Comment détecter le passage des robots via Google Analytics

Par Daniel Roch

Domaine :	Recherche	Référencement
Niveau :	Pour tous	Avancé

Par défaut, les outils de WebAnalytics permettent d'analyser en temps réel et sur le long terme le trafic des visiteurs sur un site web : on y analyse par exemple le canal d'acquisition, leur comportement ou encore les URL visitées. En référencement naturel, on peut également s'intéresser au trafic des robots (spiders, crawlers) de Google et des autres moteurs de recherche pour mieux comprendre leur indexation et trouver des leviers d'amélioration SEO. Nous allons voir ainsi comment détourner les scripts des outils de Webanalytics pour analyser les visites de Googlebot et de ses « collègues », sans pour autant avoir besoin d'analyser les logs de son serveur.

L'intérêt de suivre les bots des moteurs de recherche

La première question à se poser est de savoir quels sont les avantages d'une telle analyse. Lorsqu'on liste les étapes pour positionner une page dans Google, elles suivent un ordre logique :

- Le crawl : Google découvre une de vos pages et l'analyse ;
- L'indexation : Google ajoute cette URL à son index ;
- Google affiche ensuite cette page lors d'une requête de l'internaute, en utilisant différents critères de pondération pour son positionnement : pertinence du contenu, popularité, qualité du code HTML…

Le fait d'analyser le passage des bots va ainsi permettre de comprendre différents éléments pour chaque page du site :

- Google connaît-il et crawle-t-il cette page ?
- Revient-il souvent sur cette dernière ?
- Quelles sont les pages les plus crawlés (donc les pages les plus populaires ou pertinentes à ses yeux) ?
- Quelles sont les pages peu crawlées (donc celles peu populaires ou peu pertinentes) ?
- Les bots crawlent-ils des pages inutiles ?
- Etc.

Pour cela, on peut utiliser différentes méthodes ou outils :

- L'analyse des logs de son serveur ;
- L'utilisation d'un outil dédié :
    - Watussi : http://box.watussi.fr/ ;
    - Botify : https://www.botify.com/ ;
    - Screaming Frog : http://www.screamingfrog.co.uk/seo-spider/
    - Etc…
- Un développement sur-mesure d'un outil de suivi ;
- L'utilisation d'un outil de webanalytics.

C'est vers cette dernière solution que nous allons nous tourner.

Les bots sont théoriquement exclus

Par défaut, tous les outils de webanalytics vont exclure les bots. Leur objectif est en effet d'analyser uniquement le comportement des visiteurs, et ils excluent ainsi toute visite par un spider (que ce soit ceux de Google, de Bing ou de tout autre moteur de recherche ou outil de crawl).

La première problématique que l'on peut rencontrer actuellement est d'avoir ce que l'on appelle du « referral spam », c'est-à-dire des sites indiqués comme étant la source de trafic de certains visiteurs, alors qu'il s'agit en réalité de bots automatisés qui vont fausser les statistiques affichées. Par exemple, vous êtes dans ce cas de figure si vous avez des visites depuis des sites comme Semalt, Econsultancy ou encore Floating-share-buttons (pour les plus connnus, ou disons les plus spammeurs).

Pour vous en débarrasser, suivez les tutoriels de ces articles :

- https://moz.com/blog/how-to-stop-spam-bots-from-ruining-your-analytics-referral-data (EN) ;
- https://econsultancy.com/blog/66848-how-to-stop-referral-spam-from-screwing-up-your-google-analytics-stats/ (EN) ;
- http://www.agence404.com/supprimer-le-spam-referents-dans-analytics/ ;
- http://www.audiaweb.com/blog/comment-supprimer-le-spam-des-faux-sites-referents-de-vos-rapports-google-analytics/.

