L'impact des Mayday, Jazz et Caffeine sur Google

Retour au sommaire de la lettre

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

Le moteur de recherche Google a connu, ces dernières semaines, trois événements majeurs : la mise en place d'une nouvelle interface, baptisée Jazz, un nouveau système de crawl et d'indexation, ayant pour nom Caffeine ainsi qu'une mise à jour de l'algorithme de pertinence, que les webmasters ont nommé Mayday. Résultat : de nombreux responsables de sites ont remarqué de profonds changements dans leur trafic renvoyé par Google. Mais quels ont été les réels impacts de ces changements sur les résultats renvoyés par le moteur de recherche majeur ? Cet article, très fouillé, argumenté et basé sur des études statistiques, tente d'apporter quelques éléments de réflexion...

Depuis quelques semaines, de profonds changements ont été observés par les webmasters sur Google. Une nouvelle interface utilisateur est apparue le 5 mai 2010. Google a annoncé officiellement le 9 juin 2010 le basculement sur leur nouvelle infrasture : Caffeine (http://actu.abondance.com/2010/06/google-annonce-officiellement-caffeine.html). Et Matt Cutts, ainsi que plusieurs autres porte-parole officiels de Google ont confirmé une mise à jour importante de l'algorithme de classement du moteur, baptisée MayDay par une communauté de webmasters US (les membres du forum Webmasterworld) en raison de la date choisie pour sa mise en place : le 1er mai 2010.

Ces trois changements ont déclenché de nombreuses réactions de la part des webmasters, en raison des changements importants qu'ils ont apportés, soit dans le trafic généré par les moteurs, dans les positions, ou dans le comportement d'exploration (crawl du moteur). Certains webmasters de sites importants ont annoncé des chutes de 15 à 20% du trafic organique renvoyé par Google.

Avec un recul de quelques semaines, il est donc intéressant d'analyser les effets réels de ces changements, et d'essayer de comprendre ce qui les a produits ainsi que les objectifs poursuivis par Google. Nous analyserons dans cet article les résultats d'une étude de l'agence Aposition qui montre les étapes du changement, et révèle à quel point les modifications dans le comportement de Google ont redistribué les cartes, en suivant un calendrier parfois éloigné de celui des annonces officielles et qui ouvre de nombreuses nouvelles questions intéressantes sur la réalité de ce qui se trame sur les serveurs du moteur de recherche.

1er MAI 2010 : La mise à jour MayDay

Dès les premières heures du mois de mai, certains webmasters ont constaté des changements brutaux dans le trafic émanant des pages de résultats du moteur Google, et dans les positions observées sur certains mots clés. Les *admins* du forum Webmasterworld ont décidé que l'on pouvait bel et bien parler d'une mise à jour très probable de l'algorithme, et comme il est de tradition sur ce forum, ils ont choisi de baptiser cette mise à jour. Le nom qu'ils ont choisi est "MayDay" en raison de la date à laquelle elle était intervenue (le 1er mai). Mais ce nom qui sonne aussi comme un appel au secours a semblé aussi par la suite faire écho aux nombreux messages de détresse envoyés par des webmasters constatant des chutes de 10%, 15% et même 20% du trafic issu des moteurs de recherche sur les mots clés de la longue traine.

Mayday : un changement d'algorithme "officiel"

Des porte-parole officiels de Google ont confirmé l'existence de cette mise à jour, et ont donné quelques détails complémentaires. Matt Cutts en particulier a évoqué cet "update" à plusieurs reprises au cours du mois de Mai : d'abord à la conférence Google I/O, ensuite lors de sa visite à Paris et également dans une video publiée le 30 mai 2010.

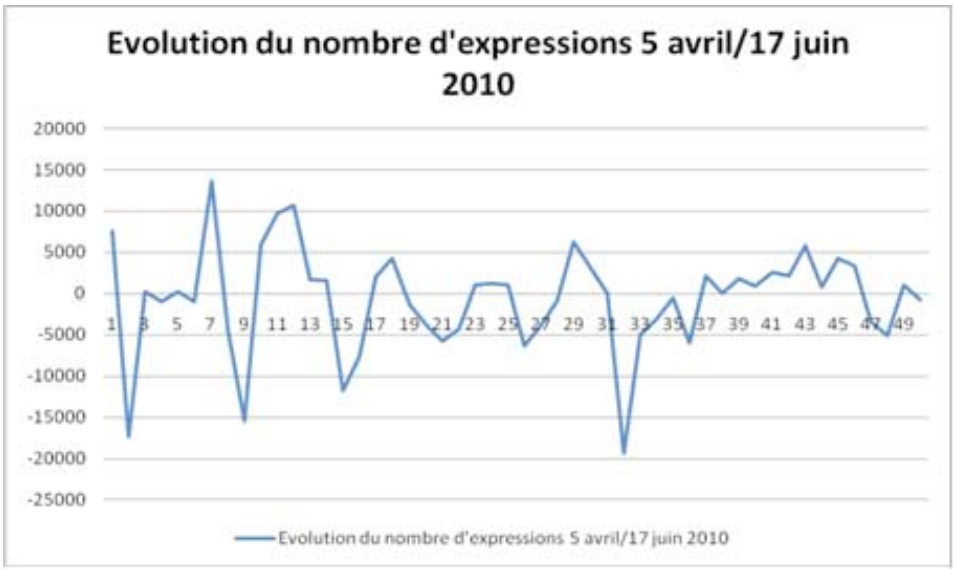
Les informations communiquées par Google sont loin d'être complètes et détaillées, mais pour une fois, on en sait un peu plus que d'habitude. Voyons ce qu'ils ont bien voulu communiquer aux webmasters :
1. Il y a bien eu une mise à jour de l'algorithme ;
2. Cette mise à jour s'est en fait étalée sur plusieurs jours, du 28 avril au 3 mai ;
3. Elle affecte les requêtes "longue traine" plus que les requêtes de "début de traine" ;
4. Ce changement est totalement indépendant du déploiement de Caffeine (voi plus loin) ;
5. L'objectif de ce changement est d'augmenter la pertinence des résultats du moteur sur les requêtes longue traine.

