Le contenu des User agents des robots : comment le comprendre et l'analyser ?

Par Daniel Roch

Domaine :	**Recherche**	**Référencement**
Niveau :	Pour tous	**Avancé**

Le User Agent permet de mieux comprendre la nature d'une visite sur un site, que ce soit un robot ou un visiteur réel. Comprendre le fonctionnement de cette donnée et savoir l'analyser peut ensuite permettre d'agir au niveau SEO, et plus globalement au niveau de tous les aspects d'un site. Nous allons voir dans cet article à quoi ressemble un User Agent, comment on peut le détecter, l'analyser, et ensuite comment l'utiliser dans une stratégie de référencement naturel.

Qu'est-ce qu'un User Agent ?

Lorsqu'un internaute ou un robot automatisé navigue sur un site, ils transmettent des informations au serveur visité. Parmi celles-ci, se trouve le User Agent. Cette dernière permet d'identifier le logiciel utilisé par le visiteur ou la nature du robot. On peut ainsi connaître différentes informations comme :

- Le nom et la version du navigateur ou du robot ;
- La compatibilité éventuelle avec certaines technologies ;
- Le système d'exploitation de l'internaute.

Pour mieux comprendre à quoi cela ressemble, voici un exemple de User Agent pour un internaute : *Mozilla/5.0 (Windows NT 5.1; rv:31.0) Gecko/20100101 Firefox/31.0.*

Détecter un User Agent

Votre User Agent

La première façon de comprendre et de détecter cela est de tout simplement vérifier quel est son propre User Agent. Visitez donc le site suivant pour connaître votre User Agent actuel : http://www.whoishostingthis.com/tools/user-agent/.