Suivre le passage des bots

La méthode pour suivre les bots est relativement simple, et réside dans l'utilisation d'un script à installer en complément du tracking habituel de Google Analytics : **« Universal Analytics for Search Bots »**. Ce script Open Source d'Adrian Vender permettra le suivi de tous les bots directement dans Google Analytics. Il est disponible à cette adresse : http://www.adrianvender.com/universal-analytics-for-search-bots/

Les étapes d'installation sont les suivantes :

- Créez une nouvelle propriété appelée « Bots » dans Analytics : cela permettra d'avoir d'un côté le suivi de vos visiteurs dans vos propriétés actuelles, et de l'autre le suivi des bots dans la nouvelle propriété que vous venez de créer ;
- Téléchargez, en suivant le lien précédent, le script d'Adrian Vender, et installez-le sur votre site en modifiant 2 éléments :
    - L'ID de suivi de Google Analytics en utilisant celui de la nouvelle propriété ;
    - Le chemin vers le script de suivi.

Pour ces différents points, le lien précédent explique en détail la procédure d'installation.

Dès lors, le script fonctionnera et pourra suivre tout type de bot : Google, Yahoo, Yandex, Bing, Majestic, Xenu…

Adapter le script

Le script est composé de 3 fichiers :

- sample.php pour paramétrer l'outil. C'est ici qu'il faut modifier les deux éléments précédemment cités ;
- ua-searchbots.php, le script qui fera fonctionner le suivi (à ne pas modifier) ;
- botconfig.php, qui liste tous les bots à suivre.

C'est dans ce dernier fichier que vous pourrez ajouter, adapter ou supprimer les bots à suivre dans Google Analytics, en prenant pour exemple les lignes suivantes qui définissent chacune un bot différent de Google :

```
'Googl(elebot)(-Image)/' => 'Google Image',
'^gsa-crawler' => 'Google',
'Googl(elebot)(-Sitemaps)/([0-9.]{1,10})?' => 'Google-Sitemaps',
```

Un cas concret, avec Frédéric Gaye

Pour bien comprendre à quoi cela sert, et surtout comment le mettre en place, Frédéric Gaye partage avec nous dans cet article son expérience en ce domaine, puisqu'il installe très régulièrement ce type de suivi à l'aide de Google Analytics, et qu'il l'utilise au quotidien pour le référencement naturel.

Qui est Frédéric Gaye ?

Frédéric Gaye est consultant en référencement naturel depuis plus de 10 ans. Certifié Google Analytics Individual Qualification, il a notamment mis en place, dans le cadre de ses prestations de référencement, le suivi du passage des moteurs de recherche via Google Analytics, pour de nombreux clients.

Quelle est la problématique rencontrée sur un site donné ?

La fréquence de passage des moteurs est un critère essentiel à la prise en compte et au classement des contenus. Nous savons très bien qu'une page sur laquelle Google ne passe pas régulièrement (et de préférence au moins tous les 7 jours) aura tendance a moins bien se positionner qu'une page équivalente d'un concurrent dont la fréquence de passage des moteurs est plus élevée.

Sur des requêtes concurrentielles (voire très concurrentielles), le moindre détail compte et connaître le jour et l'heure du passage de Google *via* son robot d'indexation Googlebot afin d'adapter le site devient nécessaire. Au-delà des outils de crawl, l'utilisation d'un outil de mesure d'audience gratuit comme Google Analytics permet, avec un peu de configuration, de répondre à ce besoin.

Dans ce contexte, nous souhaitions connaître, pour un de nos clients, si ce critère devait être optimisé afin de mettre d'éventuelles actions correctives en place.

Quelle est la solution mise en place ?

Nous avions précédemment mis en place, pour ce client, un plan de marquage avec Google Analytics. L'utilisation de cet outil rapide à mettre en place permet d'accéder à des données basiques telles que la fréquence de crawl des pages par Google ainsi que le jour et l'heure de chaque passage du bot.

Nous nous sommes basés sur le script original UA for Search Bots d'Adrian Vender. Ce script PHP permet d'enregistrer simplement les informations dont nous avions besoin.

A noter que ce script enregistre le passage de Google mais également de n'importe quel robot (y compris d'ailleurs les robots des outils de crawl comme Xenu ou autres, ce qui peut être intéressant aussi si vous souhaitez savoir si votre site est crawlé par un tiers).

Quels sont les résultats obtenus ?