Quel est l'impact réel de ce changement d'algorithme ?

Il faut relativiser l'impact de ce changement d'agorithme. Il n'affecte que la longue traine, et la plupart des acteurs ne remarquent qu'une chute de 15 à 20% maximum sur le trafic issu de ce type de requêtes. Cela signifie que l'impact sur les sites peut être très faible si la part de la longue traine est faible dans le trafic mesuré. Et que si cette part est importante, la baisse est "diluée" dans le reste du trafic issu des moteurs. N'oublions pas par ailleurs qu'à chaque fois que des sites perdent des positions, d'autres en gagnent...

Néanmoins, de telles baisses, même limitées en volume, peuvent représenter des pertes substantielles dans le chiffre d'affaires de sites marchands !

Il semble par ailleurs que ce changement marque un tournant conceptuel pour l'équipe en charge de la qualité de l'algorithme (l'équipe Qualité de recherche d'Amit Singhal semble être derrière ce changement, si on en croit les confidences de Matt Cutts faites lors de son passage à Paris fin mai).



Cette courbe est issue d'une etude d'Aposition.
En ordonnée, le nombre de positions gagnées ou perdues dans les 100 premiers résultats de Google pour les 50 sites français les plus visibles (classement du 17 juin 2010), entre le 5 avril et le 17 juin. Tous les sites ne sont pas affectés de la même façon, mais certains connaissent des chutes ou des gains spectaculaires. Mais un tel changement a déjà été observé en janvier 2010 : est-ce MayDay, ou Caffeine, ou autre chose ?

Qu'y a t-il vraiment derrière ce changement ?

Plusieurs hypothèses ont été émises par les observateurs pour expliquer les changements dans les classements :

- Une pondération différente dans les liens : en effet, les pages des sites affectés semblent pour la plupart n'avoir pour backlinks que des liens internes. Donner moins de poids à des pages qui n'ont pas de liens externes dans les backlinks pourrait produire évidemment une redistribution des cartes sur les requêtes longue traine.

- La prise en compte des syntagmes dans l'algorithme : c'est l'hypothèse de Tedster, admin de Webmasterworld. Une préférence pour les pages qui ont l'expression exacte dans leur contenu serait à l'origine des changements. On trouvera plus de détails sur l'indexation des syntagmes dans notre article du mois dernier de la lettre pro d'Abondance. Mais cette hypothèse présente deux défauts : tout d'abord elle part du principe que la gestion des syntagmes est quelque chose de neuf dans l'algorithme, ce qui est probablement faux. Ensuite, de tels changements ne déclasseraient pas toutes les pages d'un site sur la longue traine, mais auraient tendance à faire monter certaines pages sur des expressions et à les faire descendre sur d'autres, ce que les webmasters n'ont pas observé.

- Une meilleure prise en compte des pages de sites faisant autorité sur un sujet.

Mais un commentaire de Maile Ohye, senior developper programs engineer, lors de la conférence SES de Toronto (9-11 juin 2010) lève peut-être un coin du voile : "*Ce que [MayDay] produit pour les requêtes longue traîne, c'est que nous les considérons maintenant juste comme des requêtes comme les autres. Nous allons donner autant d'importance à ces résultats de recherche que toutes les autres pages de résultat*".

Au cours de cette intervention, Maile Ohye a précisé que l'objectif était notamment d'éviter que des sites puissent apparaître en tête des résultats sur les requêtes avec des méthodes automatisées. En effet, de nombreux sites importants (importants à la fois par leur popularité sur le web et par leur volume de pages) ont pris l'habitude de générer des pages ou des liens pour se positionner sur des requêtes longue traine, en utilisant des scripts pour "couvrir" tout l'univers des expressions clés pouvant rapporter du trafic. Cette stratégie s'avérait souvent payante, mais pouvait pour certains sites aboutir à positionner des pages dont le contenu avait une utilité douteuse en tant que réponse à une requête de l'internaute.

