Le référencement de sites web dynamiques (1ère partie)

Retour au sommaire de la lettre

Le référencement de sites web dynamiques est l'une des principales sources d'interrogations des webmasters actuels. Après avoir longtemps été un facteur totalement bloquant pour les moteurs de recherche, la situation s'est assouplie depuis quelques mois. Quels sont les types de pages qui sont compatibles avec les moteurs et ceux qui ne le sont pas ? Comment contourner les obstacles ? Voici quelques éléments de réponse dans cette première partie d'un article consacré à ce vaste sujet...

Il existe, et cela est vrai depuis que les moteurs de recherche existent, un certain décalage de temps entre le moment où les techniques de création de sites web sont utilisées et la façon dont les moteurs de recherche les indexent.

Cela s'est vérifié pour les frames (souvenez-vous d'Excite qui ignorait totalement les sites ainsi réalisés), puis pour le Flash ou le Javascript, par exemple. Cela se vérifie encore avec les sites web dynamiques, qui ont longtemps représenté un obstacle rédhibitoire pour les moteurs. La situation semble, certes, s'améliorer aujourd'hui, mais elle n'est pas encore parfaite, loin de là.

Dans cette série d'articles, nous allons passer en revue les différents obstacles que représente ce type de site pour les moteurs, ainsi que les différentes solutions proposées par les sociétés de référencement, avant de faire un "focus" sur les techniques d'"url rewriting", qui représentent souvent la meilleure solution face à ce phénomène.

Qu'est-ce qu'un site dynamique ?

Avant d'aller plus loin dans cet article, il est nécessaire de définir ce qu'est un site dynamique, par opposition à un site statique. Le site statique gère des pages créées au préalable. Il va avoir à sa disposition, sur son disque, des pages HTML dites "statiques", qu'il va afficher "telles quelles" dès qu'un internaute les demande. Les pages sont donc créés à l'aide d'un éditeur HTML, puis stockées pour être affichées sous leur forme initiale.

Le site dynamique, pour sa part, puise ses informations dans une base de données (qui peut être d'origines diverses) et crée des pages "à la volée", en fonction d'une action ou d'un évènement. Par exemple : une saisie effectuée par un internaute. L'exemple type de site dynamique, est... le moteur de recherche !

En effet, un internaute, lorsqu'il arrive sur un moteur, saisit une requête dans un formulaire et l'outil, sur la base des mots clés demandés, va créer une page de résultats "sur mesure" en fonction des termes demandés. Bien entendu, cette page n'existe pas en tant que tel sur le disque dur du moteur, et elle est donc créée "à la volée". Un moteur de recherche est donc un site "dynamique" !

Il en sera de même avec des sites web d'E-commerce, par exemple dans le cadre d'un catalogue en ligne, mais également la consultation d'archives de presse, etc.

Ce qui bloque le plus souvent les moteurs de recherche est représenté par l'url des pages, qui contient, pour ce type de sites, deux caractères spécifiques et représentatifs des sites dynamiques : le point d'interrogation (?) et l'esperluette (&).

Format d'une url de site dynamique

En effet, l'url d'une page émanant d'un site dynamique est le plus souvent affichée sous une forme du type :

http://www.sitedynamique.com/prog.cgi?kw=motcle&langue=fr&zone=france&encodage=ISO-8859-1

Cette adresse peut s'interpréter ainsi : "sur le site www.sitedynamique.com, on a lancé le programe nommé prog.cgi en lui passant comme paramètres les variables *kw* (de valeur "motcle"), *langue* (de valeur "fr"), *zone* (de valeur "france") et *encodage* (de valeur "ISO-8859-1").

