

# Enquêtes autour de la taille des moteurs

Béatrice Foenix-Riou

Le 8 août dernier, Tim Mayer, de Yahoo! Search, publiait discrètement dans le blog du moteur<sup>(1)</sup> un billet intitulé *"Our Blog is Growing Up - And So Has Our Index"*.

On y apprenait que l'index de ce dernier venait de croître significativement et donnait désormais accès à plus de 20 milliards de documents. Tim Mayer n'omettait pas de préciser, *"for those who are curious"* – on le serait à moins ! –, que ces "vingt milliards et plus" étaient composés de 19,2 milliards de pages Web indexées, de 1,6 milliard d'images et de 50 millions de fichiers audio et vidéo.

Comme l'on pouvait s'y attendre, cette annonce a suscité diverses réactions dans le monde des outils de recherche.

La surprise tout d'abord, car Yahoo! avait jusque-là refusé de communiquer sur la taille de son index, arguant – à juste titre – que le volume n'était qu'un des éléments de la qualité d'un moteur...

L'étonnement ensuite, au regard de l'importance de la mise à jour ; jusqu'ici en effet, les experts estimaient que Yahoo! indexait environ six milliards de pages.

Annoncer soudainement une base de 19,2 milliards de pages Web signifiait donc que Yahoo! avait plus que triplé la taille de son index, et que celui-ci représentait rien moins que le double de celui de Google !

L'effet de surprise passé, doutes et scepticisme firent leur apparition et engendrèrent tests, études et commentaires...

Différentes méthodes furent à ce titre employées et montrèrent toute la difficulté – voire l'impossibilité – qu'il y a à vérifier les assertions des moteurs...

Ainsi, des chercheurs du NCSA – National Center for Supercomputing Applications – et de l'université de l'Illinois à Urbana-Champaign, ont tenté de comparer la taille des index des deux principaux moteurs : Google et Yahoo!<sup>(2)</sup>

Pour ce faire, ils choisirent au hasard des mots dans le dictionnaire *Ispell* et les présentèrent par couples aux moteurs, afin d'obtenir moins de 1 000 résultats par requête, seule façon de vérifier la pertinence de chacun ; sur Yahoo! comme sur Google en effet, seuls les 1 000 premiers résultats peuvent être affichés.

A partir de l'analyse des réponses à 10 012 requêtes, ils conclurent qu'on obtenait en moyenne ... 166,9 % de résultats supplémentaires sur Google et que lors des 10 012 questions tests, Google avait retourné davantage de résultats dans 96,6 % des cas. Ils émirent donc des suspicions sérieuses sur la taille annoncée de l'index de Yahoo! – suspicions par ailleurs formulées explicitement par un responsable de Google, lors d'une discussion avec John Battelle<sup>(3)</sup>. *"Our scientists are not seeing the increase claimed in the Yahoo! index. The data we have doesn't support the 19.2 (billion page) claim and we're confused by that."*

Conséquence logique, eu égard à la vitesse de propagation de ce genre d'information

sur le Net, Yahoo! fut au cœur de multiples débats ; l'étude fut notamment citée dans la presse (rien moins que *The New York Times*, Associated Press...) et commentée par de multiples sites et blogs, tels que Slashdot, SE Roundtable, Search Engine Watch Blog, John's Battelle SearchBlog, InternetNews...

Si la plupart des sources se contentèrent de reprendre les résultats de l'étude, d'autres en revanche "égratignèrent" sa fiabilité en mettant en cause la méthodologie employée.

Ainsi, dans des billets intitulés *"Yahoo: Pages manquantes ?"* (épisodes 2, 3 et 4) postés sur son blog *Technologies du Langage*<sup>(4)</sup>, Jean Véronis n'hésite pas à écrire que *"cette façon de procéder conduit à une absurdité"*.

Comme il le souligne justement, la probabilité d'identifier des "documents réels" contenant deux mots choisis au hasard dans un dictionnaire est infiniment faible ; les résultats obtenus lors de recherches de ce type seront donc constitués en majorité de spams et de listes de mots. Il suffit de tester les requêtes utilisées par les chercheurs pour le vérifier.

Par conséquent, de tels résultats ne peuvent être utilisés pour une quelconque estimation de la taille des index des moteurs.

Tout au plus peuvent-ils rendre compte de l'efficacité des capacités de filtrage des deux moteurs qui, dans ces tests, donnent l'avantage à Yahoo!...

Au-delà de la polémique sur la taille des index des moteurs, cet épisode est révélateur de plusieurs phénomènes :

■ **plus que jamais, la vérification des informations trouvées sur le Net doit être le premier réflexe des internautes.**

Alors que les conclusions de l'étude contredisaient l'annonce faite par un acteur reconnu (Yahoo!), le fait que cette dernière soit chapeautée par une université et un organisme réputé sérieux a suffi pour que les résultats soient immédiatement repris par des milliers de sources ; parmi celles-ci, seul un pourcentage minime s'est interrogé sur la méthodologie, la cohérence des résultats, etc.

