Analyse quantitative des résultats de Google

Google propose, sur sa page de résultats, le nombre de résultats trouvés pour une requête donnée. Mais ce nombre est-il fiable ? Pour le savoir, nous avons saisi, chaque jour du mois de novembre 2002, les 40 mêmes requêtes sur le moteur www.google.fr et noté, à chaque fois, le nombre de résultats affiché. Un fichier Excel regroupant tous les chiffres du mois peut être téléchargé ici :

http://abonnes.abondance.com/archives/2002-12/google.xls

A la vue de ces résultats, que pouvons-nous conclure ? Tout d'abord, que les mots clés ou expressions saisis se subdivisent en trois catégories :

- Ceux pour lesquels le nombre de résultats fournis par Google a été **quasiment stable** sur les 30 jours (moins de 10% d'écart) :

Mot clé	Nombre de résultats minimum affiché par Google	Nombre de résultats maximum affiché par Google	Nombre de résultats moyen sur les 30 jours	Ecart
fabricant amortisseur	713	777	767	8,34%
nestlé environnement	5 550	6 070	5 924	8,78%
exemples business plan	18 200	19 800	19 353	8,26%
marégraphe	514	566	548	9,49%

- Ceux pour lesquels le nombre de résultats fournis par Google **a fluctué de façon assez forte** sur les 30 jours (entre 10 et 50% d'écart) :

Mot clé	Nombre de résultats minimum affiché par Google	Nombre de résultats maximum affiché par Google	Nombre de résultats moyen sur les 30 jours	Ecart
résultats tournoi ATP	5 340	8 550	7 889	40,69%
musées paris	90 300	106 000	101 307	15,50%
virus sircam	80 200	118 000	108 533	34,82%
location studio paris	105 000	141 000	132 900	27,09%
convertisseur euro	42 000	58 500	52 377	31,50%
cartes voeux	70 000	84 800	81 100	18,25%
fond d'écran	80 800	111 000	103 710	29,12%
référencement	150 000	204 000	187 967	28,73%
cinéma	2 020 000	2 850 000	2 581 667	32,15%
klevener	642	802	718	22,28%
heiligenstein	6 290	7 640	7 448	18,13%
arènes arles	2 900	3 330	3 259	13,19%
photos paris	1 110 000	1 620 000	1 511 000	33,75%
vin colmar	4 490	5 190	5 050	13,86%
astérix officiel	5 950	7 460	6 799	22,21%
téléphone portable santé	18 200	22 400	21 657	19,39%
recette couscous	10 900	12 500	12 137	13,18%
nimbo-stratus	251	279	267	10,49%
photos paris	1 110 000	1 620 000	1 511 000	33,75%
sonneries GSM	32 000	54 200	49 963	44,43%

- Ceux pour lesquels le nombre de résultats fournis par Google **a très nettement fluctué** sur les 30 jours (plus de 50% d'écart) :

Mot clé	Nombre de résultats minimum affiché par Google	Nombre de résultats maximum affiché par Google	Nombre de résultats moyen sur les 30 jours	Ecart
eye tracking	153 000	282 000	221 200	58,32%
vidéos humour	267 000	496 000	441 500	51,87%
mp3	3 880 000	29 400 000	22 651 000	112,67%
jeux vidéo	748 000	1 510 000	1 330 100	57,29%
warrants	833 000	1 760 000	1 452 733	63,81%
hotmail	4 800 000	17 200 000	14 027 333	88,40%
sexe	136 000	32 300 000	17 498 433	183,81%
webcam	1 130 000	13 600 000	7 355 667	169,53%
immobilier	63 800	1 340 000	933 880	136,66%
euro	3 730 000	21 900 000	17 140 000	106,01%
lofstory	15 500	83 600	76 910	88,55%
bourse	120 000	1 740 000	1 405 400	115,27%
googlefight	451	15 800	13 784	111,35
france	4 950 000	42 200 000	31 227 000	119,29%
star academy	232 000	409 000	293 667	60,27%
popstars	80 900	313 000	231 827	100,12%
réduve	110	217	123	86,99%

L'écart a été calculé ainsi : ((nombre max-nombre min)*100)/(valeur moyenne).