Il en est exactement de même sur Google. Si vous allez sur le site http://www.google.fr/ et que vous tapez le mot clé "abondance", l'url de la page de résultat aura comme intitulé :

http://www.google.fr/search?q=abondance&ie=ISO-8859-1&hl=fr&btnG=Recherche+Google&meta=

Sur Google, c'est le programme nommé "search" qui a été lancé, avec pour paramètres :

- *q* = abondance (le mot clé)
- *ie* = ISO-8859-1 (l'encodage des caractères)
- *hl* = fr (la zone lingusitique)
- *btnG* = Recherche Google (le nom du bouton de validation de Google)
- *meta* = (autre infomation - vide dans ce cas - pour le moteur).

> ***Nota*** : Google utilise pour son formulaire de recherche la méthode "GET" (passage de paramètres dans l'url) contrairement à un moteur comme celui de Free, par exemple, qui utilise, sur sa page d'accueil, la méthode "POST". Dans ce cas, la page de résultat a une url identique quel que soit le mot clé recherché (http://search1-2.free.fr/google.pl). La méthode "POST" est rédhibitoire pour l'indexation des pages dynamiques puisqu'une seule url est proposée aux robots pour chaque page. L'adresse des documents n'est donc plus différentiatrice de leur contenu...

Dans une url de site dynamique :

- ***Le point d'interrogation (?)*** va donc signifier un ***passage de paramètres à un programme***.

- ***L'esperluette (&)*** va ***séparer les différents paramètres***, et leur valeurs, entre eux.

Voici quelques exemples (réels) d'urls dynamiques :

http://www.nova-cinema.com/main.php?page=search.en.htm

http://canadapost.internic.ca/search.asp?lang=fr

http://www.rcsec.org/ns/french/search.cfm?V=search

http://www.gladnet.org/index.cfm?fuseaction=research.search&CFID=145633&CFTOKEN=178

http://c.ekzay.org/codemaster/tdj/modules.php?op=modload&name=Search&file=index

On pourrait ainsi multiplier les exemples à l'infini. **Retenons, pour l'instant, qu'une url dynamique contient un point d'interrogation (?) qui marque le début du passage de paramètres à un programme, chacun des ces paramètres étant séparé par une esperluette (&).**

Le plus souvent, les sites dynamiques sont créés sur la base de technologies de programmation comme PHP, ASP ou CFM. Mais ils peuvent également être bâtis au travers de produits propriétaires (qui poseront plus ou moins de problèmes supplémentaires) comme Lotus Notes, Vignette, Broadvision, etc.

Pourquoi les moteurs de recherche n'indexent-ils pas les sites dynamiques ?

Le fait que les urls dynamiques aient un format spécifique ne nous explique pas pourquoi elles sont refusées par les moteurs de recherche. Il y a en fait plusieurs explications à cela :

- Tout d'abord, les robots ne savent faire, grosso modo, que deux choses : lire, puis stocker, du code HTML et suivre des liens. Ils ne savent pas taper des mots clés dans des formulaires pour obtenir des pages de résultats. Il sera donc difficile, pour les robots des moteurs, d'indexer des pages de résultats de Google, par exemple, si la seule façon d'afficher ces dernières consiste à

taper des mots clés dans un formulaire (en revanche, ils sauront suivre un lien qui pointe sur une de ces pages et contenant, donc, les paramètres de la recherche dans leur url)...

- Le nombre de pages créées "à la volée" par un site dynamique peut être quasi infini. En effet, prenez un catalogue du type de ceux d'Amazon ou de la Redoute, multipliez le nombre d'articles par le nombre d'options possibles (délai d'envoi, couleur, taille pour des vêtements, autres possibilités diverses et variées) et vous obtenez rapidement, pour un seul site, plusieurs centaines de milliers, voire millions de pages web potentielles présentant chaque produit de façon unique. Difficile, pour un moteur, de les indexer toutes ou, en cas contraire, de savoir où s'arrêter.

- Un site web dynamique a la possibilité de créer, en quelques secondes, des milliers de pages "à la volée". Il s'agit également là d'un système à haut risque pour ce qui concerne le spam contre les moteurs. Dans ce cas, ces derniers "se méfient" et, parfois, optent pour l'option la moins risquée... Ils préfèrent ne prendre en compte aucune page plutôt que de courir le risque de devenir un "réservoir à spam" au travers de techniques de création incessante de pages... un peu trop optimisées...

