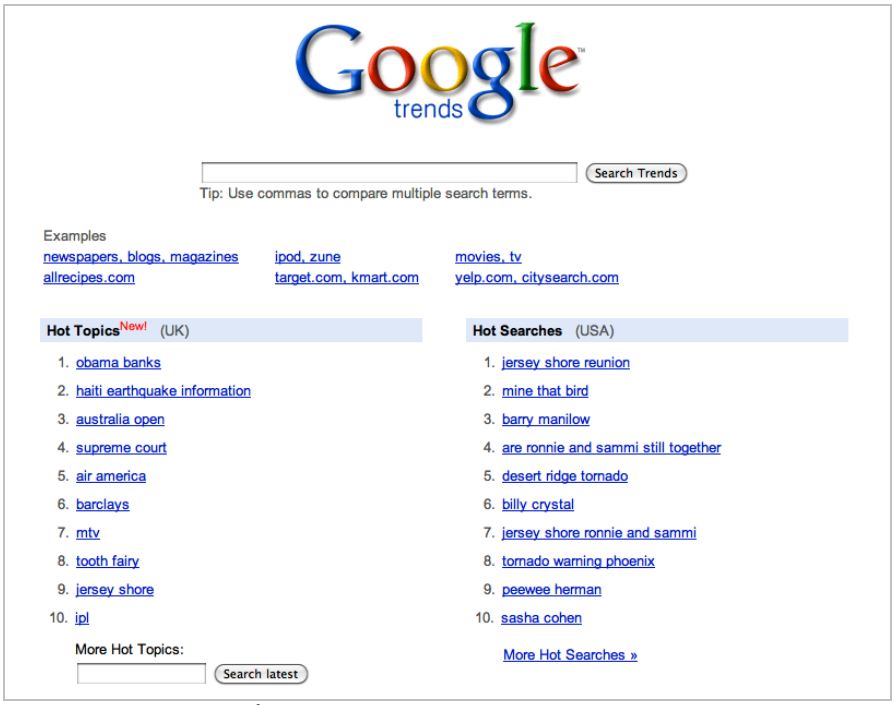
L'impact de la mise à jour « Caffeine 2.0 » sur les sites américains selon Searchmetrics
http://blog.searchmetrics.com/us/2011/11/06/google-freshness-update-many-winners-few-losers/

Les requêtes QDF (Query Deserves Freshness)

En réalité, Google a déjà modifié son algorithme depuis longtemps pour afficher des résultats frais sur certains types de requêtes. Ce comportement "différencié" a été révélé par Amit Singhal, le responsable de l'algorithme de classement chez Google, dans un article du New York Times daté du 3 juin 2007 : "*Google Keeps Tweaking Its Search Engine*" (http://www.nytimes.com/2007/06/03/business/yourmoney/03google.html?pagewanted=all).

Voilà ce qu'Amit Singhal dévoilait dans son article :

