

Le patron d'une grande société d'étude a dit un jour "Une bonne mesure est une mesure inflationniste".

« Faut-il entendre par là qu'il vaut toujours mieux annoncer des gros chiffres plutôt que des petits ? Suite aux e-mails que nous recevons chez Weborama, il est curieux de remarquer que lorsqu'on a deux systèmes de mesure d'audience, on écrit rarement pour se plaindre à celui qui compte trop.

Je rêverai de lire des courriers comme : "Bonjour monsieur, je vous écris car vous comptez beaucoup trop de trafic sur mon site, je suis sûr que je n'ai pas autant de visiteurs !" Mais non, étrangement nous ne recevons que des e-mails du type, "Je ne comprends pas avec mon analyseur de Log, j'avais 6 fois plus de pages vues !!!"... effectivement la différence est notable et mérite quelques explications. »

Sunny Paris, Associé et Directeur de la recherche et du développement chez Weborama.



La mesure d'audience par l'analyse des Tag et des Log : Quelle méthodologie pour quelle fiabilité ?

Introduction

Il existe, en effet, des écarts considérables entre la mesure par l'analyse des fichiers Log et la mesure par l'analyse des marqueurs (Tag) placés sur les pages d'un site. Concrètement, là où un analyseur de Log mesure 23 240 pages vues, un analyseur de Tag en mesure seulement 5 689. Mais ces écarts de calcul ne sont pas vraiment surprenants, car ils relèvent en fait de deux méthodologies de comptage bien différentes.

A travers ce Livre Blanc nous souhaitons apporter des explications sur les différences observées entre ces deux méthodologies. En effet, la question de la fiabilité se pose et doit être résolue pour que le marché de la mesure d'audience des sites Internet soit plus transparent.

L'étude que nous vous délivrons a été également réalisée par OJD-Diffusion / Contrôle, lorsque cet organisme a décidé de certifier une méthodologie de comptage, en l'occurrence celle par l'analyse des Tags.

1. Qu'est-ce qu'un Log ?

Historiquement, l'analyse des fichiers Log est la première méthodologie qui a été développée. Les serveurs Web génèrent automatiquement des fichiers qui rendent compte de la totalité des connexions. Ces fichiers, de type texte, sont appelés fichiers Log. Les Log comptabilisent toutes les adresses IP qui se sont connectées sur le serveur. Pour être exploitables, les données qu'ils fournissent doivent être lues et analysées par un logiciel qui les traduit en statistiques. Cette approche nécessite donc l'installation d'un logiciel particulier sur le serveur audité.

Le principe est le suivant, le navigateur qu'utilise un internaute pour visiter un site, fait une requête auprès du serveur, celui-ci envoie alors tous les fichiers demandés (le frameset, les 3 cadres associés, la feuille de style, les 2 fichiers .js, l'animation en flash, les images, etc. par exemple). Le serveur Web enregistre ces informations dans un fichier (le Log) et un programme l'analyse ensuite : pour 53 requêtes effectuées, 4 pages ont été réellement vues.

Par contre, ce qu'un analyseur mesure particulièrement bien, ce sont le nombre des fichiers téléchargés sur le serveur, le volume total de données transférées, le nombre de connections à la seconde, etc.

2. Pourquoi la mesure des fichiers Log n'est pas fiable ?

L'analyse des données recueillies à partir des fichiers Log est la plus ancienne et la moins coûteuse. En revanche, cet outil manque d'exhaustivité et les analyses statistiques qui en découlent sont imprécises. Nous allons passer en revue les principaux biais induits par la mesure des Log :

- La mémoire cache des serveurs proxy et des navigateurs.
- La multiplicité des serveurs (cluster).
- Les adresses IP dynamiques.
- Les Firewall d'entreprises.
- Les frames HTML.
- Le trafic artificiel généré par les robots et les outils de monitoring.

2.1. La mémoire cache des serveurs proxy et des navigateurs

L'analyseur de Log connaît des limites inhérentes à sa conception : pour limiter le volume des données qui transite sur le réseau, les serveurs proxies et la mémoire cache des navigateurs stockent les pages et font chuter les statistiques de consultation de 10% à 20%.

Un serveur proxy est un serveur relais, qui permet à un fournisseur ou un intermédiaire d'accès de stocker les pages Web qui font l'objet des requêtes les plus fréquentes. Sur le plan technique, l'utilisation de proxies permet d'optimiser les échanges en allégeant la consommation de bande passante.

Quant on fait une analyse des fichiers Log, les requêtes qui se trouvent dans le cache du navigateur ou qui sont gérées par les serveurs proxy ne sont pas signalées au site principal. Elles n'apparaissent pas dans le fichier Log et ne sont donc pas comptabilisées. L'utilisation de marqueurs (Tag) permet de pallier ce biais.

2.2. La multiplicité des serveurs (cluster).

Pour les sites qui ont un trafic très important et qui nécessitent la mise en place de cluster, des difficultés de comptage apparaissent également. Dès lors qu'un site utilise plusieurs serveurs différents, répartis chez des hébergeurs distincts, le regroupement des fichiers Log peut s'avérer très lourd, voire impossible.

