PAR FRANÇOIS BOURDONCLE (84)

Après une carrière de recherche aux États-Unis, il a fondé en 2000 Exalead, moteur de recherche pour les entreprises, mais aussi pour le Web avec www.exalead.com, moteur de recherche mondial qui compte huit milliards de pages référencées

Les **moteurs** de recherche, acteurs **stratégiques**

Devenus un outil aussi indispensable que le téléphone, les moteurs de recherche présentent bien des pièges : risques de manipulation, espionnage de brevets, atteinte à la vie privée. Au service de la recherche, les moteurs de demain mettront en contact les internautes qui partagent les mêmes centres d'intérêt.

Les moteurs de recherche sont stratégiques à plus d'un titre. Ils sont devenus pour beaucoup un outil aussi indispensable dans la vie quotidienne que le téléphone. Il est difficile d'imaginer comment l'on faisait « avant », tout comme il est difficile d'imaginer comment l'on faisait « avant » l'avènement de la téléphonie mobile. Que ce soit pour faire ses achats en ligne, organiser ses voyages, faire ses réservations, rechercher de l'information à usage professionnel, ou organiser une activité de veille, les moteurs de recherche sont le point de passage obligé.

Le modèle d'affaires de l'économie numérique

Mais, l'autre raison qui fait des moteurs de recherche des acteurs stratégiques du monde Internet c'est qu'ils sont les modèles d'affaires de toute l'économie numérique. Les régies publicitaires des géants Google, Yahoo, et plus récemment, Microsoft, sont en effet indispensables pour « monétiser » les services Internet, c'est-à-dire permettre à ces services de gagner de l'argent grâce à la publicité, exactement comme la télévision privée se finance par la publicité.

Le paiement à la performance voit les annonceurs se bousculer pour surenchérir pour l'achat de mots clefs pour afficher leur publicité de manière contextuelle. Par exemple, Renault va vouloir à tout prix acheter le mot « voiture » pour que chaque fois qu'un utilisateur fait une recherche du genre « voiture rouge », une publicité pour les voitures Renault s'affiche.

Un moteur populaire vendra donc plus cher le clic sur chacun de ses liens commerciaux, et il sera affiché (et donc cliqué) plus souvent. De plus, un plus grand nombre d'annonceurs implique un plus grand nombre de mots différents achetés, et donc un pourcentage plus important des recherches qui donnent lieu à l'affichage de liens commerciaux (on appelle cela le « taux de couverture » de la régie). Enfin, le nombre d'annonceurs est aussi directement lié à la pertinence des liens commerciaux, ce qui augmente la probabilité qu'un utilisateur clique sur le lien (c'est le « taux de clics »), ce qui génère là aussi plus de revenus.

Le revenu d'un moteur de recherche est donc fonction du produit de son trafic, du prix moyen du lien commercial, du taux de couverture, et du taux de clics des utilisateurs sur les liens, ce qui fait que la rentabilité d'une bonne régie peut facilement être plus de trois à cinq fois supérieure à celle d'une régie médiocre.

REPÈRES

La performance des régies, qui commercialisent à la fois des bannières (publicité sous forme graphique) et des liens commerciaux (qui apparaissent au-dessus, et parfois à droite des résultats dits « organiques »), est étroitement liée à la qualité et au trafic du moteur associé. Les liens commerciaux sont, d'une part, achetés aux enchères par les annonceurs, et d'autre part ne rapportent de l'argent aux sites qui affichent la publicité que si ces liens sont cliqués (c'est ce qu'on appelle le « paiement à la performance »).

Des risques de manipulation

Dans le contexte de la veille, les moteurs de recherche présentent bien des pièges dont peu de professionnels semblent conscients. Au-delà du débat sur l'exhaustivité de l'indexation des moteurs de recherche et sur la taille du Web visible ou caché (voir plus loin), il n'est pas illégitime de se demander si les moteurs de recherche indexent l'information de manière totalement neutre, et si certaines informations sensibles y sont ou non référencées. De plus, le classement des résultats peut être manipulé de plusieurs manières, et des sociétés spécialisées, ou des particuliers particulièrement doués, ont ainsi réussi, pendant un certain temps, à faire apparaître le site officiel du président George W. Bush en tête des résultats du moteur Google sur la requête « miserable failure ».

