PowerBoutique®

Mesure d'audience sur Internet

différences fondamentales entre les solution basées sur les journaux et celles utilisant le marquage de site.

L'objectif de ce document est de répondre à la question légitime : « Pourquoi existent-il des écarts de visites entre Webalizer et Google Analytics ? »

En effet, lors du processus complexe d'analyse des performances d'un site web, on peut mettre en œuvre différents outils qui semblent fournir les mêmes informations (par exemple les visites). Or ces informations ne sont pas du même ordre de grandeur car ne mesurent pas les performances selon les mêmes critères. Nous allons tenter de vous expliquer simplement les raisons de ces écarts.

1. Navigateur et Serveur web

Lors de la consultation d'un site internet, deux éléments principaux sont mis en jeu : votre navigateur d'un coté et un *serveur web* de l'autre.

1.1. Le serveur web

Le serveur web est un logiciel, développé selon une norme : le protocole HTTP (rfc 2616). Cette norme permet à tous les navigateurs de discuter avec les serveurs Web en définissant un langage simple commun à tous.

Cette norme définit :

- L'adressage des *ressources* sur le réseau internet : les URL (Uniform Ressource Locator),
- La façon d'écrire les dates/heures,
- Le langage à utiliser pour interroger un serveur,
- …

1.2. Le navigateur

Le navigateur dialogue avec les serveurs : il est le logiciel qui interroge divers serveurs dans le but de générer un affichage compréhensible par l'utilisateur des différentes ressources.

1.3. Les requêtes sur le serveur

Les serveurs web ne comprennent pas les notions de *session de navigation*, de *page d'entrée* (ou de sortie), et ne font pas de différences entre une page, ses dépendances ou tout autre forme de document téléchargeable. Pour eux, un site Internet est composé de ressources (images, css, pdf, …) auxquelles on accède par des requêtes unitaires (ou *Hits*). On appelle requête l'action d'interrogation effectuée par un navigateur vers une ressource du serveur.

1.4. Les journaux d'accès (access log)

Tout au long de son fonctionnement, un serveur web va maintenir, pour des raisons de traçabilité, un journal détaillé des requêtes qui lui ont été faites, et des réponses qu'il y a apportées. Ces

journaux d'accès (*access log*) sont stockés et archivés sur le serveur. Ils contiennent tout un lot d'informations, par exemple :

- Qui a fait la demande ?
- Quand a-t-elle été faite ?
- Sur quelles ressources elle portait ?
- Quel code a été renvoyé par le serveur en retour (Par exemple : code 200 => ok, code 301 => ressource déplacée, code 404 => ressource introuvable, …) ?
- Quelle est la taille des données envoyées à l'internaute ?
- Quel était le "référant", c'est-à-dire l'emplacement où le lien vers la ressource a été trouvé (par exemple le nom de la page pour une image de fond, ou la page précédente pour une nouvelle page, …) ?
- Quel était le navigateur utilisé (UserAgent) ?
- …

Extrait de journal :

```
127.0.0.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326
"http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"
```

2. Les analyseurs de log web

La première méthode de mesure d'audience existante a été basée sur ces journaux, il s'agit des outils de type « d'analyseurs de log web ».

Exemple d'analyseur de log web : les « détails statistiques serveur » de PowerBoutique, autrement appelées « Statistiques Webalizer »

2.1. Méthode de mesure d'audience des analyseurs de log web

Ces outils vont parcourir le journal produit par le serveur pour en extraire des informations utiles, les notions de *pages vues* ou de *visites* par exemple. Toutefois tous les outils disponibles sur le marché n'identifient pas forcement les pages de la même façon, pas plus qu'ils ne déduisent les visites en utilisant les mêmes critères.

De façon générale pour cette famille de logiciels, une *page vue* est l'accès à une ressource identifiée par l'administrateur ou par les concepteurs du logiciel comme tel, et une visite est l'ensemble des requêtes à différentes *pages* effectuées par un même couple adresse IP + UserAgent, et ce avec un temps limité entre deux accès à une page (ex moins de 30minutes entre deux requêtes).

2.2. Les limites des analyseurs de logs web

Si elle reflète un état exact de l'activité générée par le serveur, cette technique n'est cependant pas parfaite puisqu'elle inclut des données pouvant être inutiles pour analyser l'activité « humaine » du site Internet.

Exemples de données comptabilisées par les analyseurs de logs web :

- Les erreurs (404 : page introuvable, pouvant alerter de visites perdues ou d'erreurs dans la conception du site, 401 : tentative d'accès à une ressource protégée, …),
- Le suivi des redirections (301/302),
- Les différentes requêtes des systèmes d'aspiration de site,
- Celles réalisées par les robots d'indexation des moteurs de recherche,

- …

Dans certaines conditions cette technique peut aussi ignorer certaines visites ou affichage de pages. En effet, quand le serveur web fourni une *date d'expiration* pour une ressource, les navigateurs ont la possibilité de stocker la ressource et l'afficher plusieurs fois sans avoir à la redemander au serveur web, et donc sans que cet affichage soit comptabilisé.

