

Les outils basés sur les données "piratées" d'AOL

[Retour au sommaire de la lettre](#)

AOL a dévoilé par erreur, il y a quelques semaines de cela, un fichier de données portant sur 36 millions de requêtes effectuées par plus de 500 000 internautes sur son moteur américain, dont la technologie est fournie par Google. Du coup, de nombreux sites ont analysé ces données et proposé des outils d'interrogation de cette énorme base d'informations. Nous avons tenté de les compiler dans cet article...

Vous en avez peut-être entendu parler : AOL a, aux Etats-Unis, publié un certain nombre de données par erreur, certaines d'entre elles permettant de retrouver des internautes utilisateurs de son moteur de recherche au travers de leurs habitudes de recherche (<http://actu.abondance.com/2006-32/aol.php>)...

AOL s'est excusé pour cette "gaffe" mais le mal était fait (les responsables ont été, malheureusement pour eux, virés...) et le fichier téléchargé plusieurs centaines de fois avant d'être retiré. L'étude était disponible ici (mais l'url ne semble plus répondre...): <http://research.aol.com/pmwiki/pmwiki.php?n=Research.500kUserQueriesSampledOver3Months>

Cette étude portait sur l'analyse de 36 millions de requêtes effectuées sur son moteur de recherche par plus de 500 000 internautes sur une période de trois mois. Et le point le plus intéressant était que le fichier contenant les requêtes était disponible en ligne. Il a vite disparu de sa source "officielle" mais ce fichier de mots clés a cependant été sauvegardé sur plusieurs sites miroirs dont vous trouverez les adresses (entre autres) ici : <http://www.gregsadetsky.com/aol-data/>

Il n'en fallait pas plus pour donner quelques idées à certains éditeurs de sites et développeurs qui ont eu l'idée de proposer des outils ou études intéressants sur la façon dont les internautes utilisent le moteur de recherche d'AOL. En effet, un travail sur 36 millions de requêtes est plus que significatif pour mieux connaître nos habitudes de recherches, AOL utilisant - qui plus est - la technologie de recherche de Google... Cette étude est restreinte malheureusement au site américain du portail, puisque AOL est "powered by" Exalead en France. Mais cette base de données et son analyse n'en restent pas moins intéressantes.

Voici donc un florilège des articles et applicatifs qui ont été bâtis sur la base de ce fichier :

Les analyses

- Répartition des clics sur les 10 premiers résultats Google

<http://www.webrankinfo.com/actualites/200608-repartition-des-clics.htm>

WebRankInfo analyse la façon dont les internautes cliquent sur les liens proposés dans la première page de résultats du moteur (exemple : le résultat en 1ère position totalise 42% des clics !).

- Nombre moyen de mots par requête

<http://www.webrankinfo.com/actualites/200608-mots-par-requete.htm>

27,5% des requêtes sont effectuées sur un seul terme et 29% sur deux mots, etc.

- Nombre de clics dans les pages de résultats

<http://www.webrankinfo.com/actualites/200608-nombre-de-clics.htm>

Exploration du nombre de clics effectués par un même utilisateur sur une page de résultats donnée (89% des utilisateurs ne consultent qu'un seul des résultats affichés...).

- Domaines et sous-domaines

http://www.webrankinfo.com/forums/viewtopic_57139.htm

Une discussion sur les domaines et sous-domaines des sites trouvés dans les résultats, sur le forum de WRI...

- Decent tools

<http://www.decenttools.com/>

Ce site en anglais propose de nombreuses statistiques basées sur le fichier d'AOL : les 500 mots clés les plus demandés (<http://www.decenttools.com/aol/top-keywords/>), les 500 domaines les plus consultés (<http://www.decenttools.com/aol/top-domains/>) et bien d'autres, proposés en page d'accueil du site.

- **SEO Sleuth** (<http://www.seosleuth.com/site/>) propose également un classement des sites web le plus souvent consultés ainsi qu'un "top keywords" (<http://www.seosleuth.com/kp/>).

- **Don't Delete** (<http://www.dontdelete.com/>) propose un "top 500" des mots clés les plus demandés avec leur nombre d'occurrences.

N'hésitez pas également à consulter les sites proposés ci-dessous. La plupart d'entre eux proposent des analyses, tableaux et graphiques basées sur les données d'AOL...

Les outils en ligne

Le fichier en lui-même est également interrogeable en ligne grâce à un ou plusieurs formulaires de recherche. Voici les principaux sites qui ont mis en place un tel outil :

- AOL Search Database

<http://www.aolsearchdatabase.com/>

Le fichier est ici interrogeable par numéro d'utilisateur, mot clé ou site web. Il semblait malheureusement ne plus fonctionner au moment où ces lignes étaient écrites... L'outil semblait de plus ne pas utiliser la base de données en entier, mais uniquement un "extract". Prudence donc...

- SEO Sleuth

<http://www.seosleuth.com/site/>,

Recherche au sein du fichier par mot clé ou nom de domaine. Indique notamment avec quels mots clés un site a été trouvé.

- Don't Delete

<http://www.dontdelete.com/>

Recherche par mots clés, numéro d'utilisateur, url, etc. Semble en revanche très lent...

- Decent Tools

<http://www.decenttools.com/>

Idem : plusieurs types de recherches sont possibles.

- Data Blunder

<http://www.datablunder.com/>

Recherche par mots clés ou url.

- Frogspy

<http://www.frogspy.com/>

Recherche par mots clés.

- AOL Search Logs

<http://data.aolsearchlogs.com/search/index.cgi>

Recherche par mot clé ou url.

- Splunk'd

<http://apophis.rhombic.net/>

Outil qui semble puissant mais qui dispose d'une interface complexe...

- AOLStalker

<http://www.aolstalker.com/>

Belle interface. Recherche par mot clé ou numéro d'utilisateur...

- AOL Psycho

<http://www.aolpsycho.com/>

Analyse de la personnalité des internautes au travers de leurs recherches, en tapant leur numéro d'utilisateur. Glups...

Logiquement, avec tous ces outils, vous devriez en savoir plus sur la façon dont les internautes américains recherchent l'information. Soit en lisant les analyses déjà effectuées, soit en utilisant les outils d'interrogation mis à votre disposition. Soit, encore mieux, en téléchargeant la base de données et en créant les vôtres... Dans ce cas, n'oubliez pas d'en mettre une version en ligne, utilisable par les autres internautes, car nous n'avons pas trouvé ce type d'outils disponible en langue française... Bon courage !