- Une même page, proposant le même contenu, peut être accessible à l'aide de deux urls différentes (ce problème est notamment crucial en ce qui concerne les identifiants de session, voir plus loin). Cela risque d'être problématique pour un moteur, qui devra alors mettre en place des procédures de "dédoublonnage" qui peuvent s'avérer complexes...

- La longueur excessive de certaines urls, passant de nombreux paramètres, peut également poser des problèmes aux moteurs. D'autre part, certains caractères apparaissant des ces adresses (#, {, [, |, @, etc.) peuvent également parfois être bloquants, tout comme les lettres accentuées, peu fréquentes dans les urls statiques, qui peuvent causer des soucis de codage.

Certains problèmes posés par les sites web dynamiques sont appelés "spider traps" : il s'agit de pages mal reconnues par les "spiders" des moteurs, qui s'y perdent parfois dans des boucles infinies et indexent alors des milliers de documents différents représentatifs de quelques pages web uniquement.

Quels formats sont rédhibitoires ?

Comment un moteur de recherche réagit-il face à une page dynamique ? Il y a de cela quelques mois, voire quelques années, elles étaient purement et simplement ignorées. Pour certains moteurs, les pages en PHP, ASP ou CFM étaient bannies, quelle que soit leur forme. Heureusement, cette période est aujourd'hui révolue... Le simple fait d'avoir été créée dans l'un de ces langages de programmation n'est plus rédhibitoire. Ouf...

En effet, à l'heure actuelle, les moteurs de recherche reconnaissent de façon bien plus optimale les pages dynamiques. Mais la situation n'est pas encore idéale et certains blocages sont encore présents. Globalement, il en existe deux très importants : le nombre de paramètres passés dans l'url et l'identifiant de session.

Nombre de paramètres passés dans l'adresse

Dans un premier temps, il semblerait que les urls contenant un ou deux paramètres ne posent pas (plus) de problèmes aux moteurs. Exemples d'adresses aujourd'hui acceptées par ces derniers :

http://www.sitedynamique.com/search.cgi?kw=motcle
http://www.sitedynamique.com/search.cgi?kw=motcle&langue=fr

Ce fait est avéré sur des moteurs comme Google et Yahoo!, par exemple.

En revanche, jusqu'en 2003, ce type d'url (passage de plus de deux paramètres dans l'adresse) était réfusé :

http://www.sitedynamique.com/search.cgi?kw=mc&langue=fr&zone=france
http://www.sitedynamique.com/search.cgi?kw=mc&langue=fr&zone=france&codage=ISO-8859-1

Il semblerait, cependant, que la situation s'améliore de ce côté. On voit de plus en plus de pages possédant trois, quatre, voire plus de paramètres dans leur url présentes dans les index respectifs de Google et de Yahoo!. Cependant, même si cette situation est meilleure en 2004, elle reste encore bloquante dans de nombreux cas. Il vous faudra donc en tenir compte lors la mise en place de votre site afin de passer le moins de paramètres possible dans vos adresses. Allez au strict minimum. Pour l'instant, on peut encore estimer que le chiffre de deux paramètres est un maximum... Au delà, il vous faudra envisager une solution technique adéquate (voir nos articles prochains).

L'identifiant de session

Le site web sur lequel vous naviguez a souvent besoin de vous "tracker", c'est-à-dire de suivre votre navigation au travers de ces pages. Exemple typique : une boutique en ligne qui doit se souvenir en pemanence de ce que vous avez mis dans votre "caddie virtuel".

