Etude sur les changements apportés par le projet "Caffeine" de Google

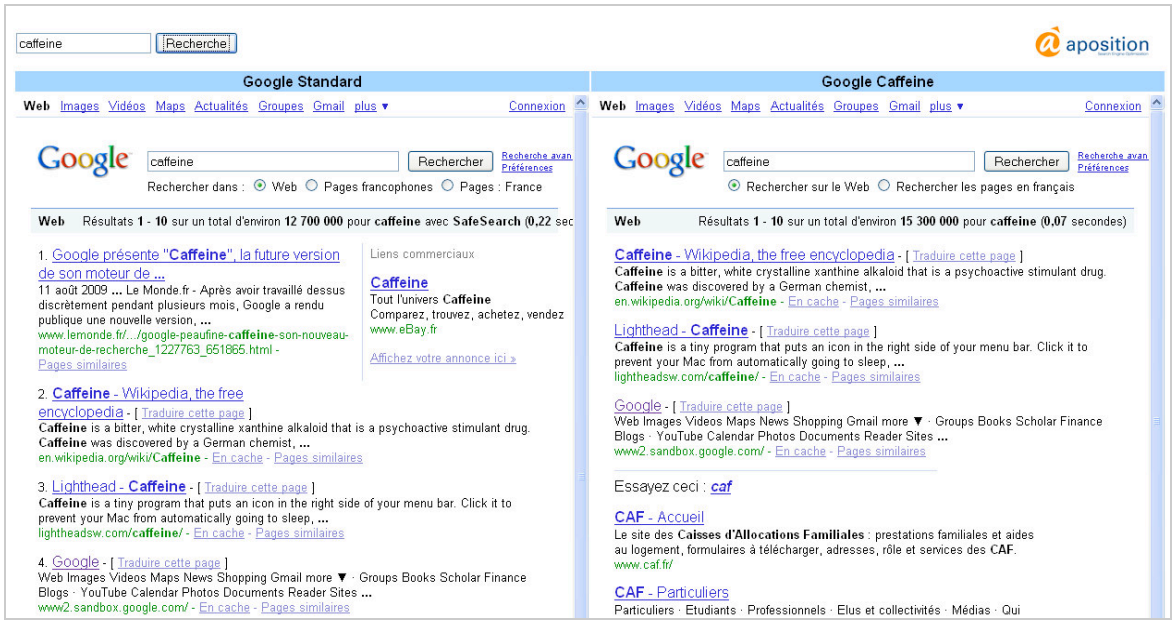
Domaine :	Recherche	**Référencement**
Niveau :	Pour tous	**Avancé**

Au mois d'août 2009, Google a communiqué sur une nouvelle infrastructure mise en place pour son moteur de recherche, recevant le nom de code "Caffeine". Une adresse était même fournie pour tester cette nouvelle et prochaine mouture du moteur qui devrait être mise en œuvre dans les semaines qui viennent. Que va-t-elle apporter ? Les résultats de recherche seront-ils réellement différents avec cette nouvelle version ? Nos sites risquent-ils de perdre ou de gagner du trafic ? Le nouvel index est-il plus important ? Pour répondre à toutes ces questions, de nombreux tests ont été effectués et les résultats peuvent parfois être assez surprenants...

Le 10 août 2009, Google a dévoilé publiquement qu'une équipe de ses ingénieurs travaillait en secret depuis plusieurs mois sur une nouvelle architecture pour son moteur de recherche. Cette « nouvelle génération » de l'architecture de Google a été affublée du nom de code : ***caffeine***. L'objectif déclaré de cette mise à jour est de gagner « en volume de données, en vitesse d'indexation, en précision, en exhaustivité et d'autres dimensions ». Voir le billet du blog Google webmaster central (http://googlewebmastercentral.blogspot.com/2009/08/help-test-some-next-generation.html) : *It's the first step in a process that will let us push the envelope on size, indexing speed, accuracy, comprehensiveness and other dimensions*. Google précise toutefois que « caffeine » est la première étape du processus qui permettra d'obtenir les gains attendus.