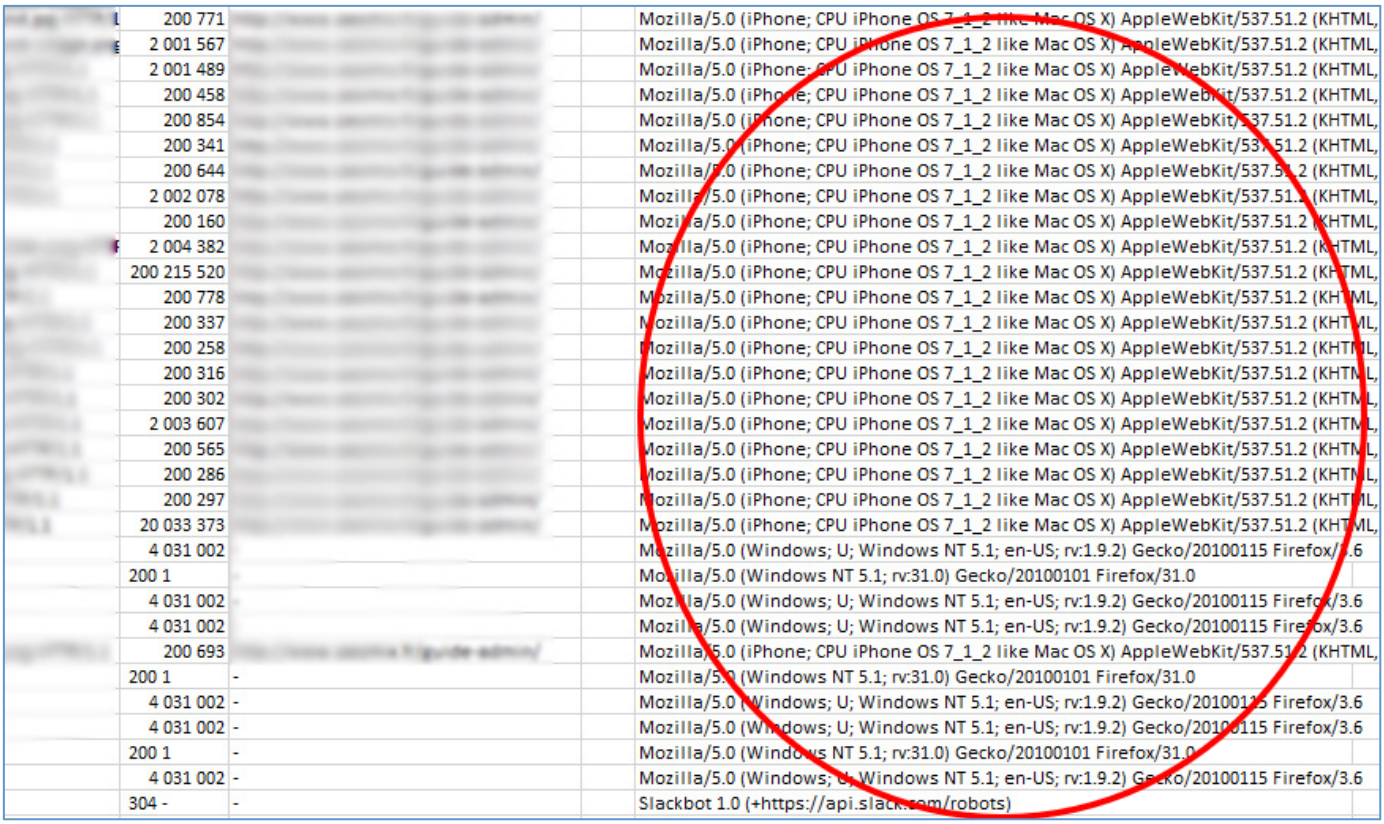
Dans les logs

Le meilleur endroit pour trouver tous les User Agents qui naviguent sur votre serveur est le fichier de logs généré par ce dernier, et qui va lister minute par minute la liste complète des robots et internautes qui consultent votre site. L'analyse de ce type de fichier est chronophage (à moins de disposer d'un outil dédié), mais peut être une réelle source d'informations à la fois pour le SEO, mais aussi pour d'autres aspects de votre site (sécurité, temps de chargement, etc.).

Pour ce type d'analyse, il faut dans un premier temps télécharger les logs sur votre ordinateur. Chaque hébergeur étant différent, vous devrez vous renseigner auprès de lui pour trouver leur emplacement. Une fois cette étape réalisée, il vous suffira d'importer dans Excel le fichier de logs. Dans notre exemple, nous importons les logs de l'hébergeur Infomaniak dans Microsoft Excel. Nous effectuons donc les actions suivantes :

- Téléchargement des logs depuis le serveur ;
- Importation des logs avec le menu « Données > Fichier Texte » d'Excel ;
- Lors de l'import, nous avons choisi le séparateur " pour séparer chaque information dans une colonne spécifique.

Excel affiche alors le détail des logs : adresse IP, heure de connexion, URL concernée sur le serveur, Etc. Ce n'est que dans la dernière colonne que le User Agent est précisé.



[illegible]	200 771	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
[illegible]	2 001 567	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	2 001 489	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 458	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 854	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 341	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 644	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	2 002 078	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 160	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
[illegible]	2 004 382	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 215 520	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 778	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 337	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 258	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 316	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 302	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	2 003 607	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 565	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 286	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 297	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	20 033 373	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	4 031 002			Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
	200 1			Mozilla/5.0 (Windows NT 5.1; rv:31.0) Gecko/20100101 Firefox/31.0
	4 031 002			Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
	4 031 002			Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
	200 693	[illegible]		Mozilla/5.0 (iPhone; CPU iPhone OS 7_1_2 like Mac OS X) AppleWebKit/537.51.2 (KHTML,
	200 1	-		Mozilla/5.0 (Windows NT 5.1; rv:31.0) Gecko/20100101 Firefox/31.0
	4 031 002	-		Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
	4 031 002	-		Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
	200 1	-		Mozilla/5.0 (Windows NT 5.1; rv:31.0) Gecko/20100101 Firefox/31.0
	4 031 002	-		Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.2) Gecko/20100115 Firefox/3.6
	304 -	-		Slackbot 1.0 (+https://api.slack.com/robots)

Fig. 1. Un exemple d'importation de logs où l'on peut consulter les User Agent

Par script

Tout dépendra du langage utilisé par votre site (PHP, Java, ASP, Javascript, etc.), mais sachez que quasiment tous permettent de récupérer très facilement cette information.

Par exemple en PHP, l'utilisation de cette variable *$_SERVER['HTTP_USER_AGENT']* permettra de récupérer tout le User Agent de l'utilisateur (internaute ou robot). Couplé à d'autres fonctions, on peut même récupérer d'autres informations comme par exemple la compatibilité de ce User Agent avec certaines fonctionnalités ou langages. En PHP, on utilisera alors la fonction

get_browser qui nous dira si le User Agent du robot ou de l'utilisateur accepte ou non les iframe, les cookies, le JavaScript, etc.