La piste des signaux manquants

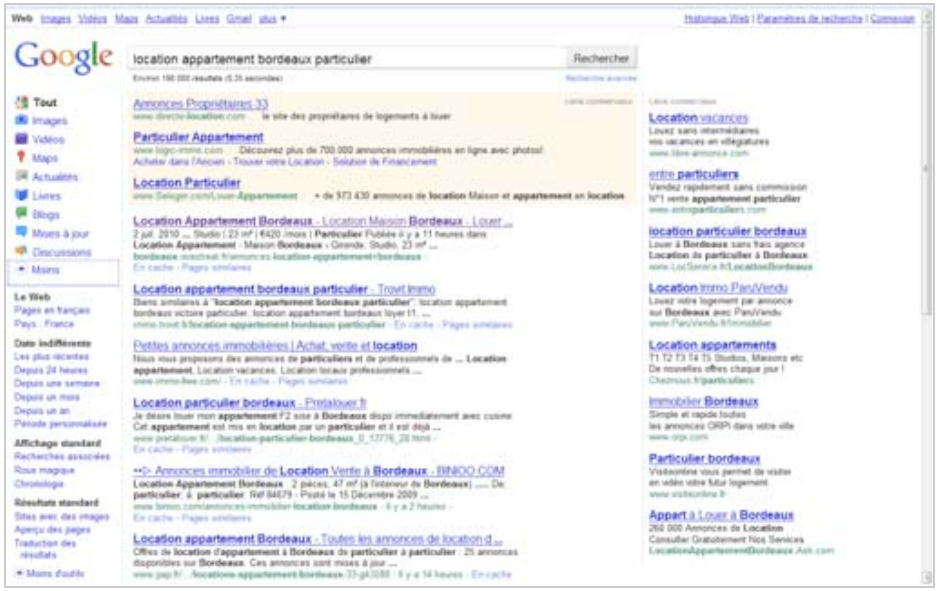
Ce commentaire ouvre une piste nouvelle. La base de données d'un moteur de recherche comme Google est complexe, et contient un nombre faramineux de données. Bien sûr cette base de données est "indexée" et de nombreux signaux sont précalculés, et les données gérées par des systèmes complexes de cache et de "*prefetching*" (le prefetching consiste à pré-extraire les données souvent réutilisées ou qui vont probablement être réutilisées, pour accélérer les temps de réponse). Mais pour économiser de la place et améliorer les temps de réponse, la plupart des moteurs font des économies en ne stockant pas tous les signaux pour les pages les moins susceptibles d'apparaître sur des requêtes concurrentielles... Ce sont justement les pages qui apparaissent généralement sur les requêtes longue traine, celles qui ont le moins de PageRank et de backlinks...

La déclaration de Mayle laisse penser que Google aurait cessé de faire ce genre de différences (qui avait amené Google dans le passé à gérer même deux index totalement différents : l'index primaire, et l'index secondaire).

Mais pour pouvoir gérer un tel changement de comportement, cela signifie que Google dispose d'une architecture permettant de stocker beaucoup plus de données qu'avant et/ou d'extraire des données plus efficacement et plus rapidement qu'avant. C'est justement ce qu'apporte Caffeine, et malgré les dénégations de Google, on se demande quand même si ce changement d'algorithme aurait été possible sans ... Caffeine.

5 MAI 2010 : Google change son interface utilisateur

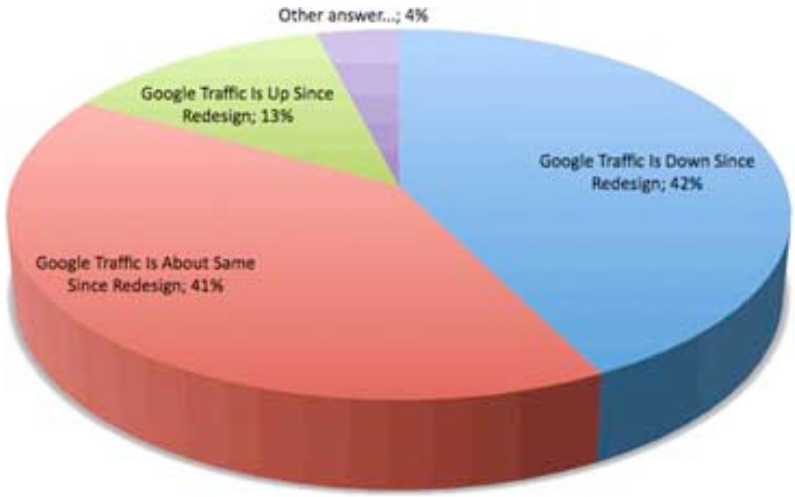
Un autre évènement important est de plus intervenu début Mai : la bascule sur la nouvelle interface utilisateur de Google (surnommée Jazz : http://actu.abondance.com/2010/05/google-annonce-officiellement-son.html).



La nouvelle interface utilisateur de Google implémentée le 5 mai 2010

Les principaux changements introduits par cette nouvelle interface sont la disposition en 3 colonnes fixes et le redimensionnement automatique des éléments sur toute la largeur disponible de la page. De tels changements dans la manière de présenter les options et les informations sur la page sont susceptibles d'entraîner des modifications dans le comportement des utilisateurs, en particulier une nouvelle répartition des clics entre les différentes zones de la page.

Il est clair que certains webmasters, qui ont vu leur trafic issu de Google changer radicalement début mai attribuent peut-être à tort à la mise à jour MayDay des phénomènes en rapport avec la nouvelle interface.



Les résultats d'un sondage (non scientifique) effectué sur le site Seroundtable.com début mai. 42% des personnes qui ont répondu pensent que leur trafic a diminué depuis le changement de design. Mais est-ce l'effet de MayDay, de Caffeine, ou de l'interface Jazz ?

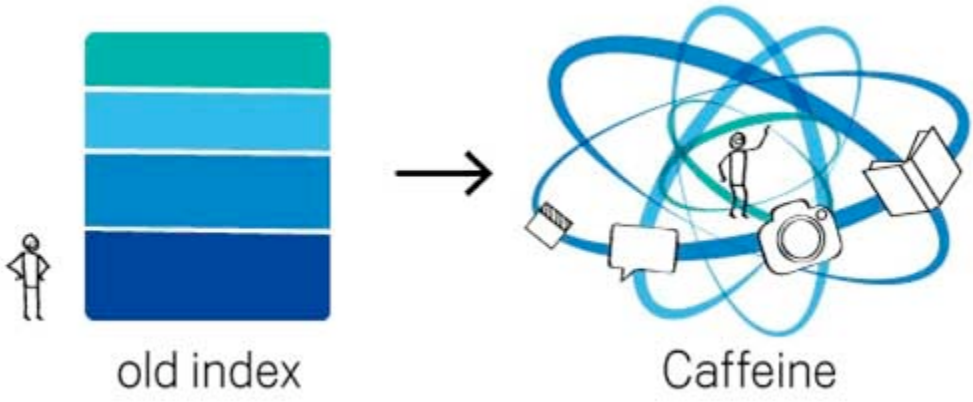
8 JUIN 2010 : Caffeine est officiellement implémenté partout