Il existe deux manières principales d'effectuer ce type de mémorisation de vos visites : le *cookie* et l'*identifiant de session*. Dans ce dernier cas, celui qui nous intéresse ici, un numéro vous est attribué à un moment donné (ce peut être dès la page d'accueil, mais ce n'est pas obligatoire). Ce numéro, unique, sera représentatif de votre visite et sera répété dans les urls de chacune des pages que vous affichez sur votre navigateur lors de votre visite. Une fois celle-ci terminée, l'identifiant est abandonné. La même page aura donc un identifiant différent, donc une url différente, si vous revenez la voir, par exemple, le lendemain... Cet identifiant est donc attribué à UN internaute donné pour UNE visite donnée.

Le paramètre d'identifiant de session, présent dans l'adresse de la page, peut prendre plusieurs noms, comme "id", "session_id", "sessionid", etc. Voici quelques exemples de telles urls :

http://delhaizewineworld.belbone.be/.../dossiers/_fr/summary.asp?dosid=24&sessionID=1637237328&language=6

http://www.maporama.com/share/default. asp?language=fr&SESSIONID=125566878

http://achat.webguideauto.com/index.php3?session_id=484590mJhvygHtE9jUHqQA256GdsiyVSGUTvxs54WSvEtvPJPARpcC0

http://www.luminus.be/Algemeen/FR/ FR?$SESSIONID$=-6471781313463408975

Etc.

Ce paramètre est redoutable pour les moteurs, car cela signifie qu'un numéro de session est indiqué dans l'url pour chaque visite, donc pour chaque prise en compte par ses robots. Si Googlebot (le robot de Google) vient chaque jour indexer une même page, un identifiant de session lui sera attribué pour chaque visite, donc une page identique aura, chaque jour, une adresse différente...

On comprend que cela pose quelques problèmes, voire quelques casse-tête sérieux, aux moteurs qui préfèrent, la plupart du temps, ignorer totalement ces pages s'ils repèrent dans leur adresse la mention "sessionid" ou un terme approchant, bref s'il y détectent un identifiant de session.

On trouve cependant, dans les index des moteurs, quelques-unes de ces pages. Tapez des requêtes comme "inurl:session_id" ou "inurl:sessionid" sur Google et il vous renverra quelques milliers, voire dizaines de milliers de pages. Ceci dit, il est clair que l'identifiant de session est un problème assez important et bloquant pour les moteurs, certainement l'un des plus bloquants à l'heure actuelle.

Certains sites contournent le problème, cependant, en adoptant en majorité trois stratégies différentes qui peuvent s'avérer complémentaires :

- Le fait d'appliquer un numéro de session le plus tard possible dans la navigation (donc en évitant ce type de système sur la page d'accueil, la page de présentation des produits et en ne l'appliquant - par exemple - qu'à partir du moment où une réelle vente est en cours).

- Le fait de plutôt utiliser les cookies, qui permettent également ce type d'action et posent moins de problèmes aux moteurs. Mais cela pose, ou peut l'imaginer, de nombreux problèmes techniques si le site a été réalisé, au départ, en tenant compte des identifiant de session... Il n'est pas toujours simple de revenir en arrière sur ce point.

- Le passage à un système d'"url rewriting" qui peut, dans certains cas, résoudre quelques problèmes (voir la troisième partie de cet article).

Conclusion

On peut dire que, depuis un à deux ans, les moteurs de recherche ont grandement amélioré la prise en compte des pages dynamiques. Cependant, celles-ci restent encore un véritable facteur bloquant dans certains cas (identifiant de sessions, sites en technologies propriétaires, nombre de paramètres trop important dans l'url, etc.).

Aussi, il est souvent nécessaire de passer par des stratégies spécifiques pour bien référencer ces sites. Dans la deuxième partie de cet article, nous verrons les différentes solutions proposées par les sociétés de référencement actuelles. Dans la troisième partie, nous ferons un "focus" sur les techniques d'url rewriting, qui résolvent bon nombre de problèmes. Enfin, sachez que nous lançons d'ores et déjà des pistes auprès des moteurs de recherche pour obtenir des réponses les plus précises possibles à ce sujet.

Au mois prochain...