De manière un peu inhabituelle, Google a proposé assez rapidement une URL de test pour permettre aux webmasters de faire remonter leurs remarques sur Caffeine. Nous avons pu étudier ce « bac à sable » avec l'aide de notre équipe de R&D, et nous sommes en mesure de vous délivrer en exclusivité les premiers résultats de nos tests et de nos investigations. Nous verrons que globalement, les pages de résultats sont peu chamboulées par cette mise à jour, mais que les différences sont suffisantes pour modifier la hiérarchie entre certains sites.

Caffeine est donc tout sauf une mise à jour mineure, et même si nous n'en voyons pas encore la version définitive, il est clair que les webmasters ont intérêt à se préparer à de réels changements.



GFS2 MapReduce et BigTable : les secrets de l'infrastructure de Google

Le dernier changement majeur d'architecture du moteur Google remonte probablement à la mise en place de « Teragoogle », alias « Big Daddy », qui s'est étalée de fin 2005 à début 2006. Depuis, peu d'indices laissaient penser que l'infrastructure avait pu évoluer.

Un certain nombre d'informations ont filtré depuis l'annonce de la mise à jour « Caffeine » sur les changements sous-jacents. Tout d'abord, Google a clairement annoncé qu'une nouvelle version de son système d'exploitation propriétaire avait été développé (ou plus exactement de son système de fichiers). Sean Quinlan (« ingénieur principal » chez Google, en charge du projet GFS après avoir fait ses premières armes chez Bell Labs ou Sawmill), a décrit dans une interview donnée au magazine de l'ACM (l'*Association for Computing Machinery* : http://queue.acm.org/detail.cfm?id=1594206) les différentes améliorations apportées par Google à GFS. Et Matt Cutts a confirmé que Caffeine serait effectivement « propulsé » par la version 2 de GFS (*Google File System*) (http://www.theregister.co.uk/2009/08/14/google_caffeine_truth/) :
« Il y a beaucoup de technologies sous le capot derrière Caffeine, et l'une des choses sur lesquelles Cafeine est construit est un système de stockage de nouvelle génération. Caffeine utilise bien ce que vous désignez sous le terme "GFS2" »

GFS2, pour résumer, est conçu pour faciliter le « cloud computing », c'est-à-dire la capacité de pouvoir faire tourner des programmes et de travailler sur des données sans avoir à se soucier de l'emplacement physique. « L'informatique nuageuse » est une évolution incontournable aujourd'hui pour tous les secteurs de l'informatique, mais qui se double pour Google de problèmes plus difficile à résoudre : les calculs à effectuer pour un moteur de recherche sont lourds, les volumétries de données à traiter sont impressionnantes et le rythme de mise à jour très élevé. GFS2 permet de travailler sur des éléments dont la granularité a beaucoup baissé (un « chunk » descend de 64 MB à 1 MB), ce qui permettra à Google de gérer des applications qui s'accommodaient mal des limitations de GFS (comme Gmail par exemple). Autre évolution importante : jusqu'ici, GFS distribuait les calculs entre des ordinateurs esclaves, mais les calculs effectués par les "maîtres" étaient traités normalement. Dorénavant, GFS2 permet aussi de gérer les logiciels tournant sur les "maîtres" en les distribuant entre futures machines.

Mais un système de fichiers n'est que la brique la plus basique de l'infrastructure d'un moteur de recherche. L'efficacité d'un moteur dans ses tâches de crawl, d'indexation, et dans sa capacité à répondre aux requêtes dépend d'autres composants logiciels. S'agissant de Google, on sait que leur système de calcul s'appuie sur deux couches logicielles clé : MapReduce et BigTable. MapReduce est une plateforme de calcul distribuée, et BigTable un système de gestion de données distribué et capable de répondre en temps réel.

Sean Quinlan a révélé que l'on était passé progressivement dans l'histoire de Google d'une sollicitation intensive de MapReduce, à une situation dans laquelle une partie des tâches et de la charge était transférée à BigTable.

Clairement, BigTable est l'une des applications clé aujourd'hui dans l'infrastructure de Google. Matt Cutts n'a pas voulu répondre à une question sur l'existence d'une version 2 de BigTable. Mais on peut supposer qu'une version améliorée de l'infrastructure passe aussi par une réécriture de certains composants clés comme MapReduce et BigTable.