Source : http://php.net/manual/fr/function.get-browser.php

Pour d'autres langages, la page Wikipédia sur le sujet détaille relativement longuement les différents codes disponibles pour le développeur : https://fr.wikipedia.org/wiki/User-Agent

Comprendre le contenu du User Agent

La plupart des User Agent suivent un schéma commun :

- Mozilla/[version]
- ([Information sur le système d'exploitation et le navigateur])
- [plateforme]
- ([détails de la plateforme])
- [extensions]

Prenons deux exemples avec les User Agent de robots, à commencer par celui de Google : *Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)*

Il se décompose donc en plusieurs parties :

- *Mozilla/5.0* : le User Agent indique qu'il est compatible avec Mozilla, le cœur du moteur Gecko utilisé par les principaux navigateurs web comme Firefox ou Chrome. Dans le cas d'un robot, cela signifie que ce dernier est compatible avec ce type de navigateur ;
- *Compatible* : un « flag » (une information) donné par le robot indiquant qu'il est compatible avec les fonctions standards des sites Internet (Cookies, iframes, Etc.) ;
- *Googlebot/2.1* : le nom du robot ;
- *+http://www.google.com/bot.html* : une URL permettant d'avoir des détails sur le robot.

Voici un second exemple avec le User Agent d'un utilisateur humain : *Mozilla/5.0 (Windows NT 5.1; rv:31.0) Gecko/20100101 Firefox/31.0*

- *Mozilla/5.0* : même signification que dans l'exemple précédent ;
- *Windows NT 5.1* : le système d'exploitation de l'utilisateur, en l'occurrence Windows XP ;
- *rv:31.0 Gecko* : la version du moteur Gecko du navigateur ;
- *Firefox/31.0* : le navigateur utilisé et sa version.

Comme on peut le constater, l'analyse manuelle est assez contraignante, surtout pour interpréter certains éléments comme la correspondance avec le nom commercial du système d'exploitation de l'utilisateur. Pour gagner du temps, on peut utiliser différents outils en ligne dans lesquels on va copier/coller le User Agent à analyser :

- http://www.useragentstring.com/ ;
- https://www.whatismybrowser.com/developers/tools/user-agent-parser/ ;

- https://useragentapi.com/.

Les User Agent des Robots

Chaque robot possède un nom précis, permettant de le reconnaître et d'éventuellement lui proposer un contenu différent ou d'en bloquer certains. Connaître ces robots peut ainsi permettre de mieux comprendre comment ces derniers crawlent votre site web. Voici donc la liste des principaux User Agent des robots liés au SEO :

- Google :
    - Googlebot ;
    - Googlebot-News ;
    - Googlebot-Image ;
    - Googlebot-Video ;
    - Googlebot-Mobile.
    - Adwords / Adsense (régie publicitaire de Google) :
        - Mediapartner ;
        - AdsBot-Google ;
        - AdsBot-Google-Mobile-Apps.
- Bing :
    - BingBot (anciennement MSNBot) ;
    - BingPreview ;
    - MsnBot –Media ;
    - AdldxBot (régie publicitaire de Bing).
- Autres moteurs de recherche :
    - Baiduspider ;
    - YandexBot, YandexImages, YandexVideo, Etc.
    - Etc.
- Outils SEO :

- AhrefsBot (Ahrefs) ;
- MJ12bot et Majestic-12 (Majestic SEO) ;
- ia_archiver (WebArchive) ;
- rogerbot (SeoMoz) ;
- HTTrack (aspirateur de site) ;

- Etc.

Ceci n'est qu'une courte liste : il existe des milliers d'autres User Agent et Robots qui parcourent le web. Pour ceux et celles qui en veulent une liste complète, une base de données contenant un très grand nombre de User Agent (dans les formats CSV, SQL et XML) est disponible à cette adresse : http://sql.sh/2290-liste-user-agents.

Les changements de noms des robots

Attention, certains robots peuvent parfois changer de nom selon les désirs et contraintes de leurs créateurs.

Si l'on prend pour exemple Google, il arrive parfois que le moteur de recherche décide de changer le nom de certains robots. En août 2015, il a ainsi changé le nom de son robot de crawl pour tous les contenus dédiés aux smartphones. Ainsi, l'ancien User Agent *Googlebot/2.1;* est devenu *Googlebot-Mobile/2.1*. Si vos scripts sont basés sur des User Agent précis, faites donc attention à vérifier ponctuellement si ces derniers ont changés ou non.