Caffeine est le nom d'une nouvelle infrastructure de Google, dont le déploiement avait été annoncé dès l'été 2009. Il ne s'agit donc pas d'un changement dans la manière dont Google classe les pages, mais dans la manière dont Google explore le web pour en extraire l'information, dans la manière dont le moteur analyse, stocke et indexe ces données, et dans la manière dont Google traite les requêtes.

Nous avons déjà traité en détail Caffeine et les autres éléments de l'infrastructure comme BigTable, et le Google File System dans un article précédent publié dans cette même lettre en septembre 2009. Nous ne reviendrons donc pas sur cet aspect, pour analyser plutôt l'impact réel de Caffeine sur le comportement de crawl d'une part, et sur les classements d'autre part.

Caffeine : un nouveau comportement de crawl

Dans l'annonce officielle publiée sur le blog de Google destiné aux webmasters (http://googlewebmastercentral.blogspot.com/2010/06/our-new-search-index-caffeine.html), apparait un petit graphe et un court commentaire, qui révèle un changement profond dans le comportement de crawl de Google.

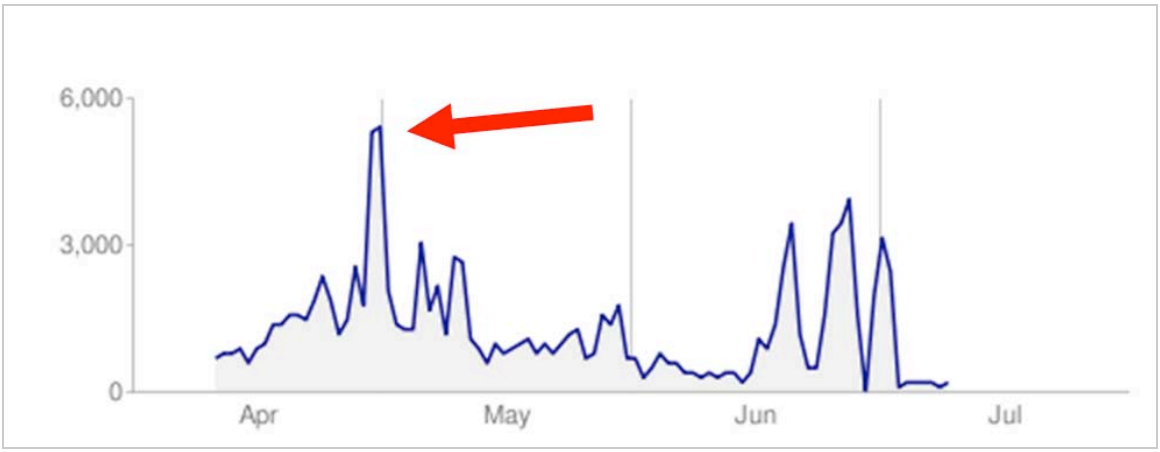


Selon le billet officiel, l'ancien index (pré-Caffeine) était structuré en différentes couches. Pour construire chaque couche, il était nécessaire de crawler l'ensemble du web à chaque fois. Certaines couches étaient mises à jour de manière plus fréquentes que d'autres, et la couche la plus importante était mise à jour toutes les deux semaines environ.

Dans le comportement de crawl post-Caffeine, le web est analysé par petits bouts, et les mises à jour se font de manière continue partout dans le monde.

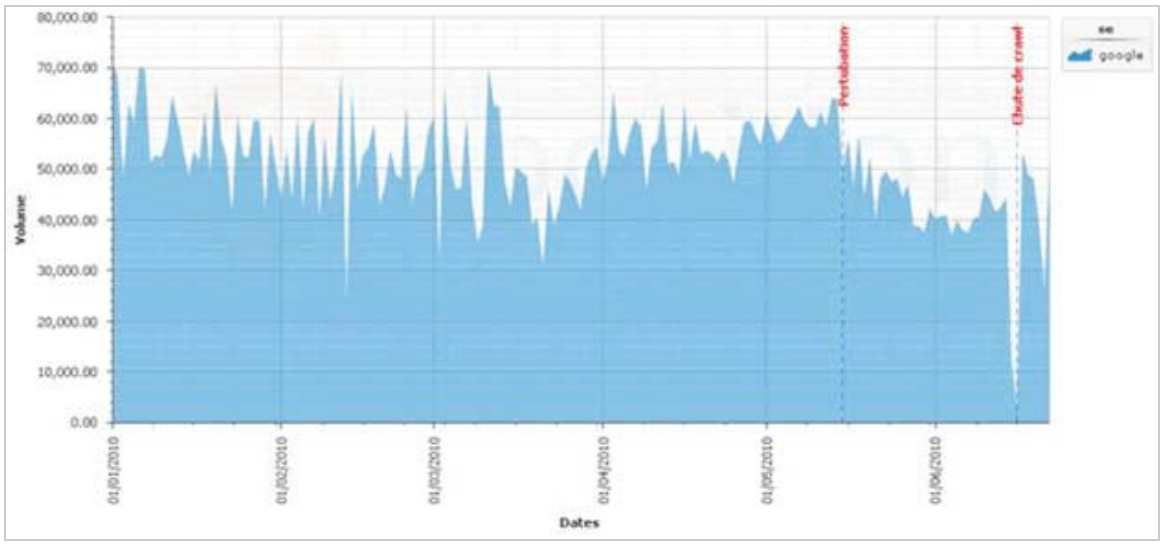
Des changements de crawl sont observables

Dans les faits, des changements de comportement de crawl sont observables. Un pic de crawl semble être intervenu sur de très nombreux sites quelques heures avant Mayday : fallait-il compléter les données ?