L'analyse de ces tableaux permettent de faire ressortir des faits marquants :

- Le nombre de résultats indiqué par Google sur sa page de résultats est donné à titre indicatif. Cela, on s'en doutait car autrement, ce serait un nombre beaucoup plus précis qui serait indiqué et non obligatoirement un multiple de 10, comme c'est le cas actuellement. De plus, la phrase affichée est on ne peut plus claire : "1 - 10 résultats, sur un total d'environ 185,000". Notez le "environ"... Ceci dit, pour le 3ème tableau, les écarts sont très significatifs et d'autres explications doivent être trouvées.

- Certains mots clés sont l'objet de ruptures fortes d'un jour à l'autre au niveau du nombre de résultats affichés par Google. La plupart de ces ruptures ont lieu le même jour, démontrant certainement soit un problème du moteur avec son index, soit une charge forte du serveur, délaissant le temps CPU chargé de calculer ce nombre pour d'autres tâches. Ainsi, les jours suivants connaissent des résultats très perturbés pour les mots clés du 3ème tableau (plus de 50% d'écart) : du 3 au 6 novembre (coïncidant avec le fin de la "Google Dance", période de la mise à jour des index du moteur, ayant commencé le 1er novembre), 11, 13, 19 novembre (très fortes perturbations ce jour-là). On observe également quelques fluctuations passagères entre le 7 et le 10, mais des résultats plutôt stables entre le 20 et le 31 du mois.

- A quoi sont donc dues les fluctuations des nombres affichés ? Trois possibilités à notre avis :

* Les index ne sont pas figés entre deux "Google Dances" et leur contenu est modifié quotidiennement, d'où de légères différences d'un jour à l'autre. Des liens issus de sites crawlés quotidiennement (sites d'actualité pour la plupart) peuvent ainsi modifier les résultats d'un jour à l'autre.
* Le temps CPU pris par la routine qui calcule le nombre de résultats est dépendant de la charge du moteur. Si il y a de très nombreuses requêtes demandées et que Google "chauffe", les résultats fournis peuvent être beaucoup plus approximatifs qu'en période "calme". On sait qu'AltaVista, par exemple, était très sensible à ce phénomène il y a quelques années de cela.
* Des changements d'index certains jours pour cause de maintenance.

Mais d'autres raisons sont peut-être envisageables...

- Les week-ends ne semblent pas avoir d'incidence : les samedi et dimanche, on trouve autant de fluctuations que lors des jours de semaine. En tout cas, on n'en trouve ni plus ni moins...

- Les écarts observés sur le mot clé "googlefight" s'expliquent : le site est très récent et lors du changement d'index, après la "Google Dance", début novembre, le nombre de pages contenant le mot (donc parlant du site) a augmenté de façon drastique. Logique puisque le site a été lancé il y a quelques semaines de cela, début octobre...

- Les nombres affichés ne sont pas calculés "au petit bonheur la chance" (heureusement :-)). Durant plusieurs jours, ils peuvent être strictement égaux pour un même mot clé, prouvant qu'il y a ainsi une certaine continuité dans les résultats et que le nombre affiché n'est pas aussi approximatif que cela. Nous avons également fait des tests sur la même requête à un quart d'heure d'intervalle sur deux heures consécutives par exemple, et les résultats étaient souvent identiques à chaque fois. Sur une période courte de quelques heures, il y a donc très souvent une bonne homogénéité du nombre de résultats affiché.

En tout cas, la conclusion globale de cette mini-étude est qu'il est difficile de faire confiance à 100% aux chiffres indiqués sur la page de résultats de Google. D'ailleurs, le moteur de recherche n'a jamais revendiqué une quelconque justesse de ce nombre, toujours considéré comme donné à titre indicatif. On en a ici la preuve...