Source : https://webmasters.googleblog.com/2014/01/a-new-googlebot-user-agent-for-crawling.html

Google n'est d'ailleurs pas le seul à faire cela, puisque Microsoft le fait à chaque changement de version majeur de son navigateur Internet Edge (anciennement Internet Explorer) : https://msdn.microsoft.com/en-us/library/hh869301(v=vs.85).aspx

Trouver plus d'informations sur un robot

Il existe plusieurs moyens d'obtenir plus d'informations sur un robot précis. Tout d'abord, vous pouvez suivre le lien présent éventuellement dans la définition du User Agent, comme c'est le cas pour Google.

Autre moyen rapide : une simple recherche sur Google peut permettre de comprendre d'où vient un User Agent inconnu (notamment dans le cas d'un nouvel outil, d'un nouveau moteur de recherche ou du changement de nom d'un User Agent).

Enfin, l'outil Useragentstring cité précédemment (http://www.useragentstring.com/) permet pour certains robots de lister des informations complémentaires liées. Si l'on prend par exemple celui de Google, on peut obtenir la liste des IP et noms d'hôtes de ce robot (figure 2).

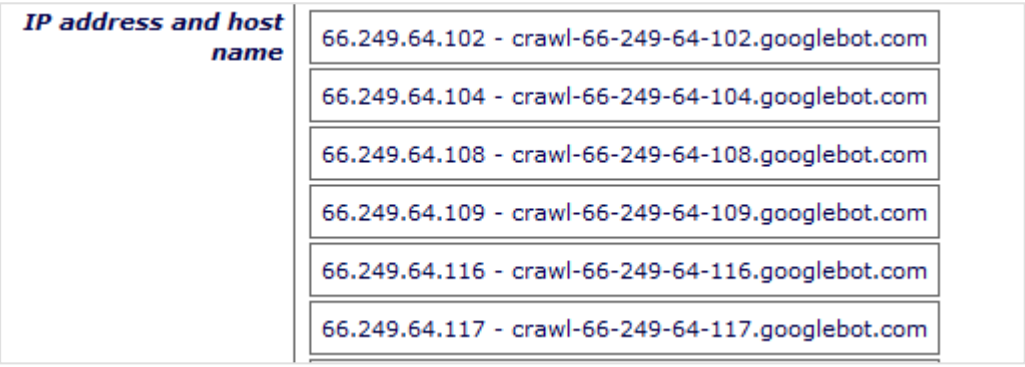


Fig. 2. Les informations complémentaires affichées par l'outil Useragentstring sur un User Agent

On peut également tout simplement se référer à la documentation officielle de chaque robot pour obtenir plus de détails sur ce dernier. Là encore, voici plusieurs exemples :

- Les User Agents liés à Windows et aux navigateurs Edge et Internet Explorer : https://msdn.microsoft.com/en-us/library/ms537503(v=vs.85).aspx ;
- Les User Agents de Chrome : https://developer.chrome.com/multidevice/user-agent ;

Pour l'IP, sachez que vous pouvez également faire une recherche par correspondance dans cet annuaire des User Agent par IP : http://www.botsvsbrowsers.com/ip/index.html

User Agent et SEO

Comprendre un User Agent et savoir le détecter est utile lorsqu'on est développeur, mais aussi pour le référenceur. Cela nous permet de pouvoir modifier notre site, comprendre comment l'internaute et les moteurs de recherche naviguent sur le site pour en améliorer le maillage interne et l'ergonomie.

Comprendre le crawl et l'indexation

C'est le premier atout : les logs vont lister toutes les URL appelées par des robots. Si vous prenez les logs sur les 30 derniers jours, vous pourrez ainsi avoir la liste complète de toutes les URL crawlées par un robot précis. L'intérêt de faire cette analyse est simple :

- Connaître les pages populaires de son site, c'est-à-dire celles que Google et les autres moteurs de recherche jugent « pertinentes » ;
- Connaître les pages non crawlés (elles ne seront pas présentes dans les logs) ;
- Connaître les pages peu crawlés, souvent par manque de liens internes ou par manque de backlinks.

Il est fréquent de se rendre compte que certains moteurs de recherche passent trop de temps sur des pages « secondaires » (par exemple les mentions légales ou la page plan du site) et à l 'inverse qu'ils ne se rendent pas assez sur des contenus pertinents. On peut également voir dans

ce type d'analyse les redirections suivies par des robots, ainsi que les pages d'erreurs qu'ils vont découvrir.