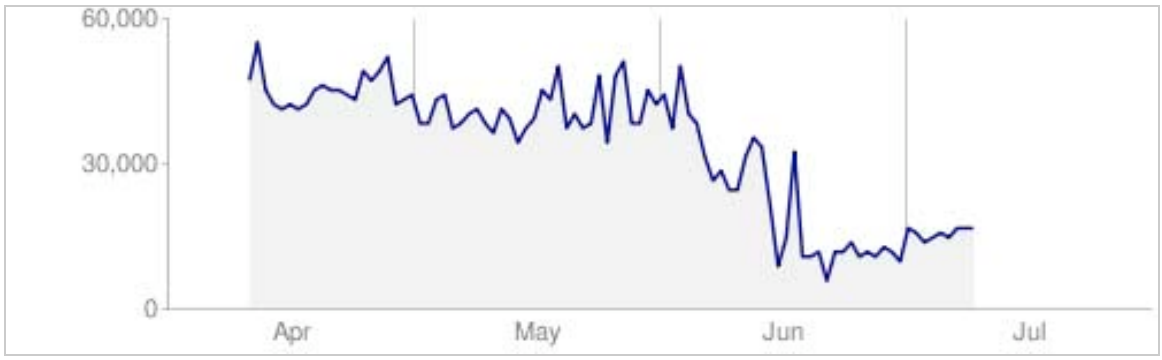
Nombre de pages crawlées par jour sur un site allemand :
graphe issu de Google Webmaster Tools.

D'autres sites ont vu les crawlers de Google faire grève subitement à la mi-juin. L'exploration des sites est revenue à la normale dans les jours qui ont suivi.



Exemple de courbe de crawl issue d'un serveur Apobox (appliance de suivi des logs commercialisée par l'agence Aposition). La chute de crawl du 14 juin est très visible, et a été observée par de très nombreux webmasters.

Mais pas pour tout le monde : sur certains sites, un nouveau comportement de crawl est apparu, Googlebot se mettant à crawler dix fois moins de pages qu'autrefois. Ce nouveau comportement de crawl semble très "ciblé" : certaines pages sont crawlées très souvent, d'autres moins souvent. Mais au total, Google semble crawler moins de pages. Les pages "statiques" sont crawlées de loin en loin, et restent plus souvent dans l'index. Toutefois, il semble que le comportement de crawl évolue encore, comme le montre la courbe ci-dessous issue des GWT (Google Webmaster Tools).



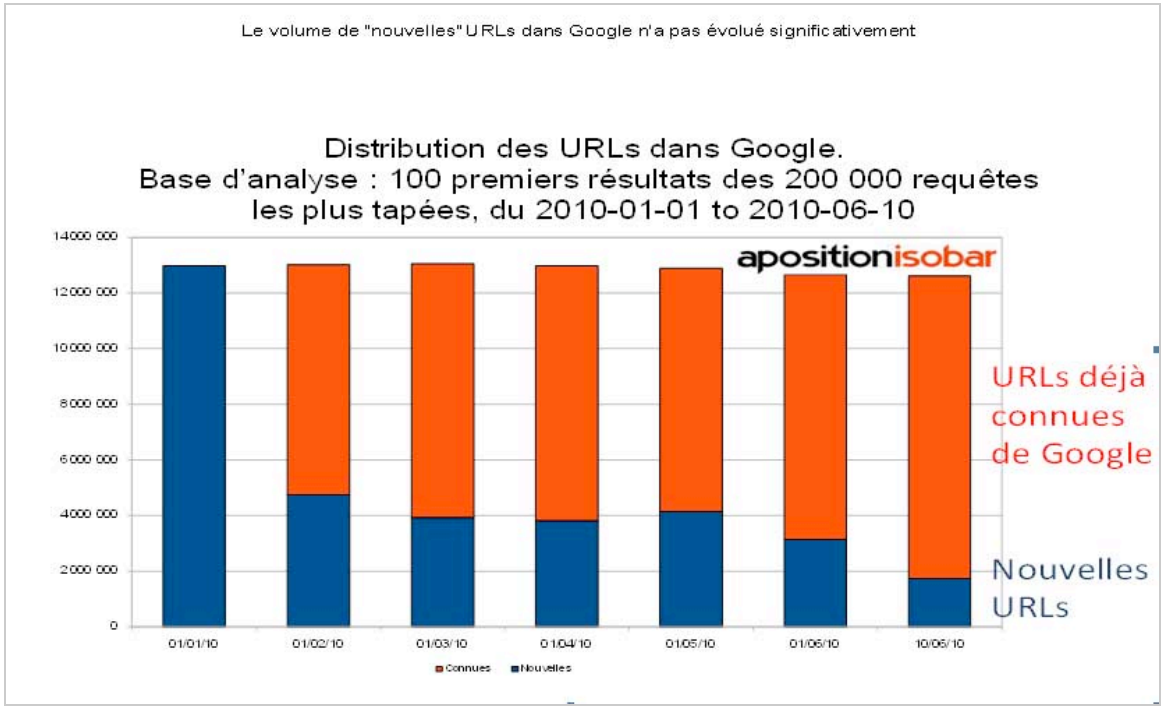
Courbe issue de GWT pour un site d'annonces français généraliste :
on voit l'arrêt du crawl le 14 juin, la reprise, puis la rechute à un niveau très bas avec une reprise lente et progressive. Le nouveau comportement de crawl semble plus ciblé et intelligent : est-ce Caffeine ? Autre chose ?