Quelques tests sur la plateforme Caffeine

La plateforme Caffeine est-elle plus rapide ?

Si on se fie aux temps de génération de pages donnés en haut et à droite des pages de résultat, ils semblent inférieurs sur Caffeine. Voici ci-dessous une comparaison entre quelques temps de réponse observés pour trois exemples de requêtes (la différence est nette et facile à reproduire).

Mots clés	Temps de réponse sur Caffeine	Temps de réponse actuel sur Google
SEO	0.10 s	0.34 s
Caffeine	0.16 s	0.32 s
Hotel Paris	0.18 s	0.80 s

Même chose si on teste les temps de réponse avec un autre outil : statistiquement, les temps de réponse de la plateforme Caffeine sont nettement inférieurs. Maintenant qu'est-ce que cela prouve ? Pas grand chose, car il est peut-être tout à fait normal que le datacenter qui répond sur les requêtes adressées à ww2.sandbox est peut-être très peu sollicité par rapport à un *datacenter* normal. Par ailleurs Caffeine n'affiche pas toujours tous les résultats de recherche universelle affichés sur www.google.fr, et on observe par moment des temps supérieurs.

Il faut donc probablement reporter notre jugement et dire que pour l'instant il est difficile de confirmer ce point.

La plateforme Caffeine présente un index plus exhaustif

On sait à quel point il faut se méfier du discours des moteurs sur la taille des index. Dans le passé, leur communication sur ce point n'a toujours été d'une honnêteté exemplaire (voir en particulier le travail de Jean Veronis sur ce point précis sur son blog Aixtal au cours de l'été 2005).

En fait, dans la mesure où il est impossible de vérifier l'intégralité des pages de résultats sur les requêtes qui retournent des millions ou des milliards de liens, il faut forcément croire Google (et ses concurrents) sur parole lorsqu'il communique un nombre de réponses à ces requêtes.

Bill Slawski (l'éditeur du site Seobythesea) a fait une étude comparative sur les mots les plus communs en anglais. Voila un extrait des résultats de cette étude, publiée le 15 août 2009 :

Comparaison du nombre de résultats entre différents moteurs *(les chiffres sont en milliards de pages)*

Query	Google Caffeine	Google	Yahoo	Bing	Ask
a	19,32	17,57	31,2	7,8	1,28
in	15,85	13,98	30,2	7,85	900
to	15,22	13,5	27,5	8,92	1,74
the	14,85	13,9	28,8	8,17	747
of	14,76	12,99	28	7,31	794
and	13,98	12,95	28	7,49	789
for	12,11	10,72	26,8	7,74	769
by	12,08	10,42	27	6,12	956
on	11,26	9,94	25,1	5,61	598
is	9,58	8,87	22,6	4,25	699
I	9,22	8,25	18,6	3,86	686
all	9,11	7,58	27,2	6,99	1,02
this	8,89	7,87	21,5	5,79	585

with	8,49	6,3	20,9	2,44	636
it	7,7	6,86	19,3	4,19	542
at	7,41	6,6	20,8	3,93	552
from	7,34	6,92	18,4	4,16	521
or	7,03	6,21	19,5	3,94	567
you	6,76	5,93	19,9	5,08	543
as	6,46	5,75	15,4	3,55	884
your	6,36	5,47	19,5	3,79	495

Pour lire le reste de l'étude : http://www.seobythesea.com/?p=2795

Evidemment, si on en croit ces chiffres, Caffeine semble donner un peu plus de réponses qu'avant. Mais on sait que ces chiffres n'ont rien à voir avec la « qualité » du moteur (qui nécessiterait que l'on se penche sur une analyse de la « précision » ou du « rappel » pour reprendre les termes consacrés).

Il est beaucoup plus judicieux de regarder ce qui se passe sur des mots rares de la langue française : là il sera possible de comparer les moteurs et de voir qui ramène le plus de pages sur un sujet (avec la possibilité de compter le « vrai » nombre de résultats et même de jauger la pertinence des réponses.