Sachez que pour ce type d'analyse il existe des logiciels ou des services en ligne, comme par exemple :

- OnCrawl : http://fr.oncrawl.com/ ;
- SpiderLog : https://spiderlog.serphacker.com/fr/ ;
- Botify : https://www.botify.com/us/

Fig. 3. Un exemple d'analyse des robots via les logs dans Botify

Source de l'image : https://www.botify.com/us/

Faire du cloaking et bloquer certains User Agents

L'autre « intérêt » d'utiliser le User Agent des robots est de pouvoir faire du cloaking, c'est-à-envoyer un contenu différent en fonction du User Agent. On peut ainsi envoyer un contenu A à Google et un contenu B aux internautes. Les référenceurs « Black Hat » utilisent régulièrement cette technique pour masquer certains contenus. On peut ainsi proposer à Google un contenu avec des liens pour améliorer son référencement naturel tout en proposant un contenu orienté uniquement vers la conversion pour l'internaute. *Rappelons que le cloaking est officiellement interdit par Google...*

On peut également choisir cette technique afin de ne pas archiver et montrer les anciennes versions de son site et de ses contenus. Le site WayBack Machine (http://archive.org/web/) est par

exemple un service en ligne qui parcourt le web et qui fait des captures à un instant T d'une URL. Des concurrents ou tout autre personne peuvent ainsi accéder facilement à d'anciennes versions de votre site web et de vos publications. En bloquant le User Agent du robot Wayback Machine, on empêche ainsi cette récupération de données.

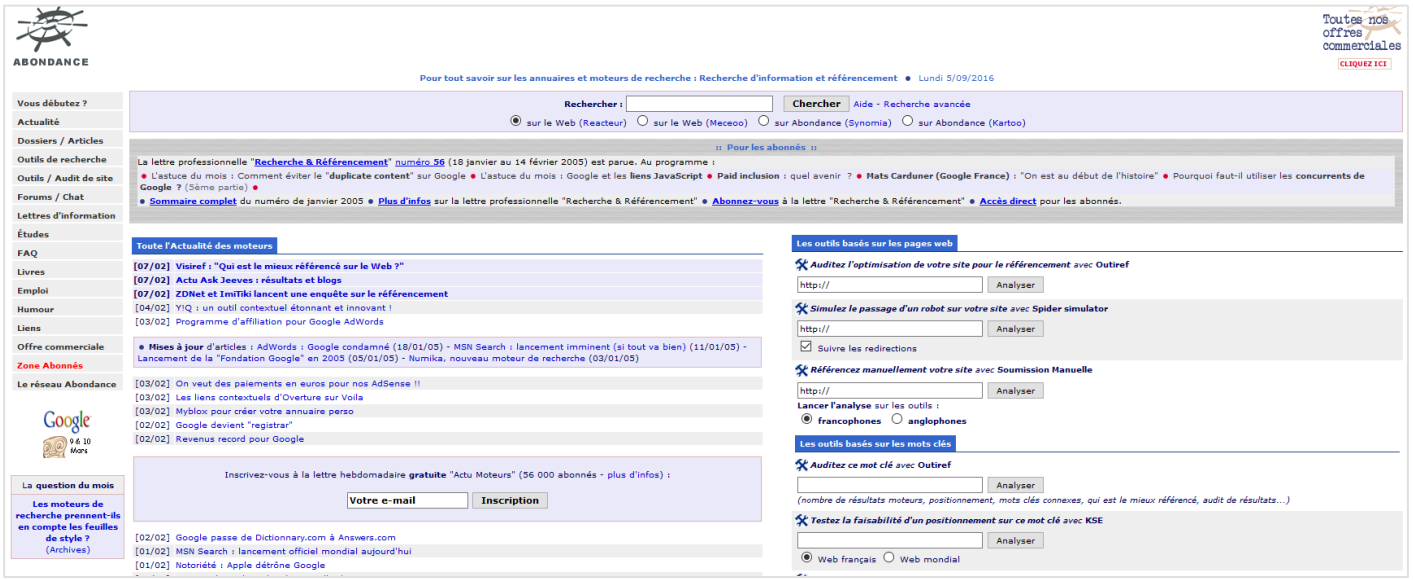


Fig. 4. Un exemple de la page d'accueil du site Abondance en 2005.

