

NETSOURCES

S O M M A I R E



MÉTHODOLOGIES

Peut-on se fier à Google ?

Enquête sur des résultats étranges

Béatrice Foenix-Riou

Avec ses lettres de couleur et son écran sobre, son critère de classement innovant et pertinent et son index volumineux, Google a su en quelques années s'imposer pour devenir la référence en terme de recherches sur Internet. A tel point que l'on voit se répandre sur le Web des néologismes comme le verbe "googler" ("to google" en anglais), qui signifie "chercher sur un moteur – Google le plus souvent – des renseignements sur quelqu'un ou quelque chose"...

Les statistiques de Xiti⁽¹⁾ sont sans appel : en France, en décembre 2005, Google était à l'origine de plus de 82 % du trafic généré par les outils de recherche !

Néanmoins, cette hégémonie du moteur sur le monde de la recherche d'information n'est pas sans conséquences. Les internautes ont désormais le "réflexe Google" et n'ont plus le recul nécessaire pour juger de la pertinence de ses résultats.

Pour beaucoup, si l'information n'est pas sur Google, c'est qu'elle n'existe pas sur le Net et il est donc inutile de poursuivre la recherche sur un autre moteur...

Hors, si Google possède des qualités indéniables, il n'est pas sans défauts. Il lui arrive en particulier de présenter des signes temporaires de dysfonctionnement, que ce soit dans l'utilisation des opérateurs booléens⁽²⁾ comme dans celui des opérateurs *intitle:* et *inurl:*.

Plus grave encore, nous avons remarqué ces dernières semaines, pour des séries de questions de plus en plus précises que nous posons périodiquement – tant pour nos formations "*Web visible et invisible : outils et méthodes pour optimiser vos recherches*" que pour la rubrique *Méthodologies de recherche* de *Netsources* – des divergences considérables et inexplicables dans les résultats.

Méthodologie de recherche

- Peut-on se fier à Google ? Enquête sur des résultats étranges, pp.1-6

Sur la Toile

- Controverses autour de l'encyclopédie Wikipedia.org, pp.10-11
- NatioMaster.com : la macro-éco à dispo., p.14

Outils de recherche

- Polymeta.com : les Hongrois à l'assaut du Web, pp.12-13

Surf sur le Net

- Tech Support Alert : les 46 meilleurs freewares, p.8
- Le blog d'Abondance : des billets sur l'évolution des moteurs, p.9

Agenda

- Veille sur le Net : outils et méthodes pour automatiser votre surveillance, p.6
- Web visible et invisible : outils et méthodes pour optimiser vos recherches, p.6
- Journée d'information DADVSI, p.9

INDEX 2005, p.15



Nous avons donc mené l'enquête pour tenter de comprendre le phénomène... Nous illustrerons l'investigation par la comparaison des réponses affichées à des dates différentes pour le traitement de la question suivante : "Comment dresser un rapide panorama des principales ressources en français sur l'obésité".

A l'origine (25 août 2004) : un article pour Netsources...

Le traitement de cette question a fait l'objet d'un article intitulé "Trucs et astuces pour un chasseur de liens", paru dans le n°51 de Netsources (juillet/août 2004) et disponible en accès libre sur notre site www.bases-publications.com.

Nous expliquions dans cet article que pour ce type de problématique – identifier une sélection de ressources sur un sujet d'ordre plutôt général –, il était intéressant de rechercher spécifiquement des "pages de liens", c'est-à-dire des pages Web listant une sélection de sites choisis pour leur couverture du sujet. Ces pages s'avèrent en effet très utiles pour identifier les sites de référence sur un thème, mais elles sont difficiles à localiser.

Nous avons alors présenté plusieurs méthodes permettant d'identifier ces ressources, parmi lesquelles celle qui consiste à rechercher sur un moteur les pages qui contiennent, en plus du mot de la requête – ici *obésité* –, le mot *liens* dans leur titre ou leur URL.

La démonstration s'est faite sur Google, qui donnait, le 25 août 2004, les résultats suivants :

- *obésité* : 208 000 documents ;
- *intitle:obésité* : 48 000 documents ;
- *obésité intitle:liens OR inurl:liens* : 460 documents.