2.3. Les adresse IP dynamiques

C'est une adresse affectée à tout appareil connecté à l'Internet, et qui est différente à chaque connexion. Lorsque l'analyseur de Log comptabilise les adresses IP qui sont venues se connecter sur un site, il n'est donc pas capable de reconnaître un visiteur qui est venu plusieurs fois dans la semaine par exemple. Il sera possible de savoir combien de connexions ont été effectuées sur une période donnée, mais il sera impossible de connaître le nombre de visiteurs sur cette même période.

2.4. Les Firewall d'entreprises

Un firewall est un ordinateur placé entre le réseau d'une entreprise et l'Internet qui permet de bloquer toutes connexions d'utilisateurs non référencés. Dans une entreprise, par exemple, tous les salariés utilisent la même adresse IP protégée par le firewall. Le problème de l'identification de l'internaute en temps que visiteur unique se pose ici aussi. La mesure des Log est quantitative mais absolument pas qualitative.

2.5. Les frames HTML

Une page Web peut contenir plusieurs frames pour être plus flexible et pour pouvoir rajouter facilement des fonctionnalités. Il y a un fichier HTML pour la page et pour chaque frame la constituant. L'analyseur de Log comptabilise uniquement les fichiers HTML et ne peut pas faire la différence entre une page et les différentes frames.

Un analyseur de Tag, comme Weboscope™, n'est pas confronté à ce problème. En effet le marqueur placé en bas d'une page HTML comptabilise la page et l'ensemble des frames la constituant comme une seule page, car c'est la page dans son intégralité qu'il est pertinent d'analyser.

2.6. Le trafic artificiel généré par les robots et les outils de monitoring

Les programmes lancés par les outils de recherche surfent sur le Web d'une façon totalement automatisée. En suivant les liens hypertextes, ils scannent le contenu des sites. Le trafic artificiel que génèrent les robots (appelés également « spiders » ou « crawlers ») et les outils de monitoring (sonde http) est très important sur le Web.

Les robots ne sont pas des internautes, et il paraît évident de les exclure de la mesure de l'audience d'un site. Mais le serveur Web connecté au navigateur a très peu de moyens à sa disposition pour savoir si derrière le navigateur il y a un véritable internaute, un robot ou une sonde.

Cette distinction internaute / robot / sonde n'est alors pas visible dans les comptes rendus statistiques des analyseurs de Log. En effet à partir du moment qu'une requête a été envoyée au serveur, et ce quelle que soit son origine, une page vue sera comptabilisée comme telle. C'est un autre facteur qui explique les différences existantes entre les deux méthodologies : Log / Tag.

Un outil comme Weboscope TM ne comptabilise pas les requêtes effectuées par des robots ou par des outils de monitoring comme des pages vues. En effet, le Tag inscrit sur la page HTML d'un site a été spécifiquement réalisé par Weborama pour ne pas pouvoir être exécuté par ce type d'outils.

Conclusion

Cette étude comparative que nous avons réalisé entre la méthodologie par Tag et celle par Log est importante, parce qu'elle permet de véritablement comprendre les enjeux de la mesure d'audience.

D'un côté, il y a les directions informatiques qui ont besoin de savoir combien de fichiers ont été téléchargés par rapport à la capacité de leurs serveurs, quel est le volume total de données transférées, le nombre de connections à la seconde, etc.

Et d'un autre côté, il y a les directions marketing / Internet, etc. qui souhaitent connaître leur audience en termes de pages réellement vues, de visiteurs uniques, de circulation sur le site, de profil des internautes, etc. Elles ont besoin de données précises et fiables pour améliorer leur site, le faire évoluer, commercialiser de l'espace publicitaire, etc. C'est pour cette raison que OJD-Diffusion / Contrôle a décidé de certifier la mesure « site centric » par marqueurs (Tag).

OJD-Diffusion / Contrôle est une organisation professionnelle tripartite qui regroupe les principaux acteurs du marché Internet français : annonceurs, éditeurs et prestataires Internet. L'association a vocation à contrôler, et à certifier la mesure de la fréquentation, de tous les sites édités sur le territoire français, de tous les sites dont l'éditeur est français et de tous les sites à contenu francophone.

Dans cette optique, OJD-Diffusion / Contrôle a défini en collaboration avec les principaux acteurs du marché, dont Weborama, un cahier des charges que les outils de mesure de trafic de sites Internet doivent rigoureusement respecter.

Pour être labellisés par OJD-Diffusion / Contrôle, les outils de mesure d'audience doivent notamment respecter les points suivants :

- Utiliser la méthodologie « site centric »
- S'affranchir des biais induits par la mémoire cache (proxy et/ou navigateur)
- Passer au travers des Firewall d'entreprises
- Exclure les robots
- Filtrer les IP internes

L'outil Weboscope™ développé par Weborama est labellisé par OJD-Diffusion / Contrôle depuis 2000. De plus, la société continue de collaborer avec OJD afin d'améliorer le Cahier des charges notamment en y incluant les notions de visiteurs uniques et la certification des « newsletters » électroniques.