Voici quelques exemples de résultat obtenus sur des mots rares :

Requêtes testés	Nombre de résultats dans Google.fr "normal"	Nombre de résultats dans Caffeine
Manichordion	644	3 240
Quilboquet	1 610	4 080
Paréchème	223	1 950
Polytypon	56	65
Homorime	509	1 300
Homéotéleute	1 340	10 600

Ce type de requête permet de tester l'exhaustivité d'un moteur. Il s'avère que dans le cas de Caffeine, l'index semble effectivement plus grand.

Mais il n'est pas certain que cela améliore de manière spectaculaire la qualité du moteur. L'étude de la requête sur « Homéotéleute » en donne un exemple flagrant : on passe de 1 340 à .... 10 600 résultats, mais si on s'intéresse à la nature des 9260 résultats nouveaux, la déception est grande. D'abord on s'aperçoit (et c'est toujours le cas) que le chiffre donné par Google sur les pages de résultat comprend les « duplicates ». En réalité Google ne présente que... 294 résultats dans sa version actuelle, et 434 dans la version Caffeine. Les autres pages sont « ignorées » parce qu'elles présentent « un contenu similaire ». Même si cela fait quand même presque 50% de pages en plus, c'est déjà moins spectaculaire, l'essentiel du volume supplémentaire de pages étant avant tout constitué de « duplicates » ou de « near duplicates ».

Une étude sur un échantillon de 50 000 mots clés

Notre service de R&D a effectué une étude comparative entre Google.fr et la version Caffeine (avec le paramètre géolocalisée en france et des résultats en Français) au cours de la première semaine de septembre 2009.

L'exercice a ses limites, car il s'agit d'une version « internationale » et de test, qui ne reflète pas forcément la version finale qui sera visible sur Google.fr. Mais faute de mieux, cela permet quand même de détecter d'éventuels changements et de s'y préparer. Les conclusions de cette étude permettent surtout de tordre le cou à certaines rumeurs apparues sur le net concernant les changements intervenus sur Caffeine.

Comparaison des "top visibilités" en première page de Google.

Le tableau ci-dessous présente un classement effectué en fonction du **nombre de requêtes** (sur l'échantillon de 50 000) **sur lesquels les domaines apparaissent en première page de résultats**. L'échantillon a été extrait aléatoirement d'une base de 250000 requêtes tapées par les internautes sur Google.fr.

Classement	DOMAINES	Caffeine	Normal	delta	Classement actuel
1	fr.wikipedia.org	18353	18919	-566	fr.wikipedia.org
2	images.google.fr	13410	6	13404	video.google.com
3	news.google.fr	6301	10	6291	maps.google.com
4	maps.google.fr	5354	44	5310	news.google.com
5	www.commentcamarche.net	5004	4625	379	www.commentcamarche.net
6	video.google.fr	4951	38	4913	www.dailymotion.com
7	www.dailymotion.com	2481	2390	91	www.youtube.com
8	www.youtube.com	2383	2335	48	en.wikipedia.org
9	www.linternaute.com	2238	2177	61	www.allocine.fr
10	www.allocine.fr	2061	2192	-131	www.linternaute.com
11	www.amazon.fr	1668	1888	-220	www.amazon.fr
12	www.infos-du-net.com	1595	1626	-31	www.evene.fr
13	en.wikipedia.org	1589	2200	-611	www.infos-du-net.com
14	www.ciao.fr	1557	1509	48	www.clubic.com
15	www.evene.fr	1536	1721	-185	www.ciao.fr
16	www.clubic.com	1499	1573	-74	www.wikio.fr
17	www.wikio.fr	1119	1374	-255	www.doctissimo.fr
18	www.doctissimo.fr	1102	1094	8	www.01net.com
19	www.01net.com	1074	1052	22	www.priceminister.com
20	www.leguide.com	1027	1000	27	www.leguide.com

Les principaux chamboulements concernent les domaines de Google (images, news, maps, video, youtube). Ces variations sont clairement des artefacts provenant des différences de géolocalisation entre les deux versions testées. Un seul domaine disparait du top 20, mais il ne perd que 4 places. En fait ce classement montre que globalement le classement du "top visibilité" est inchangé avec Caffeine.