Un autre exemple avec un site d'annonces français du secteur immobilier qui, lui, est crawlé beaucoup plus souvent qu'avant et plus profondément. Dans ce cas, on voit l'arrêt du crawl le 14 juin et un autre le 28 juin, mais pas de pic "pré-MayDay".

Il semble que le comportement de crawl ait changé après le 14 mai, sans que de véritables changements aient été apportés sur le site.

L'index est-il plus frais après Caffeine ?

Une étude menée par Aposition et qui a été publiée sous forme de video humoristique adressée à Matt Cutts a démontré que cela ne changeait pas vraiment l'âge des pages (*is Google Caffeine a decaf? Message to Matt Cutts (Google)* : voir lien en reference à la fin de l'article).

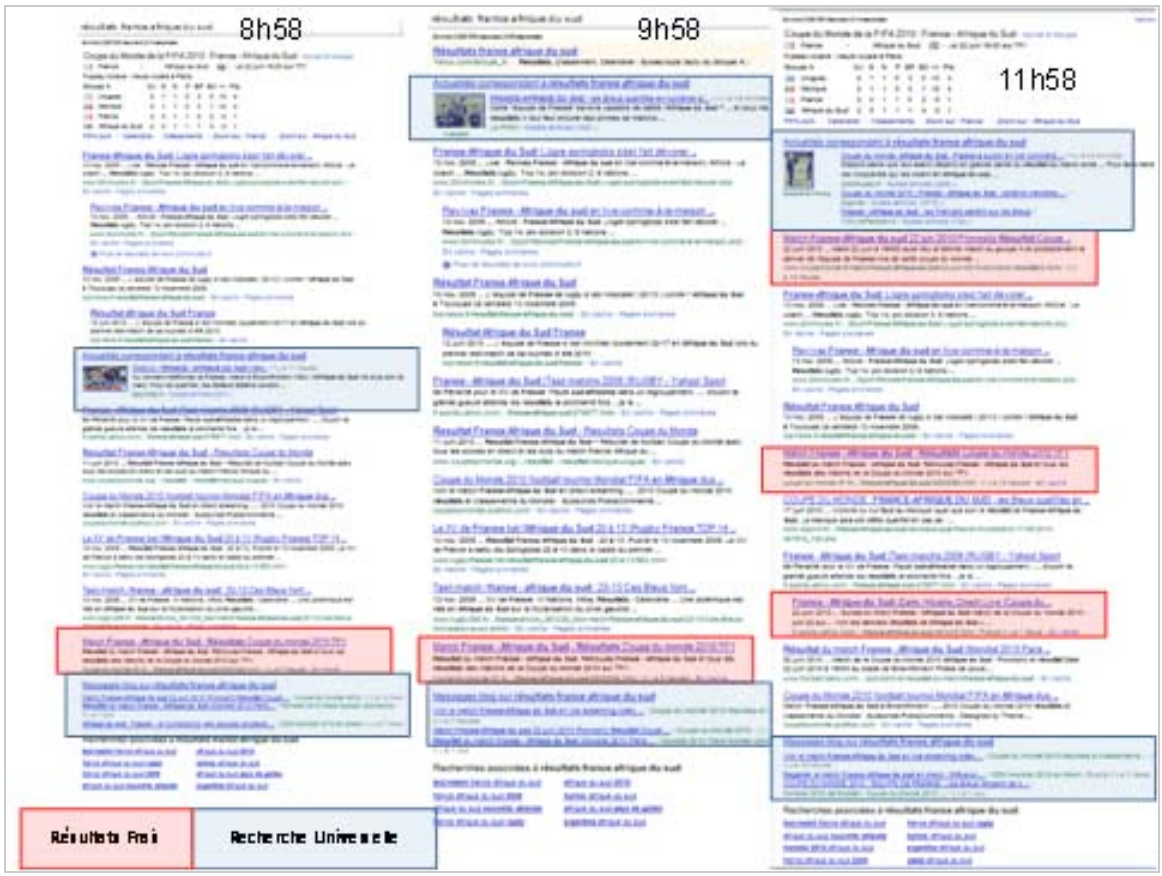


Matt Cutts a publié un commentaire sur cette video. Il précise qu'avec Caffeine, les résultats mettent moins de temps à entrer dans l'index, et c'est pour cela qu'ils sont plus frais.

Mais de quelle fraîcheur faut-il parler ? De l'âge des pages ? De la proportion des pages "à jour" figurant dans l'index ? Ou de la capacité à être réactif en cas d'évènements ?

Sur ce dernier point, un autre test mené par l'équipe d'Aposition semble démontrer que oui, Caffeine semble changer quelque chose.

Prenons par exemple la requête "France Afrique du Sud" correspondant à un évènement prévu le 16 juin à 16h (c'est le premier match de l'équipe de France à l'occasion de la Coupe de Monde de Football, précisons-le pour ceux qui auraient vécu sur la planète Mars ces dernières semaines).



Les résultats frais sont surlignés en rouge.
Les résultats issus de la recherche universelle en bleu.
Résultats observés à différents moments pendant la matinée du 16 juin 2010.