Cette histoire, qui est une illustration de ce que l'on appelle le « Google bombing », a fait le tour de l'Internet, mais le classement des résultats tient compte d'un nombre de paramètres tellement important (popularité du site, texte de la page, texte du titre, texte des liens pointant sur la page, graphe des liens hypertextes, descripteurs sémantiques, etc.), que le résultat de la formule est difficile à prévoir, et sa manipulation éventuelle quasi impossible à prouver.

Brevets et vie privée

Enfin, un dernier point important dans un contexte de veille mais aussi de respect de la vie privée est la traçabilité de plus en plus grande de l'ensemble des activités en ligne : chaque recherche effectuée sur un moteur de recherche est archivée avec l'adresse IP de l'ordinateur d'où est issue la recherche. Si cet ordinateur est le pare-feu d'une grande entreprise, alors il est possible de savoir qu'un salarié de cette entreprise a tel ou tel centre d'intérêt, ce qui peut être grave si l'entreprise en question est en train de déposer des brevets sur ce sujet.

Au niveau de la vie privée, ce qui est préoccupant, c'est le croisement des bases de données contenant des informations personnelles, et la facilité qu'il y a à faire des croisements entre ces bases de données quand celles-ci appartiennent à la même société (recherche d'informations, courrier électronique, paiement en ligne, blog, etc.), car l'utilisateur a le même identifiant, le même mot de passe, et la société dispose peut-être même de son identité réelle s'il a donné à un moment donné son numéro de carte bancaire pour le paiement d'une transaction. Il est donc important, pour le professionnel comme pour le particulier, de ne pas mettre tous ses œufs dans le même panier et d'utiliser des services de plusieurs fournisseurs différents. Il est également légitime pour l'Europe de se poser la question de son indépendance stratégique en matière d'accès à l'information, comme elle le fait par exemple pour le GPS avec Galileo.

Un principe commun

Tous les moteurs de recherche fonctionnent aujourd'hui fondamentalement sur le même principe. Dans une première phase, ils recensent toutes les pages Web auxquelles ils ont

La prime à la taille

La prime à la plus grosse régie (Google) est considérable et tend mécaniquement à renforcer encore plus son avantage concurrentiel, car les sites et les annonceurs vont naturellement préférer travailler avec elle. La tâche est donc très dure pour le numéro deux (Yahoo), et *a fortiori*, pour le numéro trois (Microsoft).
L'ironie de l'histoire est que le brevet sur les liens commerciaux est détenu par Yahoo qui a accordé une licence pour quelques centaines de millions de dollars seulement juste avant l'entrée en Bourse de Google, quand cette société n'était pas encore aussi dominante qu'elle ne l'est aujourd'hui. Mais Google ayant un moteur plus performant que celui de Yahoo, sa régie publicitaire s'est renforcée au point de devenir incontournable. De ce point de vue, le rachat de Yahoo par Microsoft, s'il était confirmé, serait une excellente nouvelle pour tous les acteurs de l'économie numérique car il permettrait à ces derniers de mettre en concurrence, pour monétiser leurs services, deux régies publicitaires aux performances comparables.
La position dominante de la régie de Google est également l'une des raisons pour lesquelles le Département de la justice américain enquête actuellement sur l'utilisation de la régie publicitaire de Google par Yahoo, en complément de sa propre régie publicitaire, afin d'améliorer la monétisation de son moteur de recherche.

Les moteurs de recherche indexent-ils l'information de manière totalement neutre ?