Identification des domaines qui progressent en visibilité : une forte poussée des sites institutionnels

Voici le classement des sites qui gagnent le plus de visibilité en première page :

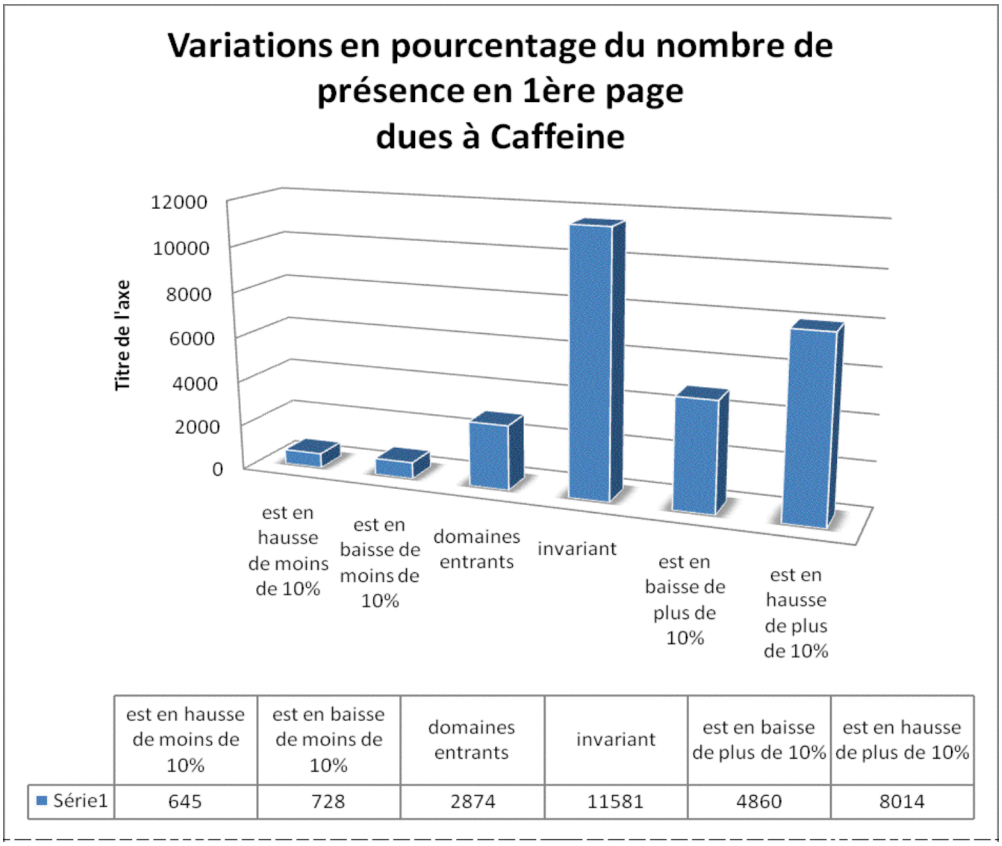
	domaine
1	books.google.com
2	profile.myspace.com
3	lachainemeteo.com

4	fr.youtube.com
5	meilleurtaux.com
6	www.fiscal.gouv.fr
7	front.webedu.men.aw.atosorigin.com
8	www.mister-wong.fr
9	m.youtube.com
10	premier-ministre.gouv.fr

En observant de près les 100 premiers du classement, on détecte une forte présence de sites gouvernementaux en .gouv.fr (12%) ou de sites faisant autorité (15%) comme des sites d'université, d'institutions ou du monde de l'éducation.