Les résultats obtenus à chaque étape étaient cohérents entre eux et l'on identifiait au final plusieurs pages de liens.

L'une d'entre elles, issue du site Gros.org, listait par exemple une vingtaine de sites en donnant un résumé de leur contenu, une autre recensait une liste de sites couvrant le domaine de la nutrition et de la santé au sens large, etc.

7 décembre 2005 : évolution cohérente des résultats

Le 7 décembre 2005, soit plus d'un an et trois mois après la rédaction de l'article, nous avons mis à jour cette recherche pour la proposer en exercice lors d'une formation. Les mêmes questions ont donné les résultats suivants :

- *obésité* : 501 000 documents, soit une augmentation de 141 % ;
- *intitle:obésité* : 79 000 documents, soit une hausse de 65 % ;
- *obésité intitle:liens OR inurl:liens* : 1 530 documents, soit une croissance de 232 %.

La comparaison des réponses obtenues à plus d'un an de distance pour ces différentes recherches peut surprendre de prime abord. Mais à la réflexion, même si l'on est surpris que les taux de croissance varient autant d'une requête à l'autre, les ordres de grandeur des résultats restent cohérents entre eux et les augmentations peuvent s'expliquer par l'accroissement notable de la taille de l'index de Google.

En effet, lors des premiers tests en août 2004, Google affichait encore "Nombre de pages Web recensées par Google : 4 285 199 774" et il fallut attendre le 11 novembre de la même année – le lendemain du lancement de MSN Search, voir Netsources n°53 –, pour que ce chiffre soit mis à jour et annonce "8 058 044 651 pages"...

L'augmentation suivante eut lieu – ou du moins fut annoncée – le 26 septembre 2005, à l'occasion du septième anniversaire de Google⁽³⁾.

Le moteur, amateur de devinettes, écrivit ce jour-là dans son blog que la taille de son nouvel index était "mille fois plus importante que lors de son lancement". Or, si l'on se réfère à la publication des deux créateurs de Google présentant le prototype du moteur⁽⁴⁾, ce dernier fut lancé avec un index de 24 millions de pages.

La taille du nouvel index de Google serait donc de 24 milliards de pages (voir Netsources n°57), soit une croissance annoncée de 200 % entre novembre 2004 et septembre 2005.

La variation du nombre de réponses obtenues lors de nos deux tests est donc tout à fait plausible.

Un mois plus tard : des résultats qui jouent au yoyo

En revanche, les résultats des mêmes recherches, effectuées un mois après le deuxième test – et répétées plusieurs fois au cours des deux premières semaines de janvier –, sont quant à eux beaucoup moins cohérents.

La requête lancée sur le simple mot *obésité* a en effet obtenu le 13 janvier 2 910 000 documents, soit une augmentation de 481 % en un mois, ou de 1 299 % en un an et quatre mois !

Les chiffres annoncés semblent avoir "explosé" en un temps très court et ce, non seulement pour cette recherche, mais aussi pour la majorité des questions que nous posons régulièrement :

- le nombre de réponses à la requête "futian pharmacy" est ainsi passé de 440 le 7 décembre à 1 690 le 13 janvier (soit une augmentation de 374 %) ;
- "laboratoires pharmaceutiques" R&D OR "recherche et développement" a obtenu respectivement à ces deux dates 17 300 et 61 700 réponses (+274 %) ;



• *statistiques "équipement informatique" ménages* est passé de 12 200 pages le 7 décembre à 23 200 pages le 13 janvier (+ 90 %)...

Peut-on pour autant en déduire que l'index de Google a augmenté dans ces proportions, ou que son annonce de septembre était prématurée et que l'augmentation n'a eu lieu de façon assez brutale qu'en décembre ?

Non car, outre le fait que ces tests sont bien trop limités pour permettre de déduire quoique ce soit, il s'est avéré que :

■ lorsque la requête était simple (un ou plusieurs mots reliés par AND ou OR), les mêmes tests répétés sur plusieurs jours ont obtenu des résultats qui pouvaient être très différents. Ainsi, la recherche sur le mot *obésité* a pu afficher – quelquefois dans la même journée – un nombre de réponses variant entre 1 820 000 et 4 330 000 ; de la même façon, le nombre de réponses à "*futian pharmacy*" a oscillé entre 144 et 1 690 !