Pour en savoir plus sur OJD, consulter leur site : www.ojd.com

Tableau récapitulatif	
LOG	TAG
I n s t a l l a t i o n	
Cette technologie nécessite l'installation d'un logiciel d'analyse de Log sur le serveur.	Cette technologie nécessite l'inscription d'un code Javascript sur chaque page du site à auditer.
F o r m a t	
Le format des informations statistiques (propre à la technologie utilisée) est difficilement lisible, interprétable et exploitable.	Les données résultant d'une analyse par Tag sont lisibles et exploitables au travers d'une console de consultation rapide et simple.
M é m o i r e s c a c h e s e t P r o x i e s	
Les informations localisées dans le cache du navigateur ou dans le serveur proxy sont estimées à 20% du volume des pages consultées.	Les Tag analysent précisément la fréquentation de chaque page en prenant en compte les informations incluses dans le cache du navigateur et du serveur proxy.
M é t h o d o l o g i e s	
L'analyse des Tag se base essentiellement sur l'analyse des adresses IP se connectant au site. L'utilisation d'un Firewall implique que derrière 100 personnes connectées il n'y a, en fait, que 1 adresse IP reconnue par l'analyse des Log.	L'analyse des Tag se base sur l'analyse des IP ainsi que sur l'utilisation de cookies, installés dans le navigateur de chaque ordinateur qui se connecte à un site. Par conséquent les firewall ne sont pas pris en considération.
En utilisant un ISP (Internet Service Provider), un internaute A utilise une IP dynamique qui peut être aussi utilisée par un internaute B. Le Log considère seulement une IP et par conséquent 1 visite au lieu de 2 visiteurs.	En plaçant un cookie dans le navigateur du visiteur, les internautes qui utilisent une même IP seront considérés comme des visiteurs distincts.
L'analyse des Log comptabilise tous les éléments d'une page (frames, gifs, jpegs, pop up, pop under,...) comme plusieurs pages.	Le Tag s'installe dans la frame principale de chaque page (frame de contenu). Les Tag prennent en compte la page dans son intégralité.
S i t e s D y n a m i q u e s	
Les Log ne sont pas adaptés à la mesure du trafic des pages dynamiques de plus en plus répandues sur le Web.	Les Tag sont parfaitement adaptés à la mesure des sites dynamiques.
N o u v e l l e s T e c h n o l o g i e s	
Les Log mesurent mal les sites utilisant les autres technologies du Web telles que : Flash, Rich Media (vidéo), Wap, etc.	Les Tag mesurent les pages sécurisées, en Flash, RichMedia, Wap, etc. Et ont une forte capacité d'adaptation à tous types d'innovation.
S i t e s c o m p l e x e s e t s i t e s d ' u n g r o u p e	
Pour les sites qui ont beaucoup de trafic, les fichiers Log sont plus lourds et plus compliqués.	Les Tag mesurent avec la même facilité les sites qui ont un petit ou un très grand trafic.
Les sites qui exploitent différents systèmes et qui ont plusieurs entrées sur différents serveurs sont très compliqués à analyser avec les Log.	La technologie Tag fonctionne avec tous les systèmes d'exploitation ainsi qu'avec toutes les plateformes existantes et indépendamment du nombre de serveurs.
Les Log ne permettent pas d'obtenir des informations globales sur l'ensemble des sites d'un même groupe.	Avec les Tag, il est facile de mutualiser des informations de trafic des sites d'un groupe. L'utilisation de cookies permet d'analyser les visiteurs uniques sur l'ensemble des sites : « audience transversale ».
R o b o t s	
Les Log ne font pas la distinction entre les robots programmés pour crawler les sites et les vrais internautes.	Les Tags ne peuvent pas être exécutés par les robots.

La mesure d'audience par l'analyse des Tag et des Log : Quelle méthodologie pour quelle fiabilité ?

Weborama – Source d'e-novation

Créée en 1998, la société Weborama offre à ses clients des services innovants pour optimiser leurs stratégies marketing on-line en termes d'efficacité et de ROI. Weborama propose, à cet effet, une gamme de produits adaptée à tous les besoins sur Internet et les nouveaux médias. Ces outils d'aide à la décision permettent aux directions marketing et opérationnelles de connaître l'impact et le retour sur investissement de toutes les actions e-marketing. Weborama, reconnue société « innovante », est soutenue par l'ANVAR. Weboscope™ 4.0 permet non seulement une parfaite mesure du trafic d'un site Internet ou Intranet, mais aussi une analyse pertinente du profil et du comportement des internautes. Weboscope™ 4.0 est labellisé par OJD-Diffusion / Contrôle. Noos, Groupe BNP Paribas, PSA, Arte, MSN, Société Générale, etc. font partie des nombreuses références de Weborama.

Contact :

Weborama SA
15 rue Clavel
75 019 Paris
Tel : +33 (0)1 53 19 21 40
Fax : +33 (0)1 53 19 21 41
weborama@weborama.com
www.weboscope.com et www.weborama.com