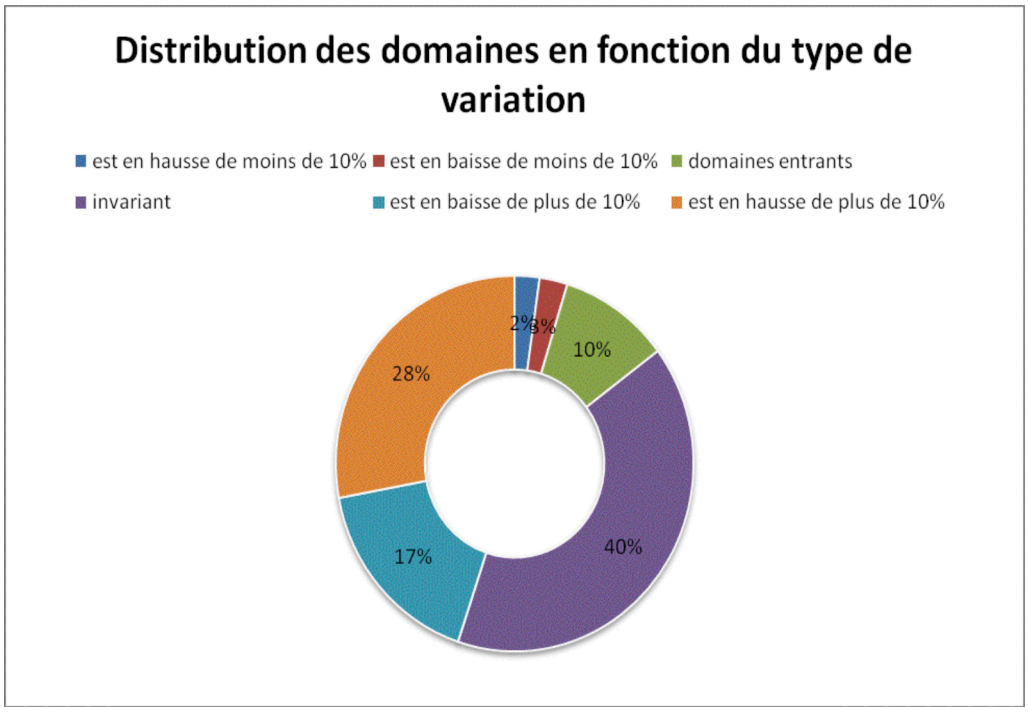
Il ne faut pas néanmoins en déduire que les sites institutionnels grimpent systématiquement dans les classements : d'autres sites du même type disparaissent des premières pages. Il semble donc que certains des critères concernant cette catégorie de sites varient de manière significative dans l'index Caffeine.

Evaluation de l'impact global



	est en hausse de moins de 10%	est en baisse de moins de 10%	domaines entrants	invariant	est en baisse de plus de 10%	est en hausse de plus de 10%
Série1	645	728	2874	11581	4860	8014

Globalement, la mise à jour Caffeine semble bien se faire à algorithme constant : peu de premières pages de résultats sont chamboulées. 40% ne connaissent pas de changements réels de visibilité, et 45% un changement nul ou faible. On reste dans le « bruit » normal dû à des changements dans les données indexées ou l'évolution normale de l'internet.



A côté de cela, un petit nombre de sites subissent des variations violentes : doublement, triplement voire plus des positions en première page, ou perte de plus de 90%.
De telles variations signifient probablement un impact sérieux sur l'audience de ces sites, en positif comme en négatif, car l'échantillon des requêtes étudiées a été prélevé sur un « top 250 000 » des requêtes les plus tapées sur le net.

Faut-il craindre cette nouvelle infrastructure de Google ?

On concluera cette étude sur le fait que le passage à l'infrastructure Caffeine peut entrainer des variations de trafic importantes pour certains sites. Mais pour presque la moitié des domaines, l'impact sera sans doute nul et négligeable.

Le changement de la taille de l'index semble produire quelques effets de bord : par exemple les sites comportant un volume de pages important semblent être favorisés dans la version étudiée.

Il est probable que les changements les plus radicaux n'arriveront pas avec la mise en place de Caffeine, mais plutôt avec la mise en ligne de nouvelles fonctionnalités que les équipes de Google ont évoqué sans en préciser clairement la nature.

En tout cas nous ne devrions pas tarder à être fixés : même si Google n'a communiqué aucune date pour la mise en service de Caffeine, beaucoup parlent d'une échéance courant septembre 2009 !

Philippe Yonnet, *Directeur Technique @Position (http://www.aposition.com) et président de l'association SEO Camp (http://www.seo-camp.org/)*

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://abonnes.abondance.com/blogpro/2009/09/etude-sur-les-changements-apportes-par.html