A partir du script original, nous avons configuré un rapport personnalisé (https://support.google.com/analytics/answer/1151300?hl=fr) ce qui nous a permis de voir instantanément quelles étaient les pages sur lesquelles Google passaient le plus souvent, et de découvrir au passage que certains types de contenus étaient mieux pris en compte que d'autres. Cette information nous a ainsi incité à demander à nos rédacteurs de rédiger plus spécifiquement certains type d'articles.

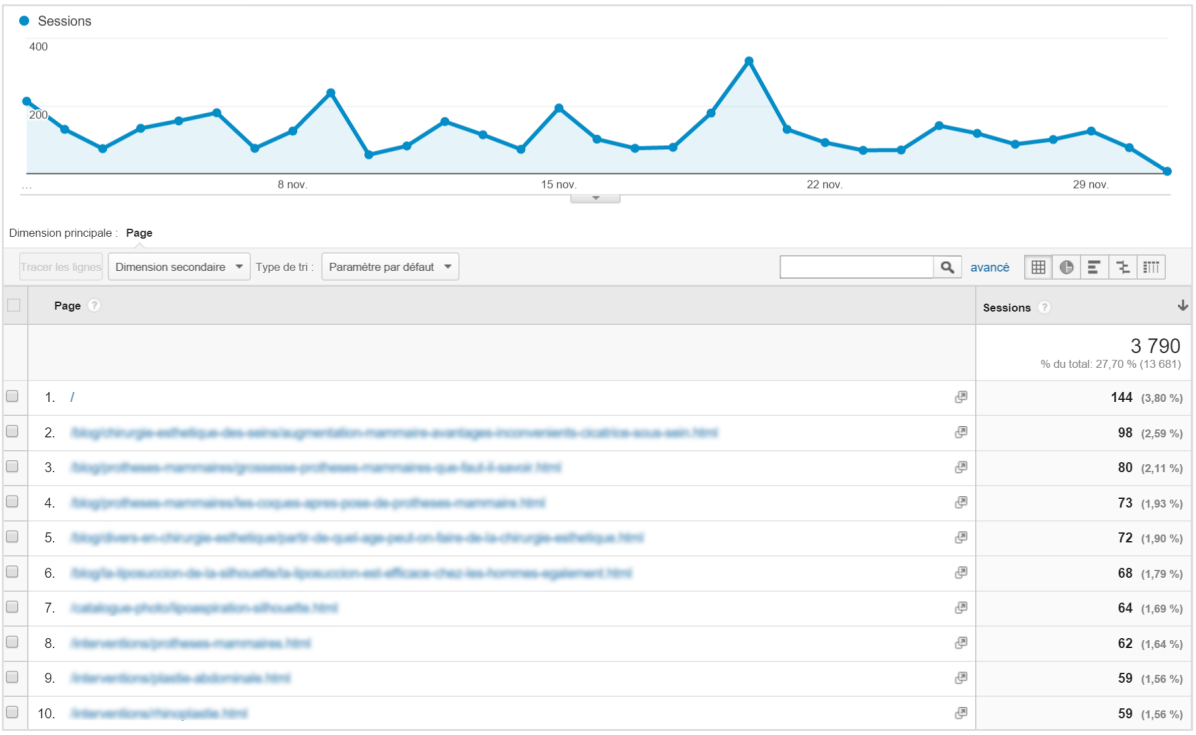

Fig 1. Tableau de bord personnalisé permettant le suivi du passage des moteurs via Google Analytics

Il est également intéressant d'associer ce type d'information avec les données de positionnement dans Google Search Console afin de voir si une fréquence de passage plus élevée entraîne plus d'impressions dans les pages de résultats. Certes, la fréquence de passage n'est qu'un critère parmi les 200 pris en compte par Google mais il est intéressant de noter une corrélation entre fréquence de passage et impressions dans les pages de résultats de recherche.

Le rapport personnalisé nous a également permis d'identifier, à partir d'une page donnée, le jour, l'heure et la minute de passage du moteur. La courbe des sessions de ce rapport indique également, sur une période donnée, par exemple le mois dernier, si le moteur passe régulièrement sur le site tant en fréquence de passage sur le mois que par jour. Les pics de fréquentation étaient alors analysés afin de comprendre ce qui favorisait une hausse de la fréquence de passage (pour reproduire dans la mesure du possible ce schéma).