> *Durant la seconde moitié de [2006], l'un des [problèmes] récurrents qui était remonté était la "fraîcheur".*
>
> *La "fraîcheur", qui mesure la proportion de pages récemment créées ou modifiées dans une page de résultat, est au centre d'un débat constant chez les spécialistes des moteurs de recherche : est-il préférable de fournir des informations récentes, ou de renvoyer des pages qui ont résisté à l'usure du temps et sont donc probablement de meilleure qualité ? Jusqu'à maintenant, Google a donné la préférence à des pages suffisamment anciennes pour avoir eu le temps d'accumuler des backlinks.*
>
> *Mais l'année dernière, Mr. Singhal à commencé à se demander si la balance ne penchait pas du mauvais côté. Quand la société a introduit son nouveau service de cotation boursière, une recherche sur "Google Finance" ne permettait pas de le trouver ! Après avoir observé des problèmes similaires, il a assemblé une équipe de trois ingénieurs pour trouver une solution.*
>
> *Au début du Printemps 2007, il a présenté les découvertes de son équipe à la réunion hebdomadaire organisée par M. Manber, et qui rassemble les meilleurs ingénieurs en qualité de recherche chargés d'évaluer les projets importants. [...]*
>
> *Mr. Singhal a présenté le problème posé par la "fraîcheur", en expliquant que changer seulement les formules de l'algorithme pour renvoyer plus de pages de résultats récentes finissait par dégrader la qualité des résultats la plupart du temps. Ensuite, il a dévoilé la solution trouvée par son équipe : un modèle mathématique qui cherche à déterminer quand les utilisateurs veulent une information "fraîche", ou non. (et oui, comme tous les projets de Google, cette solution avait un nom : QDF, qui signifie "query deserves freshness", "la requête demande de la fraîcheur").*
>
> *[...]*
>
> *La solution QDF consiste à déterminer si un sujet est "chaud" (hot). Si des sites d'actualités ou des blogueurs traitent activement d'un sujet, le modèle détermine qu'il s'agit d'un sujet sur lequel les utilisateurs sont plus enclins à chercher de l'information actualisée. Le modèle examine également le flot composé par les milliards de requêtes tapées sur Google, et que Mr Singhal considère comme une encore meilleure source d'information pour détecter un enthousiasme mondial à propos d'un sujet déterminé.*
>
> *Il cite comme exemple ce qui se passe quand une ville subit une panne d'électricité : "quand il y a une panne générale à New York, les premiers articles sont rédigés au bout de 15 minutes; nous voyons apparaître des requêtes au bout de deux secondes".*



Hot topics identifiés en novembre 2010 au Royaume Uni et aux USA

> *Mr Singhal indique qu'il a testé QDF pour une application simple : décider s'il faut inclure quelques titres d'actualité parmi les résultats normaux quand les gens tapent des requêtes avec des scores QDF élevés. Bien que Google ait déjà un système différent pour inclure des titres d'actualités sur certaines pages, QDF fournissait des résultats plus sophistiqués, en plaçant les actualités en haut de la page pour certaines requêtes, et en les plaçant au milieu ou en bas de la page pour d'autres."*

Le principe des requêtes QDF s'inscrit donc plus largement dans ce que l'on appelle dans le monde de la recherche d'information, le QIR : "*Query Intent Resolution*" (détermination de l'intention derrière la requête). L'analyse des logs de requêtes et l'analyse statistique des sessions de recherche permet aux moteurs de recherche de découvrir des comportements d'utilisateur récurrents. Cela permet de :
- classer les requêtes par types, en fonction de l'intention de l'utilisateur.
- créer des algorithmes différenciés, capables de fournir de meilleurs résultats en fonction de l'intention supposée de l'utilisateur.

Comment reconnaître une requête QDF ?

Certains indices permettent de reconnaître que le classement de la page de résultats est influencée par le critère de fraîcheur. En particulier, l'omniprésence de l'horodatage dans les snippets.

Voici un exemple de requête QDF :

Une requête sur présidentielle RDC le 11/12/2011 renvoyait une très grande majorité de résultats frais, comme le révèle l'horodatage des résultats.

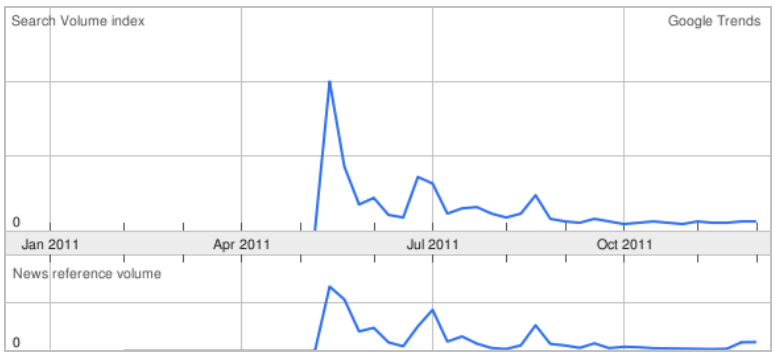
Les différents types de requêtes QDF

En fait, les requêtes QDF ne constituent pas un tout monolithique... Les requêtes "demandent de la fraîcheur" pour différentes raisons, qui conduisent à créer une typologie de requêtes plus fine, et qui appelle une différence de traitement dans l'algorithme de classement.