Cela étant, on peut se demander si ce pourcentage n'aurait pas été plus élevé si les conclusions de l'étude avaient été contraires et avaient remis en cause la "supériorité" de Google...

■ **le pouvoir des blogs et leur rôle dans la circulation de l'information, ne peut plus être ignoré.**

Ce nouveau média, qui permet à tout un chacun de publier instantanément et extrêmement facilement une information sur le Web, mais aussi de commenter et de citer les billets écrits par d'autres, joue un rôle croissant dans la diffusion de l'information.

Les résultats de l'étude, repris par des milliers de blogs, ont ainsi été le centre de moult discussions sur la Toile. Et les remises en cause de sa fiabilité – émises notamment sur les blogs *Technologies du Langage* de Jean Véronis et *Infthought* de Seth Finkelstein<sup>(5)</sup> et reprises par de nombreuses sources – ont eu des conséquences presque immédiates.

A la surprise générale, le NCSA, soucieux de son image, s'est en effet purement et simplement désengagé de l'étude !

Quelques jours après le début de la polémique, l'étude incriminée – qui précisait clairement qu'elle était réalisée par "Matthew Cheney and Mike Perry, two researchers working for the National Center for Supercomputing Applications (NCSA) under the supervision of Associate Director of Humanities and Social Sciences Dr. Orville Vernon Burton", fut en effet mise à jour.

La nouvelle version indiqua en prémisses "The following study was completed by two of Professor Vernon Burton's students at the University of Illinois. Though one of the students previously worked with Professor Burton at the National Center for Supercomputing Applications (NCSA), the study was done outside the scope of any NCSA core projects. When first published online, staff at the NCSA noted several issues with the study, and some revisions have been made to the study to reflect several of these concerns.

Please note again that this study is not an NCSA publication and was not conducted as part of any NCSA project or under the supervision of NCSA."

On ne peut être plus clair !

Cela étant, devant une telle mise au point, on trouve étonnant et regrettable que la première version ait pu être éditée en arborant fièrement le logo du NCSA (supprimé dans la mise à jour...).

Mais le désaveu ne s'est pas arrêté là. Après le NCSA, c'est le professeur Vernon Burton qui s'est désolidarisé de l'étude...

La dernière version en date précise ainsi en introduction "The following study was completed by two of Professor Vernon Burton's former students at the University of Illinois. Although he agreed to host the report on his webpage in the interests of encouraging the debate on the relative strengths of the different search engines, neither Professor Burton nor NCSA had any direct involvement in the study.

*This version is a followup study to the original study that was done to address some legitimate concerns about the inclusion of "wordlists" and "dictionaries" in the study results. The followup study again sampled ~10,000 search queries of Google and Yahoo (excluding dictionaries and wordlists) and found similar results to the original study."*

Bref, les critiques ont été entendues, mais n'ont pas forcément été comprises ; la nouvelle méthode n'est guère plus satisfaisante, puisque les requêtes-tests sont constituées de deux mots tirés au hasard dans le dictionnaire et d'un troisième mot que cette fois-ci l'on exclut ; cela donne, par exemple :

- *blowhole coppersmith -births,*
- *principles preemphasizer -Napoleonizes,*
- ou encore *giraffe Beethoven -jiggered...*

Les résultats de l'étude – qui donnent toujours la supériorité à Google, mais de façon moins frappante – sont donc aussi faussés, mais ce ne sont plus désormais que les travaux de deux étudiants...

■ **il est impossible aujourd'hui de déterminer précisément la taille réelle des index des moteurs** et ce pour plusieurs raisons.

La première est que les moteurs de recherche limitent en général l'affichage de leurs réponses aux 1 000 premiers résultats (pour Yahoo! et Google), voire aux 250 premiers pour MSN.

Lorsque l'on compare les résultats de deux outils, on ne peut donc que se fier aux chiffres – ou plutôt aux estimations – indiqué(e)s par les moteurs. Or, des incohérences dans le nombre de résultats affichés ont à plusieurs reprises été constatées, que ce soit sur Google (voir Netsources n°54), sur Yahoo!<sup>(6)</sup> ou sur MSN<sup>(7)</sup>.

D'ailleurs, il n'est pas rare que les moteurs eux-mêmes se "contredisent" au fur et à mesure de l'affichage des réponses.

Ainsi, pour une recherche sur le mot “diaphragmatocèle”, on obtient<sup>(8)</sup> :

■ **sur Google** : 1 510 résultats.

Mais en fait, après la 132ème réponse, Google affiche “*Pour limiter les résultats aux pages les plus pertinentes (total : 132), Google a ignoré certaines pages à contenu similaire. Si vous le souhaitez, vous pouvez relancer la recherche en incluant les pages ignorées.*” En cliquant sur le lien proposé, on peut afficher cette fois-ci 599 résultats (soit près de 40% des 1 510 promis), comprenant un grand nombre de doublons.