Quelle est la méthode à utiliser pour bien utiliser les rapports et données ?

La première chose à faire, une fois le script installé sur le serveur, est de configurer Google Analytics *via* la mise en place d'un rapport personnalisé de type explorateur. Le point important à comprendre est qu'un moteur est assimilé à une source et un support. Le support est par défaut « bot ». Il faut donc bien veiller lors de la création du rapport, si l'on veut dans ce dernier isoler le comportement de Google, à ajouter un

filtre de type *inclure* sur la statistique « Source/support » avec comme paramètre « Google / bot » comme indiqué dans la figure 2.

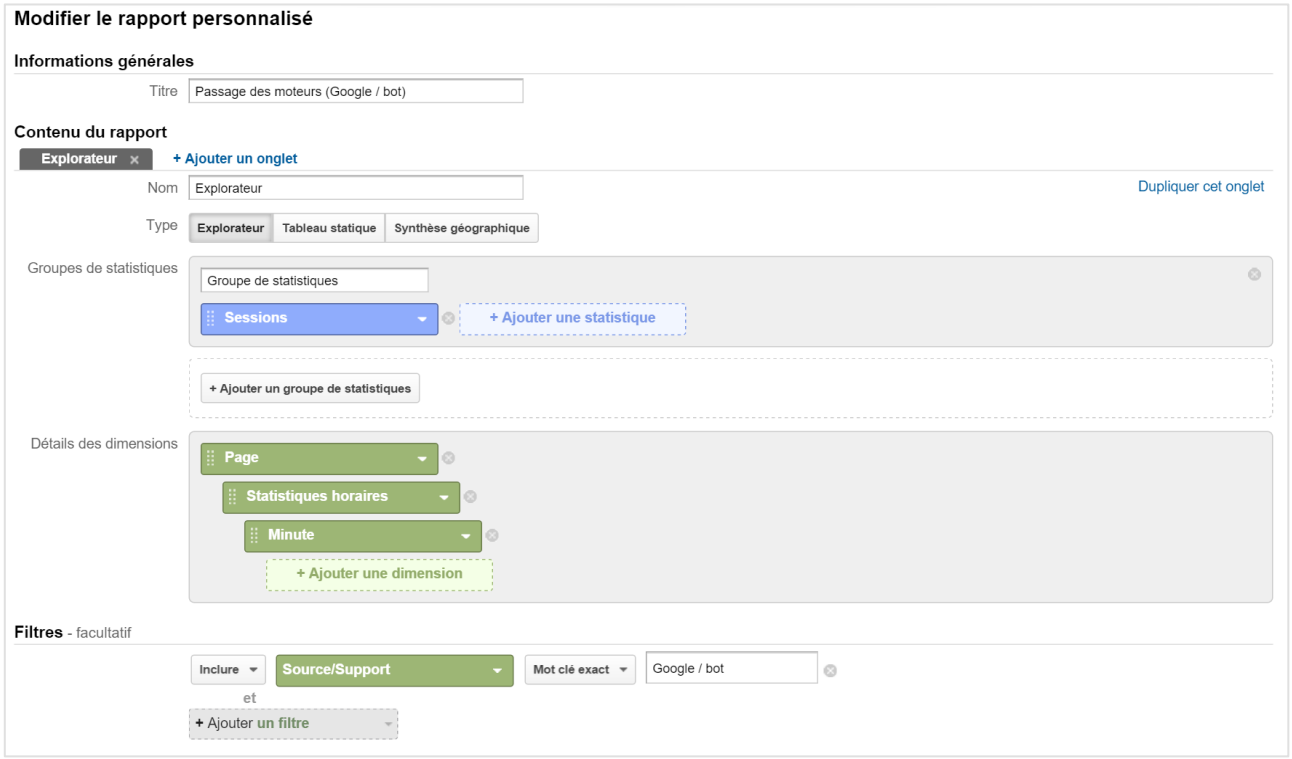

Fig 2. Règles de paramétrage du tableau de bord personnalisé permettant le suivi des moteurs

Dans l'exemple de la figure 3, la configuration en mode explorateur permet au clic sur la page souhaitée d'afficher alors le jour et l'heure du crawl, puis les minutes. Il est également possible d'utiliser le mode « Tableau statistique » afin d'avoir un aperçu sous forme de tableau de données avec du coup la liste des pages par jour de passage les unes sous les autres.