1. Les requêtes sur des évènements récents ou sur des sujets "brûlants d'actualité"

C'est la première famille de requêtes QDF que Google a voulu détecter et prendre en compte dans son algorithme. Comme expliqué plus haut, la détection se fait en analysant la fréquence de publications dans les sites et les blogs d'actualités et en surveillant les pics de requêtes sur le moteur de recherche.

Il faut noter qu'une telle requête devient temporairement QDF : elle ne remplissait pas les critères avant, et peut ne plus remplir les critères au bout de quelques jours...



Evolution du trafic sur la requête « affaire DSK ».
Avant le 14 mai 2011, ce n'est pas un hot topic …

2. Les évènements récurrents

Les évènements annuels (ou qui ont lieu tous les quatre ans comme les Jeux Olympiques) risquent de devenir des sujets "brûlants" lorsque l'on se rapprochera de la date du prochain évènement. Ils seront détectés comme tels, et en toute probabilité, classés dans la première catégorie ci-dessus.

Mais en attendant : une requête sur les "jeux olympiques" (sans autre précision) en décembre 2011 est probablement tapée par un utilisateur qui cherche des informations sur London 2012 que sur Tokyo 1964. Améliorer les résultats passe donc par faire remonter dans les classements les pages qui parlent de la dernière session ou de la prochaine session de l'évènement. Cela suppose néanmoins que trois choses ont été réglées en amont :
- identifier les évènements récurrents et les requêtes associées.
- identifier la date des évènements, pour savoir quel comportement adopter.
- identifier les pages qui parlent d'une session particulière.

3. Les informations fréquemment mises à jour

Si une requête appelle une information susceptible de changer tous les jours, voire toutes les heures, il convient évidemment de fournir des résultats particulièrement frais.

C'est le cas pour une requête comme *"nfl scores"* (les scores des matchs de football américain) : il est clair que la plupart des utilisateurs s'intéressent aux scores de la veille, pas à ceux de 1973. Même chose pour la requête "*cours du dollar*", ou pour "*Prix Iphone 4S*".

S'agissant des requêtes sur les prix, les bonnes affaires, les réductions etc., il est intéressant de constater que les mêmes requêtes en anglais sur google.com produisent beaucoup plus de résultats horodatés que les requêtes sur google.fr et en français. Mais seule une étude approfondie permettrait de déterminer si l'algorithme a plus ou moins d'impact en fonction des index pays et des langues.

Il est intéressant de remarquer que seule la première catégorie de requêtes était réellement gérée par la solution QDF de 2007. La mise à jour intervenue début novembre gère clairement aussi les deux autres catégories, c'est donc une extension du concept QDF.

Maintenant, on peut se demander si ces trois catégories de requêtes représentent réellement 35% des requêtes tapées par les internautes. Evidemment non, c'est absurde !

Google a changé l'équilibre de son algorithme.

Pour impacter 35% des requêtes, il est quasiment obligatoire d'introduire une prise en compte plus importante du critère de fraîcheur sur la plupart des requêtes.

Après, l'impact réel peut être quasi indécelable sur des requêtes non QDF, d'où le deuxième chiffre communiqué par Google, indiquant que seul 6 à 10% des requêtes ont vu les classements dans les pages de résultats chamboulés "notablement". Peut-être le principe d'un algorithme modifié sur les requêtes QDF est-il préservé.

Mais c'est bien à un rééquilibrage violent entre "importance des pages" (autrement dit : le pagerank) et "fraîcheur des pages" que l'on a assisté, et au profit du critère "fraicheur".

Pourquoi un changement d'algorithme aussi massif ?

Evidemment, on peut se demander ce qui a conduit Google a procéder maintenant à un changement aussi radical.