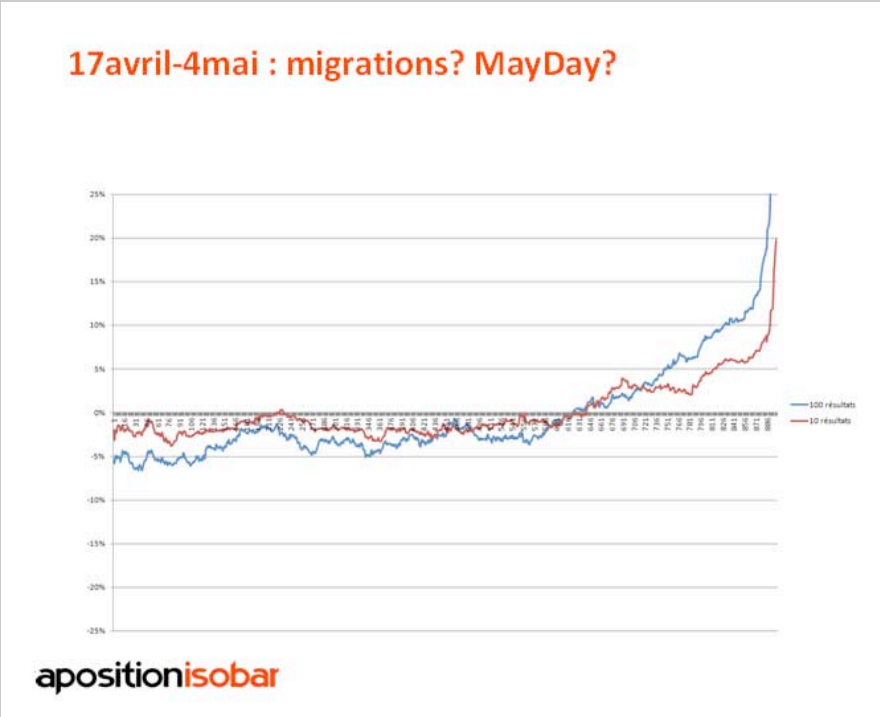
Et pendant le match lui même! Des résultats de recherche universelle apparaissent, mais aussi des pages très récentes dans l'index "normal"

Comparé au comportement passé, l'intégration de nouvelles pages semblent effectivement s'effectuer à un rythme clairement accéléré avec Caffeine.

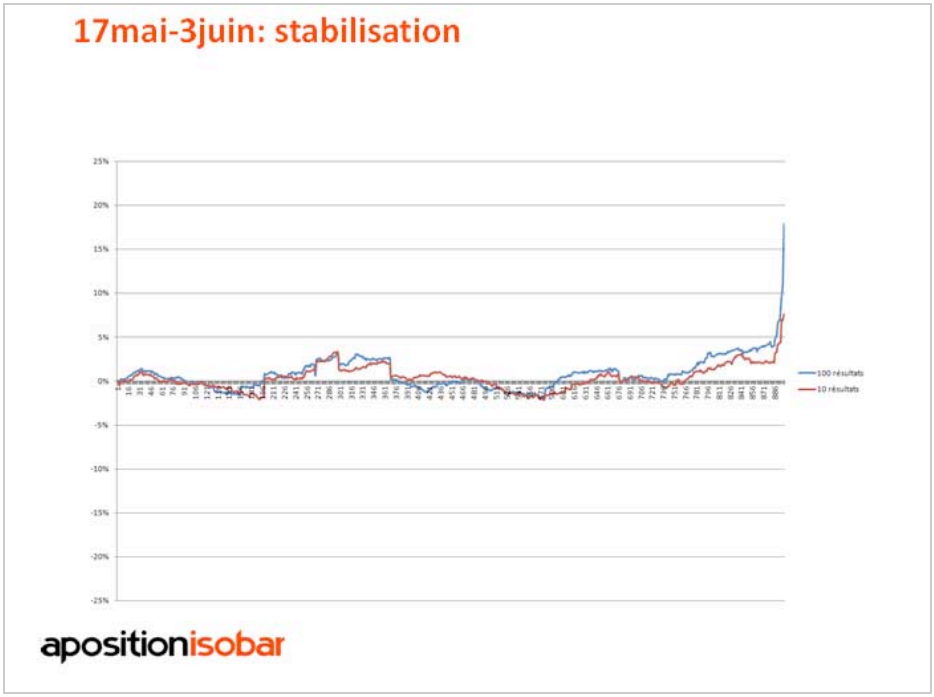
Caffeine a-t-il chamboulé les classements ?

Si l'on observe l'évolution des classements sur les six derniers mois, on s'aperçoit qu'à deux reprises, ceux-ci ont été chamboulés. Est-ce lié à Caffeine ? C'est une hypothèse…

Une analyse des positions mesurées par Aposition sur les 200 000 expressions clés les plus tapées sur l'index Google.fr, et pour les 500 sites ayant le plus de positions classées parmi les cent premiers résultats, révèle une chronologie intéressante :
• Entre le 20 décembre et le 22 janvier, les changements de position correspondent au "bruit" habituel ;
• Entre le 22 janvier et le 18 février : on voit apparaître une première série de changements importants ;
• Entre le 18 février et le 17 mars : retour à la normale ;
• Entre le 17 mars et le 5 avril : les classements sont à nouveau chamboulés ;
• Entre le 5 avril et le 17 avril : les changements continuent ;
• Entre le 17 avril et le 4 mai : les changements s'amplifient ;
• Les changements se poursuivent jusqu'au début du mois de juin pour se stabiliser ensuite.