■ à l'inverse, lorsque la requête utilisait un opérateur (*intitle:*, *inurl:*...), le nombre de réponses restait relativement stable et présentait une croissance bien plus raisonnable – voire une décroissance – par rapport aux tests précédents. Par exemple, *intitle:obésité* obtint 81 200 résultats, soit une croissance de 2,8 % en un mois ou de 69,2 % en un an et quatre mois et *obésité intitle:liens OR inurl:liens* afficha 756 réponses, soit une diminution de 51 % en un mois ou une croissance de 64 % en un an et quatre mois...

Nous nous sommes alors remémorés les "mésaventures" que nous avons rencontrées avec Google en 2003 et qui inspirèrent l'article "*Google : quand les "data centers" n'en font qu'à leur tête...*"⁽⁵⁾

Des tests répétés pendant plusieurs semaines avaient montré que le fonctionnement des opérateurs *intitle:* et *inurl:*

s'avérait à l'époque aléatoire et que le moteur affichait sporadiquement des résultats montrant clairement qu'il n'avait pas tenu compte de la limitation de champ.

Pour tenter de comprendre le phénomène, nous avons testé le fonctionnement des différents centres de données (*data centers*) de Google et nous avons mis en évidence que seul l'un d'entre eux semblait fournir des résultats logiques.

Il faut en effet savoir que Google possède la particularité de découper son index en morceaux sur des milliers d'ordinateurs, hébergés dans plusieurs centres de données, chacun des centres disposant d'une copie de la totalité de l'index. Ces centres regroupent des serveurs traditionnels, disposant de leurs propres adresses IP. Le *data center* de Washington DC par exemple est accessible depuis les adresses <http://216.239.39.99>, <http://216.239.39.104>...

Les centres les plus anciens sont localisés aux Etats-Unis (Californie, Virginie,

Washington DC...) et d'autres ont été ouverts en Suisse (Zurich) et en Irlande (Dublin). Google continuant son expansion, on compte aujourd'hui au moins dix-huit centres de données, alors qu'on en recensait neuf en mai 2003.

Lorsque l'internaute interroge Google, le moteur dirige automatiquement la question vers l'un des centres de données, choisi comme étant le plus rapide en fonction de l'encombrement, de la localisation géographique, etc.

Mais les index des centres ne sont pas toujours identiques, ce qui explique qu'une même question sur Google peut obtenir, le même jour et depuis un même poste, des résultats variables selon qu'elle est posée à un centre ou à un autre.

Au cœur des data centers

Nous avons donc interrogé directement les différents centres de données, afin de comparer les résultats de chacun à nos questions-tests.





Nous avons utilisé pour cela le précieux **Outil d'analyse des data centers de Google**, offert sur le site **Webrankinfo.com**⁽⁶⁾.

Conçu pour les webmasters et les référents, cet outil permet de lancer simultanément une recherche sur les différents centres de données de Google. Son principal objectif était à l'origine de suivre la répercussion de la *Google Dance* sur chacun des *data centers* – la *Google Dance* étant la période pendant laquelle le moteur recalcule le *PageRank* de toutes les pages de son index.

Cette *Google Dance*, qui était au départ mensuelle, se fait désormais de façon beaucoup plus continue.

Mais l'*Outil d'analyse des data centers de Google* reste fort utile lorsque l'on souhaite – c'est le cas ici – vérifier la cohérence de certains résultats du moteur.

Le tableau ci-dessous confirme nos premières impressions : lorsque la requête contient un opérateur (*intitle;*, *inurl:*...), les résultats des différents centres de données sont très proches, à l'exception des réponses fournies par le premier centre, baptisé Bigdaddy. En revanche, une recherche "simple" peut générer un nombre de réponses très variable selon le centre interrogé.

Si les réponses variables des centres de données restent incompréhensibles et sont sans doute le signe d'un dysfonction-

nement de certains serveurs de Google, les résultats propres à Bigdaddy s'expliquent en revanche très bien.