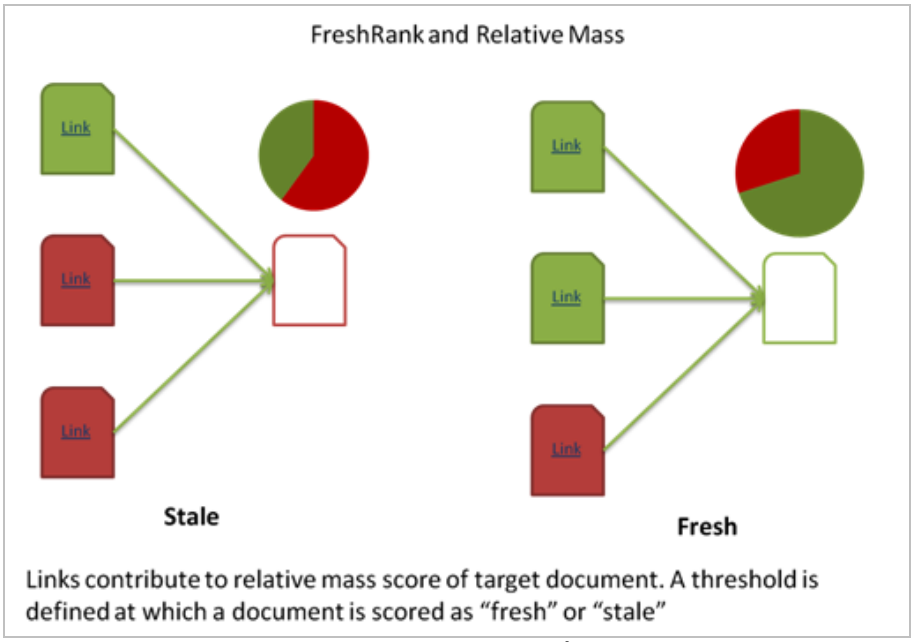
On peut trouver deux explications principales : Caffeine et le changement de comportement des internautes.

Caffeine : le problème et la solution du problème

Le déploiement de Caffeine en 2010 a permis à Google de crawler beaucoup plus de pages, plus vite. Evidemment, cette nouvelle infrastructure était conçue pour améliorer la "fraîcheur" de l'index.

Paradoxalement, Caffeine a eu des effets de bord contraires : d'abord, en crawlant plus de pages, Caffeine a favorisé les "gros" sites, comportant des millions de pages, par rapport aux petits. Il a donc fallu rectifier le tir à l'aide de changements d'algorithmes comme "Mayday". Ensuite, la taille considérablement élargie de l'index de Google n'a fait que renforcer la mauvaise qualité des signaux renvoyés par le "graphe des liens". Déjà, la pertinence du pagerank avait été durement mise à mal par la montée en puissance des réseaux sociaux, et par la disparition progressive de certaines formes de linking. Et les signaux renvoyés par les nouvelles pages ramenées par Google étaient plutôt de mauvaise qualité.

Bref, il est devenu indispensable pour Google de modifier son algorithme pour tenir moins compte de la popularité par les liens, et d'essayer d'introduire une notion de "popularité sociale". Or cette "popularité sociale" se porte avant tout sur des résultats "frais" : le microblogging, ou le graphe social se développe essentiellement autour d'évènements ou d'informations "récents".



Principe du « fresh rank » : pour corriger certains défauts du Pagerank, notamment sa propension à « doper » des pages obsolètes ou délaissées par les internautes d'aujourd'hui, les liens présents sur la toile se voient attribuer un score de fraîcheur transmissible. L'idée est ensuite de combiner les scores de freshrank et de pagerank dans l'algorithme du moteur.