■ **sur Yahoo!** : 1 170 résultats.

Là encore, après l’affichage des 58 premiers, Yahoo! indique “*Afin de ne vous montrer que les résultats les plus pertinents, nous avons omis certains résultats très similaires à ceux déjà affichés.*”

Pour voir l’ensemble des résultats, vous pouvez relancer la recherche en y incluant les résultats occultés. Et l’on peut alors afficher 397 réponses, soit 34 % de ce qui était promis !

Ces différences importantes entre le nombre de résultats “estimé” par le moteur et celui réellement affiché peuvent avoir plusieurs causes. Une hypothèse probable est celle d’un filtrage des résultats à chaque requête, pour éviter les pages indésirables et notamment le spam.

Les moteurs doivent en effet faire face à de multiples tentatives de “spamindexing” de la part d’éditeurs peu scrupuleux, qui cherchent à améliorer le classement de leurs pages ; pour ce faire, ces derniers n’hésitent pas à employer des méthodes contraires à la “nétiquette”, comme l’insertion dans leurs balises de listes de mots sans rapport avec le contenu...

Pour lutter contre cette technique – qui nuit à la pertinence des résultats –, les moteurs utilisent des algorithmes de détection et “blacklistent” de leurs index les pages identifiées comme “spammeuses”.

D’après Jean Véronis<sup>(9)</sup>, ce mécanisme de filtrage, effectué au fur et à mesure de l’affichage, explique les “disparitions” dans les résultats...

Mais d’autres raisons empêchent de connaître précisément la taille des index des moteurs. Ainsi, il faut savoir que ces derniers n’indexent pas forcément les documents importants dans leur intégralité – notamment les documents PDF, qui peuvent comporter de nombreuses pages (thèses...). Google par exemple s’est longtemps restreint aux premiers 101 Ko d’un document, avant de reculer notablement ses limites. Yahoo! quant à lui n’indexe quelquefois qu’une partie minime des documents PDF, Word, etc.

D’autre part, on estime généralement que l’index de Google est composé d’une portion significative (environ 40 %) de pages “repérées” mais non encore indexées ; on identifie ces dernières dans les résultats au fait que contrairement aux autres, seul le titre de la page est indiqué. Il n’y a ni extrait pertinent, ni URL. Par conséquent, le nombre de pages identifié dans l’index “réel” serait quelquefois “gonflé” pour tenir compte des pages non (encore) indexées (voir Netsources n°54).

Bref, ces divers éléments expliquent qu’il est aujourd’hui difficile de connaître la taille réelle des index des moteurs et qu’en tout état de cause, on ne peut se fier aux nombres de résultats estimés par chacun<sup>(10)</sup>.

Mais au fond, qu’importe ces batailles autour de la taille du Web – et des moteurs. Il est rare en effet de visualiser plus de quelques dizaines de résultats...

Les premiers critères de qualité d’un moteur restent la pertinence de son classement et la fraîcheur de son index.

On aimerait bien que sur ces points là justement, Google, Yahoo! & Co. soient un peu plus diserts...

(1) [www.ysearchblog.com](http://www.ysearchblog.com) - A look inside the world of search from the people of Yahoo!

(2) “A Comparison of the Size of the Yahoo! and Google Indices”  
<http://vburton.ncsa.uiuc.edu/indexsize.html>

(3) <http://battellemedia.com/archives/001790.php>

(4) <http://aixtal.blogspot.com>

(5) <http://sethf.com/infothought/blog/>

(6) <http://aixtal.blogspot.com/2005/03/web-yahoo-double-ses-comptes.html>

(7) <http://aixtal.blogspot.com/2005/02/web-msn-triche-t-il-aussi.html>

(8) tests effectués le 28 septembre 2005

(9) <http://aixtal.blogspot.com/2005/08/yahoo-pages-manquantes-3.html>

(10) Le dernier épisode de la “guerre des moteurs” vient de s’achever, au moment où nous mettons sous presse. Le 27 septembre en effet, jour de son anniversaire (7 ans), Google a supprimé l’indication du nombre de pages recensées, qui figurait dans le bas de son écran. Il indique sur son blog que :

- la taille de son nouvel index est 1 000 fois plus importante que lors de son lancement il y a sept ans – Google a été lancé avec un index de 24 millions de pages, ce qui porterait le nouvel index à 24 milliards de pages ; cette estimation – qui sous-entend un triplement de son index – est cohérente avec divers tests effectués par des experts ;
- la taille de son index est trois fois plus importante que celle de n’importe quel autre outil de recherche – cette affirmation revient à nier la précédente communication de Yahoo!... La hache de guerre ne sera peut-être pas enterrée très longtemps...

(voir <http://googleblog.blogspot.com> et