On peut également souhaiter masquer aux concurrents les liens externes que l'on a mis en place. Le fait de détecter les User Agents permet donc de masquer aux outils comme Majestic SEO ou Ahrefs les liens que l'on aurait pu placer dans nos différents contenus, le tout pour éviter que nos concurrents puissent comprendre et analyser notre stratégie. Par exemple, avec un fichier htaccess, il est possible de mettre en place des règles qui permettent de rediriger le robot ou lui afficher un page d'erreur spécifique.

Dans le code suivant, on affiche une page d'erreur pour le robot d'Ahrefs :

```
<IfModule mod_rewrite.c>
RewriteEngine On
RewriteCond %{HTTP_USER_AGENT} AhrefsBot [NC,OR]
RewriteRule ^(.*)$ - [F]
</IfModule>
```

Source : http://www.renardudezert.com/2012/01/17/freinez-les-traceurs-de-backlinks.html

Changer de User Agent

Lors de vos tests et de votre travail quotidien de référenceur, il peut être très intéressant de pouvoir visualiser le contenu comme un robot précis le verrait. Pour cela, rien de plus simple puisqu'il existe des extensions dédiées dans la plupart des navigateurs web. Il suffit d'installer cets outils et puis de choisir le User Agent désiré :

- Pour Firefox : https://addons.mozilla.org/fr/firefox/addon/user-agent-switcher/

- Pour Chrome : https://chrome.google.com/webstore/detail/user-agent-switcher-for-c/djflhoibgkdhkhhcedjiklpkjnoahfmg

- Pour Opera : https://addons.opera.com/fr/extensions/details/user-agent-switcher/?display=en

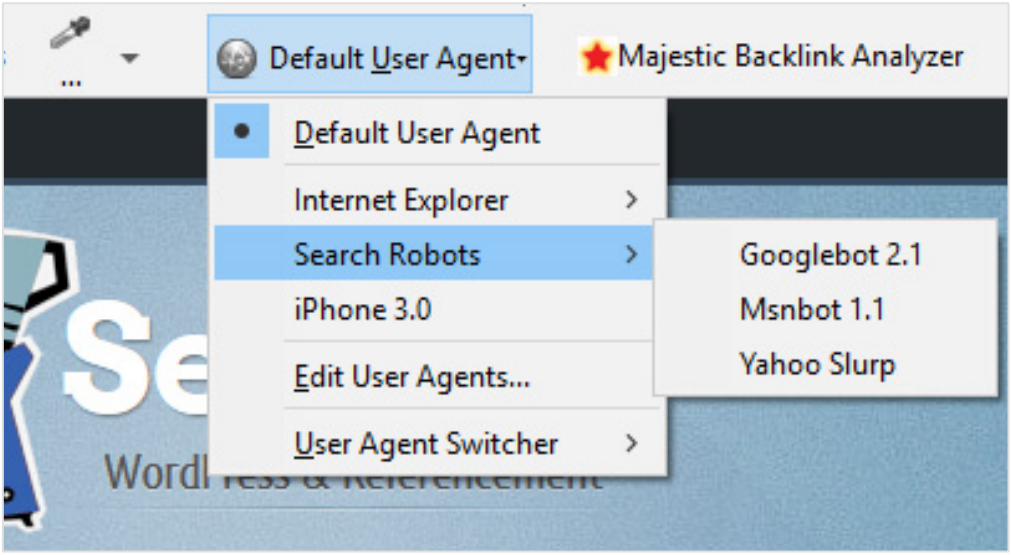


Fig. 5. Le menu de l'extension « User Agent Switcher » dans Mozilla Firefox

Conclusion

Le User Agent permet de comprendre la provenance et la nature d'une visite, qu'elle soit issue d'un robot ou d'un internaute.

Bien utilisé, il permet de faire des merveilles pour mieux comprendre comment un site est utilisé ou crawlé, permettant ainsi de prendre des décisions utiles quand à vos contenus, la structure de ces derniers et tout votre maillage interne.

Mais les défauts des User Agents sont assez simples à comprendre : ils sont plus ou moins difficiles à analyser, mais surtout ils peuvent être modifiés ou servir pour masquer et remplacer des contenus. Vous ne devez donc jamais avoir une confiance absolue dans le User Agent affiché d'un robot ou d'un utilisateur, ni dans ce qu'un site vous affiche, et toujours vérifier ce qu'il en est *in fine*.

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