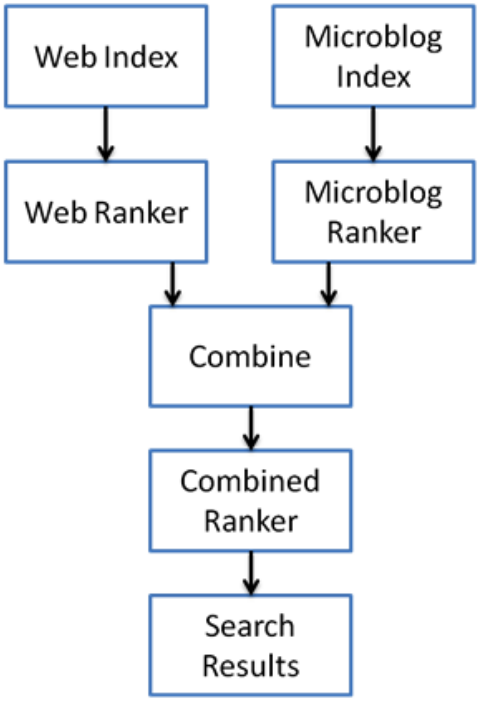
Par ailleurs, Caffeine est en même temps la solution au problème créé : puisque les pages "importantes" sont en général plus "fraiches" qu'autrefois, il faut renvoyer plus de pages "récentes", ce qui n'est plus un problème depuis Caffeine, puisque Googlebot crawle et recrawle plus, plus vite et plus intelligemment !

Le comportement des utilisateurs a changé : ils cherchent plus volontiers des résultats frais qu'auparavant

S'il existe a un domaine dans lequel l'offre crée effectivement la demande, c'est bien dans le domaine de la recherche d'information. Aujourd'hui, il est plus aisé de savoir si sa ligne de RER préférée est en panne (et pourquoi) en suivant les messages sur Twitter que sur le site de la

RATP. L'afflux d'informations d'actualité produites dans le graphe social conduit de plus en plus d'utilisateurs à chercher des informations sur les moteurs de recherche qu'ils ne cherchaient pas auparavant, parce qu'ils pensaient que cette information n'était pas (encore) disponible.

Nous n'avons pas vu d'études précises encore sur le sujet, mais il semble bien que la proportion des requêtes d'actualité dans les requêtes tapées ait bel et bien augmenté, notamment si l'on croit les remarques relevées dans plusieurs articles publiés par les équipes de recherche de Yahoo!.



Principe d'un moteur intégrant un score de microblogging dans son algorithme, décrit dans un brevet de Yahoo !

Quelles conséquences pour le référencement des sites ?

Ce "rééquilibrage" de l'algorithme a eu un impact plus ou moins fort sur les sites selon la typologie des requêtes sur lesquelles ils étaient positionnés. Cela ne change donc pas grand chose si vos pages constituent des réponses à des réponses non QDF.

A l'inverse, si les requêtes qui vous rapportent du trafic sont des requêtes de type QDF (ou apparentées), il convient de vérifier :

- qu'aucun obstacle technique ne vient à perturber le crawl et l'indexation RAPIDE par google de vos pages et une bonne compréhension de la date de création de vos documents, et du rythme de leur changements :
    * renvoyer la bonne date de création.
    * avoir des dates de publication repérables et lisibles.
    * apporter des changements dans les pages en suivant toujours la même logique cohérentes.
    * renvoyer la bonne date de mise à jour en cas de changement.

- que vous renvoyez bien des signaux montrant que votre contenu est :
    * à jour.
    * pas en voie d'obsolescence.
    * susceptible de continuer à intéresser les internautes.

Il est important ici, en particulier, de s'assurer que les contenus reçoivent régulièrement des liens nouveaux.

Ensuite, d'une façon générale, il est important de développer une activité sociale autour du site, afin de développer les signaux montrant que de vrais humains s'intéressent en ce moment à vos contenus.

Un changement d'échelle, mais pas une nouveauté !

Le principe à l'œuvre dans la mise à jour que les SEO ont surnommée Caffeine 2.0 est loin d'être nouveau. Le principe des requêtes QDF date de 2006, sa mise en oeuvre de 2007.

Ce qui frappe dans cette mise à jour, c'est le changement d'échelle. En 2011, Google a osé procéder à des changements majeurs, qui affectent une proportion énorme des pages de résultats (on pense aussi à Panda).