De ce fait, elle montre de grosses faiblesses pour distinguer les visiteurs « humains » et analyser leurs comportements.

3. Le marquage des sites web

Une autre technique a donc vu le jour, basée sur le marquage des sites web et popularisée par des outils tels que "Google Analytics" ou "Webtrend",.

3.1. Méthode de mesure d'audience par marquage des sites web

Le principe est le suivant : l'outil (par exemple Google Analytics) fournit un bloc de code html à intégrer sur les différentes pages de votre site. Lors du passage d'un internaute, ce code présent sur la page va déclencher une requête vers les serveurs de la solution de mesure (par exemple les serveurs de Google Analytics). Les informations transmises vont être stockées dans une base de données sous une forme adaptée à l'analyse.

Selon les techniques utilisées pour le marquage (javascript, cookies, web-bug, …) il va être possible de stocker plus d'informations.
Exemples d'informations pouvant être fournies par cette méthode :

- La taille de l'écran,
- L'analyse du comportement de la souris,
- Le suivi d'un visiteur sur plusieurs sites,
- La prise en compte du changement d'adresse de la connexion Internet de l'internaute,
- Le "retour" d'un internaute sur un site,
- …

Une chose ne change pas toutefois, l'analyse de la *session*, qui sert de base au comptage des visiteurs, de l'internaute se fait toujours en se basant sur des requêtes espacées de moins de XX minutes (souvent 30).

3.2. Les limites du marquage des sites

De part sa conception, cette technique d'analyse s'appuie uniquement sur l'interprétation qu'en fait le navigateur client. Les données issues de cette analyse ne peuvent donc être traitées comme des données absolues puisqu'elles dépendent à la fois des paramétrages mais aussi des possibilités offertes par le navigateur.

Exemples de cas où la requêtes peut ne pas être prises en compte dans ce mode d'analyse :

- Navigation via un navigateur en mode texte, ou autre système dédié,
- Lorsque le javascript ou les images sont désactivées sur le poste de l'internaute,
- En cas d'utilisation de plugins anti-pub ou de protection de vie privée,
- En cas de requêtes effectuées sans navigateur (robots des moteurs de recherche, système de

copie de sites, …),
- En cas de requêtes effectuées par des sites externes sur vos images,

De plus, comme cette méthode dépend d'un service externe au site web visité, le navigateur de l'internaute doit, pour contacter le service de collecte (par exemple Google Analytics), effectuer une résolution DNS supplémentaire et ouvrir une connexion à un autre serveur. Cette action peut entraîner un ralentissement de l'affichage du site, ou empêcher la collecte d'informations.

4. Avantages de chaque méthode

Nous avons vu le fonctionnement de chaque méthode, ce qui va maintenant nous aider à mieux comprendre leurs avantages.

4.1. Les avantages de la solution d'analyse de journaux

- Les serveurs web génèrent déjà les fichiers de journaux, donc la donnée est accessible "de base", sans ajout de code supplémentaire sur le site web.
- Le serveur web stocke de façon fiable toutes les opérations réalisées, incluant la fourniture de documents PDF, d'images, … et ne nécessite pas la coopération du navigateur.
- Les données d'analyse sont stockées sur les serveurs hébergeant le site, dans un format standard, non propriétaire. Cela permet un changement simple de logiciel d'analyse, l'utilisation de plusieurs outils, ou de refaire une analyse d'anciennes données avec un nouveau logiciel.
- Les journaux contiennent des informations de visites des robots (qui en général ne traitent pas le JavaScript et donc ne sont pas comptabilisées par "marquage"). Bien que ces visites ne devraient pas être incluses dans les flux "humains", ces informations sont des données utiles pour les personnes travaillant sur le référencement du site Internet.
- Il n'y a pas d'appel vers des serveurs externes pouvant ralentir le chargement des pages ou empêcher le comptage de certaines visites.

4.2. Les avantages de la solution de marquage des sites

- Lorsqu'un internaute visite une page qui avait été mise en cache par son navigateur, cette visite est comptabilisée (alors que pour l'analyseur de logs, la mise en cache évite de renouveler les requêtes vers le serveur, donc la visite n'est pas comptabilisée).
- Le contrôle des données d'audience est plus efficace car plus proche de la réelle activité humaine sur le site.
- Des informations supplémentaires sont collectées : taille de l'écran, articles achetés, …
- Cette méthode peut offrir des possibilités de suivi avancé, n'impliquant pas le serveur web
- La traçabilité des visiteurs est plus performante : meilleur suivi des visiteurs utilisant une adresse IP dynamique, possibilité de détection des retours d'anciens visiteurs (grâce aux Cookies), …
- L'interface de consultation des données est souvent plus ergonomique et offre davantage de possibilités sur la mesure de l'évolution des différents indicateurs.

5. Conclusion

Chaque méthode a son propre mode de fonctionnement et comptabilise les données selon ses

propres critères. Il est donc logique que ces outils remontent des valeurs différentes pour des informations à priori identiques (par exemple pour les *visites*).

Au regard de leurs avantages et limites respectives, il est conseillé d'utiliser les deux méthodes, en connaissance de cause, selon les critères que vous jugez important d'analyser.