De fil en aiguille

De nombreuses améliorations des moteurs de recherche sont possibles, notamment lorsqu'on prend en compte la composante humaine de l'activité de recherche, qui est essentielle, notamment dans un processus de veille. Certains moteurs comme Exalead proposent également de naviguer dans les résultats grâce à une technique brevetée appelée « recherche par sérendipité » qui, à l'aide d'une sorte de table des matières contextuelle, permet de rechercher un peu comme on lit un dictionnaire ou une encyclopédie, en commençant par un mot, et de fil en aiguille, en trouvant le mot ou le concept le plus intéressant que l'on n'avait pas forcément présent à l'esprit en commençant la lecture.

accès, en démarrant par la page d'accueil des plus gros sites existants (des portails comme Yahoo par exemple), et en suivant les liens hypertextes qui apparaissent dans les pages rencontrées. Ils mettent ces liens dans une liste d'attente (les nouveaux liens à la fin de la liste) et parcourent le Web un peu à la manière de l'onde qui se propage à la surface d'un lac quand on laisse tomber une goutte d'eau en son centre. On appelle cela le parcours « en largeur d'abord » du graphe constitué par les liens hypertextes. Ce parcours essaie d'éviter de recenser trop de pages d'un site au détriment des autres. Le système s'arrête soit quand les pages n'existent pas, sont protégées, par exemple, par des mots de passe, soit encore si elles ne sont pas accessibles en suivant des liens, ce qui est par exemple le cas des pages qui sont stockées dans une base de données accessible par un formulaire que le moteur ne peut pas remplir seul sans assistance humaine (c'est ce que l'on appelle parfois le Web invisible, ou le Web caché).

Le parcours « en largeur d'abord » évite de recenser trop de pages d'un site au détriment des autres

Une fois les pages recensées et numérotées, elles sont stockées, puis indexées de manière à associer à chaque mot la liste ordonnée des pages où ce mot apparaît, et les positions de ce mot sur chacune des pages. Quand un utilisateur fait une recherche à plusieurs mots, par exemple « vache folle », le moteur met en correspondance les deux listes des occurrences de vache et de folle et cherche tous les documents dans lesquels vache et folle apparaissent, et où, de plus, si vache apparaît à la position *n* dans un document, alors folle apparaît à la position *n* + 1 dans le même document. Cet algorithme est linéaire dans la taille des deux listes.

Une composante humaine

Mais le futur des moteurs de recherche est sans doute ailleurs encore. En effet, il y a fondamentalement trois moyens de chercher sur Internet : sa mémoire (par exemple, les favoris de son navigateur), les moteurs de recherche (qu'ils soient généralistes ou spécialisés, comme ceux des sites de commerce électronique), et enfin, il y a des amis à qui l'on peut demander conseil. Or les moteurs de recherche ne prennent pas du tout en compte cette troisième composante, humaine, de l'activité de recherche, qui est pourtant essentielle, notamment dans un processus de veille. Le service Baagz est le premier moteur de recherche permettant à l'utilisateur d'entrer automatiquement en contact avec d'autres utilisateurs partageant les mêmes centres d'intérêt que lui et qui seront les mieux à même de l'aider à trouver des réponses à des questions complexes qu'il est difficile de poser à un moteur de recherche traditionnel.

Des dossiers intelligents

La manière de fonctionner d'un système comme Baagz consiste à permettre à l'utilisateur de créer et d'organiser ses favoris dans des « sacs » et de partager s'il le souhaite certains de ses sacs avec d'autres utilisateurs. Ces sacs sont en réalité des dossiers intelligents qui utilisent la description sémantique que le moteur Exalead associe à chaque site Internet pour comprendre les centres d'intérêt de l'utilisateur et associer automatiquement les sacs de ce dernier aux communautés qui sont les plus à même de l'intéresser et de l'aider dans ses recherches. ■

BIBLIOGRAPHIE

« Recherche d'aiguilles dans une botte de liens », François Bourdoncle et Patrice Bertin, *La Recherche*, 328 (février 2000), page 66.

EN SAVOIR PLUS SUR INTERNET

www.exalead.com
www.baagz.com