Evolution des positions pré et post-Mayday : la courbe dresse la répartition statistique des variations de positions entre les deux dates, à partir de calculs de moyennes mobiles, pour les 500 sites français les plus visibles. Lorsque les classements sont stables, les courbes restent proches de l'abscisse, lorsqu'ils évoluent beaucoup, les courbes s'en éloignent. Le pic à droite de la courbe est un artefact du à la méthode de calcul.



Les chamboulements dans les classements semblent diminuer dans les semaines qui ont suivi

… pour aboutir à un nouveau léger gain de positions en juin et une plus grande stabilité des résultats.

Ces résultats amènent évidemment leur lot de questionnements...

Ces changements de positions sont ils liés à Caffeine ?

Peut-être. Cela semble en tout cas cohérent avec le calendrier normal de déploiement de la nouvelle infrastructure. Mais cela peut-être le reflet d'autres changements dans Google.

S'agit-il de l'update Mayday ?

Pour une partie de ces changements, c'est possible. Mais pour la plupart d'entre eux, sûrement pas, ces données montrent que les résultats de Google ont changé bien avant MayDay et après la mise à jour.

La situation Post MayDay – Post Caffeine : une nouvelle donne pour le référencement

Il est clair que tous ces changements redistribuent, même légèrement, les cartes.

Tout d'abord, le nouveau comportement de crawl de Google tend à favoriser les sites d'actualité, qui peuvent voir leur contenu (mais aussi leurs images et leurs videos) être découverts beaucoup plus rapidement et être indexés, si besoin est, dans les minutes qui suivent leur publication. A l'inverse, certains sites risquent de subir des problèmes d'indexation dus à un crawl plus paresseux et plus sélectif. Pour ces derniers, en cas de chute du trafic, il sera indispensable d'analyser l'évolution de la liste des pages qui reçoivent du trafic, la quantité de trafic apporté par telle ou telle catégorie de requêtes, ainsi que la liste des pages explorées par Googlebot et la fréquence de recrawl de ces mêmes pages. Ces analyses peuvent vous permettre de comprendre quelles sont les pages que privilégient Google sur votre site et de corriger éventuellement le tir si certaines pages importantes sont "oubliées" par le moteur.

Ensuite, il devient plus difficile de se positionner sur des requêtes longue traine. Visiblement, avoir une page qui "matche" sur la requête accompagné de quelques backlinks internes ne suffit plus. Google nous indique que la "qualité" des sites et des pages est mieux prise en considération pour les requêtes de la "longue traine". S'agit-il d'avoir un contenu unique ? S'agit-il d'avoir plus de contenu, un contenu plus utile et informatif ? Ou d'être la page ayant le plus "d'autorité" sur la requête, ce qui peut vouloir dire avoir plus de backlinks externes pointant directement vers cette page que les concurrents ? A ce stade, aucune logique claire n'a encore été dégagée. Disons qu'améliorer son contenu sur ces différents points ne peut qu'améliorer vos classements.

Par ailleurs, avec l'arrivée de Caffeine, il est important de surveiller mieux le temps de téléchargement de ses pages, car Google a déclaré officiellement que ce critère était à présent intégré dans l'algorithme de classement (et qu'un temps de téléchargement trop élevé pénalise le crawl).

Est-ce la fin des changements ?

Beaucoup de choses ont donc changé dans Google ces dernières semaines, et même ces derniers mois. D'une manière générale, le rythme des changements s'était déjà accéléré dès l'été 2008, pour culminer au cours du premier semestre de cette année. La plupart des changements attendus sont à présents déployés. Est-ce le début d'une période d'accalmie et de stabilité ? Rien n'est moins sûr.

D'abord, les changements importants qui viennent d'être apportés au moteur de recherche vont nécessiter probablement toute une série de calages, de réglages et de corrections.
Ensuite, on peut être sûr que les possibilités ouvertes par Caffeine pour traiter "plus de données plus vite" vont rendre possible le déploiement de nouvelles fonctionnalités.
Et enfin la nécessité de garder de l'avance sur des concurrents comme Bing semble pousser de plus en plus Google à lancer des nouveautés sur son interface et à essayer de nouvelles approches.

Pour les référenceurs, l'environnement change à un rythme accéléré maintenant, et il faut s'habituer à l'idée que, d'un mois sur l'autre, il faille imaginer de nouvelles stratégies et savoir réagir à un changement de comportement du moteur. Mais cela rend l'exercice encore plus intéressant.

REFERENCES ET LIENS COMPLEMENTAIRES

Le thread de Webmasterworld à propos de MayDay

http://www.webmasterworld.com/google/4125460.htm

Une video de Matt Cutts à propos de la mise à jour MayDay
http://www.youtube.com/watch?v=WJ6CtBmaIQM

Billet de Vanessa Fox sur SearchEngineLand
http://searchengineland.com/google-confirms-mayday-update-impacts-long-tail-traffic-43054

Our New Search : Caffeine (billet du blog officiel de Google)
http://googleblog.blogspot.com/2010/06/our-new-search-index-caffeine.html

is Google Caffeine a decaf? Message to Matt Cutts (Google)
http://www.youtube.com/watch?v=8bTjDTDIZ00

Philippe Yonnet, *Global SEO Strategist, WEB DMUK (Londres) – Easyroommate / Vivastreet*
Jérôme Spiral, *Directeur Associé et responsable R&D de l'agence Aposition*

Réagissez à cet article sur le blog des abonnés d'Abondance :
http://blog-abonnes.abondance.com/2010/07/limpact-des-mayday-jazz-et-caffeine-sur.html