En officialisant le lancement de ce centre sur son blog le 4 janvier⁽⁷⁾ Matt Cutts, l'un des ingénieurs de Google, nous a appris que ce *data center* fournissait des résultats différents de ceux des autres centres. Bigdaddy repose sur une nouvelle infrastructure et utilise à la fois des données et un algorithme de classement qui lui sont propres ; il prépare en fait le cadre des améliorations qui seront apportées à Google au cours de l'année 2006.

D'après Matt Cutts, les résultats de Bigdaddy devraient se retrouver d'ici un mois ou deux sur les autres *data centers*.

TESTS EFFECTUES LE 13 JANVIER SUR L'OUTIL D'ANALYSE DES DATA CENTERS (Webrankinfo.com)

(www.webrankinfo.com/outils/google-dance/google-dance3.php)

	<i>obésité</i>	<i>obésité intitle:liens OR inurl:liens</i>	<i>intitle:obésité</i>	<i>intitle:obesite</i>	<i>consommation carburants</i>	<i>"futian pharmacy"</i>
66.249.93 (Bigdaddy)	1 820 000	756	81 200	12 800	2 180 000	1 690
216.239.37 (www-va)	2 910 000	543	54 600	9 820	2 240 000	463
216.239.39 (www-dc)	2 910 000	543	54 600	9 820	2 240 000	463
216.239.53 (www-in)	4 330 000	553	55 000	9 790	1 640 000	465
216.239.57 (www-cw)	4 330 000	553	55 000	9 790	1 640 000	465
216.239.59 (www-gv)	4 370 000	545	54 700	9 840	1 400 000	144
216.239.63	3 240 000	553	55 000	9 790	1 590 000	465
64.233.161	2 910 000	546	54 600	9 820	2 240 000	463
64.233.167	3 590 000	546	54 700	9 840	1 420 000	463
64.233.171	2 490 000	545	54 700	9 840	1 460 000	463
64.233.179	2 490 000	545	54 700	9 840	1 460 000	463
64.233.183	3 760 000	553	55 000	9 790	1 490 000	146
64.233.185	2 490 000	545	54 700	9 840	1 460 000	463
64.233.187	2 540 000	546	54 700	9 840	1 380 000	463
64.233.189	4 330 000	553	55 000	9 790	1 590 000	465
66.102.7 (www-mc)	4 330 000	553	55 000	9 790	1 640 000	465
66.102.9 (www-lm)	2 620 000	539	51 000	9 610	1 330 000	144
66.102.11(www-kr)	2 620 000	539	51 000	9 610	1 330 000	144



Cette information apporte un éclairage nouveau aux recherches sur Google.

L'outil de Webrankinfo nous a en effet permis de vérifier que lors des tests effectués depuis Google.fr, les résultats obtenus étaient souvent issus de Bigdaddy. Or, la comparaison de ses résultats avec ceux des autres centres de données montre une variation importante de la quantité, mais aussi de la qualité des informations fournies.

Sur le point de la quantité déjà, nos tests sur *"futian pharmacy"* ont obtenu des résultats proches de 144 sur 4 *data centers*, de 460 sur 13 centres et 1 690 sur Bigdaddy – pour comparaison, la même recherche sur Yahoo! a affiché 796 réponses.

Sur le plan de la qualité, les résultats diffèrent aussi : le site de la société Futian Pharmacy (www.xylyitol-cn.com) était ainsi affiché en 2ème position par Bigdaddy (et par Yahoo!), mais n'apparaissait qu'en ... 39ème position sur les autres centres – et encore, c'est la page *"Produits"* et non la page d'accueil du site qui était présentée !

On peut donc se demander s'il ne serait pas judicieux, tant que les résultats de Bigdaddy n'alimentent pas les autres centres, d'interroger Google depuis l'adresse IP de Bigdaddy (<http://66.249.93.104>) et non depuis l'interface Google.com ou Google.fr ...

Cela étant, si Bigdaddy permet de comprendre les différences de résultats obtenus pour une même question à plusieurs moments de la journée, il n'explique pas l'accroissement considérable du nombre de réponses en un mois, pour des questions "simples"... Peut-être trouverons-nous ultérieurement des explications sur le blog de Google ou sur celui de Matt Cutts...

Nos interrogations ne se limitant pas au nombre de résultats de Google, nous avons poursuivi nos investigations et remarqué d'autres dysfonctionnements.