Il semble que l'on rentre dans une période où des idées, énoncées dans des articles scientifiques ou des brevets il y a des années, sont enfin concrètement mises en oeuvre. Tout se passe comme si Google avait réussi à éliminer les obstacles l'empêchant de traiter les volumes de données gigantesques nécessaires pour déployer ces idées de façon efficace. C'est vrai avec l'infrastructure Caffeine, c'est vrai aussi avec Panda et les algorithmes d'apprentissage automatique. On peut aussi noter que ce type d'algorithmes est souvent décrit comme une solution élégante pour parvenir à "classer" correctement les requêtes QDF.

Tout laisse penser que d'autres changements majeurs restent encore à introduire dans l'algorithme, en particulier une meilleure prise en compte des signaux "sociaux". D'autres grands changements sont encore possibles... En attendant, il va falloir veiller à la "fraîcheur" de vos contenus.

BIBLIOGRAPHIE ET REFERENCES

L'annonce de la mise à jour de l'algorithme :
http://insidesearch.blogspot.com/2011/11/giving-you-fresher-more-recent-search.html

Pour approfondir le problème de la fraîcheur, un article de Justin Briggs :
http://justinbriggs.org/methods-for-evaluating-freshness

Articles

Confucius and Its Intelligent Disciples: Integrating Social with Search
http://infolab.stanford.edu/~echang/Confucius-VLDB10.pdf
Xiance Si, Edward Y. Chang, Zoltan Gyongyi, Google, Maosong Sun, Tsinghua University Beijing - China

Towards Recency Ranking in Web Search
http://www.wsdm-conference.org/2010/proceedings/docs/p11.pdf
Anlei Dong Yi Chang Zhaohui Zheng Gilad Mishne
Jing Bai Ruiqiang Zhang Karolina Buchner Ciya Liao Fernando Diaz
Yahoo! Inc.

Brevets

Ranking User Generated Web Content
Invented by: Xiance Si, Jian Gong Deng, Huacheng Ke, Dong Zhang, Zoltan I. Gyongyi, and Edward Y. Chang
Applicant: Google
Publication Number WO/2011/050495
Publication Date: May 5, 2011
International Filing Date: October 29, 2009

Information retrieval based on historical data

http://patft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=7,346,839.PN.&OS=PN/7,346,839&RS=PN/7,346,839
Inventors: Acharya; Anurag (Campbell, CA), Cutts; Matt (Mountain View, CA), Dean; Jeffrey (Palo Alto, CA), Haahr; Paul (San Francisco, CA), Henzinger; Monika (Lausanne, CH), Hoelzle; Urs (Palo Alto, CA), Lawrence; Steve (Mountain View, CA), Pfleger; Karl (Mountain View, CA), Sercinoglu; Olcan (Mountain View, CA), Tong; Simon (Mountain View, CA)
Assignee: Google Inc. (Mountain View, CA)
Appl. No.: 10/748,664
Filed: December 31, 2003

DOCUMENT SCORING BASED ON DOCUMENT CONTENT UPDATE
http://appft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.html&r=1&p=1&f=G&l=50&d=PG01&S1=20110258185.PGNR.&OS=dn/20110258185&RS=DN/20110258185
Inventors: ACHARYA; Anurag; (Campbell, CA) ; DEAN; Jeffrey; (Palo Alto, CA) ; HAAHR; Paul; (San Francisco, CA) ; HENZINGER; Monika; (Corseaux, CH) ; LAWRENCE; Steve; (Mountain View, CA) ; PFLEGER; Karl; (Mountain View, CA) ; TONG; Simon; (Mountain View, CA)
Assignee: GOOGLE INC. Mountain View CA
Serial No.: 174243
Series Code: 13
Filed: June 30, 2011

Incorporating Recency in Network Search Using Machine Learning - Anlei Dong et al
http://www.google.com/patents/about/12_579_855_Incorporating_Recency_in_Netw.html?id=nrXLAQAAEBAJ
Application number: 12/579,855
Publication number: US 2011/0093459 A1
Filing date: Oct 15, 2009

***Philippe YONNET** , Directeur SEO international, Twenga.*

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2011/12/les-requetes-qdf-et-le-nouvel.html