Par date

avancé

Page	Statistiques horaires	Minute	Sessions
1. /	2015110100	04	1 (0,03 %)
2. /	2015110104	39	1 (0,03 %)
3. /	2015110111	27	1 (0,03 %)
4. /	2015110113	21	1 (0,03 %)
5. /	2015110114	35	1 (0,03 %)
6. /	2015110122	41	1 (0,03 %)
7. /	2015110210	54	1 (0,03 %)
8. /	2015110211	14	1 (0,03 %)
9. /	2015110215	08	1 (0,03 %)
10. /	2015110219	14	1 (0,03 %)

Fig 3. Présentation du rapport en vue "Tableau Statistique" permettant d'afficher les jours et heures de passage des robots de moteurs par page.

Le suivi du passage des robots de Google ou des autres moteurs doit, au final, permettre d'optimiser le trafic en provenance du référencement naturel. Et, ici encore, l'utilisation d'un outil comme Google Analytics ou comme Summarix (www.summarix.com) qui analysent les données d'un compte Google Analytics (pour indiquer, dans un langage clair, les leviers les plus performants) permettra de surveiller l'impact de ces modifications et de prendre les bonnes décisions.

On peut ainsi mesurer l'impact d'une modification du site sur le crawl de Google, et le coupler ensuite avec les données de positionnement du Centre Webmaster de Google ou d'un outil dédié de positionnement (Ranks, MyPoseo…).

Merci à Frédéric Gaye (www.lereferencementnaturel.org) pour l'utilisation des visuels et pour ses réponses.

Les problématiques techniques

Attention cependant, le script ne fonctionnera pas dans un cas de figure : les pages mises en cache par des extensions. Dans l'image suivante, on voit justement la forte baisse des statistiques de crawl suite à la mise en place d'une extension de cache (voir fig. 4).

Fig 4. Suite à l'activité d'une extension de cache, on perd une grosse partie des données de crawl.

La raison à ce bug est simple : le script fonctionne avec du code PHP exécuté sur votre site, et qui détecte puis enregistre les données. Le souci est que les extensions de cache vont générer une version HTML statique de vos contenus, et ces contenus sont servis *via* des règles htaccess. A aucun moment, le code PHP ne peut donc être exécuté, et l'on perdra ainsi systématiquement les données de crawl de l'ensemble des bots, à moins de désactiver temporairement l'extension de cache.

Autre problématique : le plan de marquage. Pour bien fonctionner, le script doit être présent sur toutes les pages. Cela implique donc de bien maîtriser son outil et de vérifier la réelle présence du script. Cela implique aussi que certains types de contenus ne pourront pas être analysés, comme par exemple les images, les vidéos, les flux RSS ou encore tous les fichiers textes, Word ou PDF qui ne peuvent exécuter de code PHP.

Conclusion

Suivre le crawl de Google est réellement intéressant pour comprendre la façon dont Google découvre et (ré)analyse chacune des URL d'un site. Cela permet de voir plus

facilement les pages inutiles, les pages populaires et surtout l'impact de vos actions sur la vitesse de crawl des moteurs de recherche.

Le script **« Universal Analytics for Search Bots »** permet facilement et rapidement de mettre en place ce type de suivi. Mais, tout comme avec les autres solutions existantes, celle utilisant Google Analytics connaît quelques lacunes techniques qui ne permettront pas d'avoir un suivi total du crawl des différents bots.

Cependant, les données seront déjà largement suffisantes pour la plupart des référenceurs afin de mesurer d'une part la qualité du crawl de Google, et d'autre part l'impact de leurs optimisations.

Daniel Roch, *Consultant WordPress, Référencement et Webmarketing chez SeoMix (http://www.seomix.fr/).*