Les opérateurs n'aiment pas les accents

Alors que le moteur précise clairement dans son aide en ligne que *"Par défaut, les recherches sur Google ne tiennent pas compte des accents ou autres signes diacritiques"*, il semble que la prise en compte des accents se fasse en réalité de façon différente, selon que l'on utilise ou non un opérateur.

Le plus souvent, le nombre de réponses diffère légèrement lorsque l'on écrit le mot-clé avec ou sans accents – en fait, la différence apparaît lorsqu'un centre de données autre que Bigdaddy est interrogé – mais dans les deux cas, les résultats contiennent à la fois les occurrences du mot avec et sans accents.

On notera cependant – nous l'avions signalé dans le n°49 de Netsources – que le comportement de Google peut varier lorsque la recherche se fait sur un mot "rare" (pour lequel il y a peu de réponses).

En revanche, lorsque le mot est recherché dans un champ particulier du document – en utilisant par exemple l'opérateur *intitle:* ou *inurl:* –, la prise en compte des accents est différente et ce, quel que soit le centre de données interrogé.

Ainsi, la recherche *intitle:obésité* a affiché 81 200 réponses, quand *intitle:obesite* n'en a obtenu que 12 800 ; on notera que l'on obtient les mêmes résultats, que l'on saisisse l'opérateur *intitle:* ou que l'on utilise la grille de recherche avancée.

De la même façon, *inurl:obesite* a sélectionné 33 500 réponses, quand *inurl:obésité* en a identifié 758.

Mais contrairement à ce qui se passe lors d'une recherche sur l'intégralité de la page, **les résultats semblent montrer que les accents sont pris en compte de façon stricte, lorsque la recherche est limitée au titre ou à l'URL.**

Cette impression est confirmée par des tests sur des mots rares, qui comprennent peu de réponses et permettent donc de vérifier chacune. A titre d'exemple, *intitle:caramélisation* n'identifie QUE 19 pages contenant le mot accentué et *intitle:caramelisation* n'affiche QUE 13 pages ayant le mot non accentué dans leur titre.

C'est là un grave dysfonctionnement, quand on sait que les internautes saisissent en majorité les mots sans accents...

Quand Google tronque, les internautes trinquent !

Autre défaut, non moins grave : Google utilise la troncature de façon implicite, mais selon son bon vouloir seulement.

Dans son aide en français, Google indique que *"Pour garantir des résultats aussi précis que possible, Google n'applique pas de « lemmatisation » (réduction des mots au masculin et/ou au singulier, à l'infinitif, etc.) et ne supporte pas les recherches à base de caractères joker/wildcard. Autrement dit, Google utilise les mots exactement tels que vous les entrez dans le champ de recherche."*

L'aide de la version internationale en revanche n'est pas aussi catégorique, puisqu'elle indique depuis deux ans (voir Netsources n°48) *"Google now uses stemming technology. Thus, when appropriate, it will search not only for your search terms, but also for words that are similar to some or all of those terms."*

If you search for pet lemur dietary needs, Google will also search for pet lemur diet needs, and other related variations of your terms. Any variants of your terms that were searched for will be highlighted in the snippet of text accompanying each result."

Google précise néanmoins que *"Sometimes you'll only want results that include an exact phrase. In this case, simply put quotation marks around your search terms."*



Si cette troncature était jusqu'ici utilisée pour des mots en anglais, il semble bien que le moteur s'en serve désormais pour les mots en français, mais quand il le juge approprié et sur certains mots seulement.

C'est du moins ce que laissent supposer les résultats des recherches suivantes, effectuées le 13 janvier 2006 :

- *consommation carburant* : 2 490 000
- *consommation "carburant"* : 2 500 000
- *consommation carburants* : 2 180 000
- *consommation "carburants"* : 664 000
- *consommation carburants OR carburant* : 2 180 000.

Ceci étant, le principe adopté par Google est loin d'être clair, car ces exemples laissent à penser :

- que *carburants* identifie des occurrences au pluriel, mais aussi au singulier – il y a des différences notables lorsque le mot est saisi entre guillemets –, quand *carburant* ne retrouve que celles au singulier ;
- que l'opérateur OR fonctionne de façon pour le moins étrange, puisque l'on obtient davantage de réponses en recherchant le mot uniquement au singulier qu'avec le mot au singulier OU au pluriel !

Quoi qu'il en soit, les quelques tests destinés à comprendre le fonctionnement de Google sur ce point nous ont montré que ... celui-ci était incompréhensible !

A titre d'exemples :

- *"carburant"* : 3 270 000
- *"carburants"* : 5 000 000
- *"carburant" OR "carburants"* : 4 320 000

ou encore :

- *carburant* : 3 310 000
- *carburant consommation* : 2 500 000
- *carburant – consommation* : 2 500 000

Bref, tout n'est pas simple au royaume de Google et il ne faut sûrement pas prendre pour argent comptant le nombre de réponses affichées...

Ces quelques tests – qui démontrent, s'il en était besoin, que le moteur préféré des Français est loin d'être infaillible – nous inspirent deux commentaires :

■ à notre avis, le problème ne réside pas tant dans les inexactitudes du nombre de réponses affichées – nos colonnes ont souvent démontré que Google ou Yahoo! pouvaient sur ce point largement surestimer le nombre réel de résultats –, que dans la confiance aveugle que semble avoir une majorité d'internautes en cet outil, au demeurant fort sympathique...

Notre crainte est que cette "confiance aveugle" les empêche de prendre conscience des limites de l'outil et de sa faillibilité... Or, sans cette prise de conscience, de nombreux internautes risquent de se contenter des 145 réponses trouvées sur "*futian pharmacy*" et de passer à côté de données importantes ! C'est aux professionnels de l'information de les sensibiliser, notamment à la nécessité qu'il y a d'interroger plusieurs outils...

■ une autre crainte tient aux nombreuses diversifications actuellement entreprises par Google (voir Netsources n°55 et 57).

Outre l'élargissement de la couverture de son index (qui contient désormais des vidéos, des blogs, des cartes, des images satellites, des livres...), Google s'intéresse aux photos, aux services emails, au wifi, à la messagerie instantanée, aux logiciels, etc.

Cette volonté de se diversifier tous azimuts, pour attirer dans ses filets une clientèle toujours plus large, ne se fait-elle pas au détriment de la qualité de son service premier, la recherche sur le Web ?

Il est frappant – et désolant – de constater que Google n'hésite pas à se lancer dans des voies peu explorées, mais néglige dans le même temps sa recherche avancée, qui pourrait être grandement améliorée (on aimerait pouvoir utiliser la troncature, disposer de termes associés comme sur

Exalead ou de listes de liens comme sur Teoma...).

De la même façon, les dysfonctionnements rencontrés dénotent une absence de contrôle qualité peu compatible avec l'image du moteur – mais peut-être ces problèmes (récents) vont-ils rapidement disparaître...

En tout état de cause, Google a indubitablement marqué la recherche sur le Web et la recherche d'information tout court.

Mais ce colosse qui domine le marché ne doit pas oublier qu'il a – comme tout moteur de recherche – des pieds d'argile... AltaVista fit en son temps les frais de cet oubli !

A l'heure où l'on annonce l'arrivée – très attendue – du moteur de recherche européen Quaero, le roi Google a tout intérêt à mettre un peu de rigueur dans son fonctionnement !

(1) <http://www.secrets2moteurs.com/barometre2005-12.html>

(2) L'article "*Quand Google ne sait plus compter*", paru dans le n°54 de Netsources (janvier/février 2005), revenait sur diverses incohérences repérées dans les résultats de Google par Jean Véronis et citées sur son blog *Technologies du Langage* (<http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html>)

(3) *We wanted something special for our birthday*. September 26, 2005. <http://googleblog.blogspot.com>

(4) *Sergey Brin and Lawrence Page. The Anatomy of a Large Scale Hypertextual Search Engine*. <http://www-db.stanford.edu/~backrub/google.html>

(5) *Netsources n°44*, mai-juin 2003, en accès libre sur www.bases-publications.com

(6) www.webrankinfo.com/outils/google-dance/google-dance3.php

(7) *Feedback on Bigdaddy data center*. January 4, 2006. www.mattcutts.com/blog/bigdaddy/