"Site Search" (1ère partie) : panorama du marché

Retour au sommaire de la lettre

Une multitude d'acteurs offrent aujourd'hui des solutions pour indexer les contenus des sites Web et déployer des fonctionnalités de recherche internes. Nous vous proposons ici un panorama du marché de ces solutions de recherche intra sites. Cet article sera suivi, les prochains mois, d'une présentation plus approfondie des offres des acteurs, ainsi que d'une présentation des principales questions qui doivent être posées en amont du choix des solutions.

Les moteurs Web majeurs comme Yahoo, MSN ou Google n'ont pas le monopole de la recherche, on l'oublie trop souvent ! Pour trouver l'information dont ils ont besoin, de nombreux internautes effectuent en effet leurs recherches en utilisant directement les moteurs "internes" de sites qu'ils connaissent déjà.

Les éditeurs, webmasters et propriétaires de sites l'ont bien compris et ils sont souvent très attentifs au choix des moteurs de recherche qu'ils proposent à leurs lecteurs pour effectuer des recherches dans leurs pages. Chaque type de recherche doit en effet répondre à des objectifs bien précis en termes de communication et de marketing.

Outre le fait qu'ils représentent un service à valeur ajoutée apprécié par les lecteurs, des moteurs intra-sites performants peuvent éviter la fuite de certains utilisateurs insatisfaits vers de grands portails de recherche. Lorsque nous jugeons les cartouches de recherche de certains sites Web inefficaces, ne sommes nous pas nombreux à leur préférer aujourd'hui les fonctionnalités avancées de "Site Search" (fonctionnalités qui permettent de restreindre une recherche à un site donné, comme l'option "site:") des grands moteurs ?

Par ailleurs, les technologies de recherche sur sites les plus évoluées sont désormais souvent enrichies de fonctionnalités d'indexation et de catégorisation permettant d'associer à chaque résultat d'autres liens pertinents. *"Notre fonctionnalité 'Smartlinks' permet de générer automatiquement des liens en rapport avec une requête, ce qui offre à l'utilisateur la possibilité d'affiner sa recherche. Pour le propriétaire du site, cela augmente le nombre de pages vues"*, s'enthousiasme ainsi Martin Grosjean, co-fondateur de Synomia.

Estimation du marché du Site Search

Contrairement au marché de la recherche d'entreprise, il existe peu d'études sur le marché du "Site Search". Cette absence de documentation sur le sujet s'explique par deux principaux facteurs : la relative "jeunesse" des fournisseurs de solutions – les principaux acteurs sont entrés sur le marché en 1998 – et le fait que les analystes englobent le plus souvent le marché de la recherche "intra-site" dans un marché plus large comprenant toutes les solutions de gestion de l'information d'entreprise.

Une chose est sûre, les acteurs sont très nombreux sur ce marché... Les deux "poids lourds" du secteur sont le norvégien Fast et le duo américain Autonomy / Verity (Autonomy ayant racheté Verity en novembre 2005). Dans l'Hexagone, s'affrontent également sur ce terrain un certain nombre de spécialistes, parmi lesquels Antidot, Exalead, Synomia, Sinequa et Go Albert…

Typologie des solutions proposées

La plupart des acteurs proposent aujourd'hui des solutions en marque blanche. Les sites peuvent donc s'approprier ces outils et apposer leurs propres marques sur les pages de résultats…

Solutions hébergées chez le client

Les solutions hébergées chez le client sont les plus répandues. Elles nécessitent l'installation d'un logiciel dédié sur un serveur (en interne ou chez un hébergeur).

A noter, toutefois, que certains acteurs, comme Fast, ont la capacité de proposer des solutions "mixtes", en offrant le choix entre un hébergement interne ou une solution en ASP, suivant les besoins du client.

Fast (http://www.fastsearch.com/) offre une solution (*Enterprise Search Platform*) répondant aux besoins de trois types de clients : plateformes d'e-commerce, sites "classiques" (presse ou information) et répertoires (types annuaires). Nécessitant l'acquisition d'une licence, cette application peut être déployée chez le client ou *"hébergée sur le serveur de Fast"*, explique Jay J.M M'Bei, directeur commercial de Fast Search&Transfer France. Il ajoute que le choix est *"fonction des infrastructures hardware et des équipes dédiées à la disposition du client"*. En France, sa solution a, entre autres, été déployée sur le site Internet de l'hebdomadaire Stratégies (http://www.strategies.fr/).

Verity (http://www.verity.com/fr/) offre une solution similaire. Elle est utilisée, par exemple, pour la version électronique du quotidien Les Echos.

Le Français Exalead propose quant à lui une solution dénommée "exalead one:enterprise" (http://corporate.exalead.com/enterprise/l=fr?p=solutions_sites-internet_index). La Société Générale l'a notamment déployée pour indexer les contenus d'une centaine de ses sites Web (http://www.socgen.com/).

Sur ce marché, Exalead est en concurrence avec deux autres acteurs nationaux : Sinequa et Antidot. Le premier commercialise deux applications "Intuition Web Content Edition" (http://www.sinequa.com/html-fr/fr-edition.web.html) et " Intuition Press Edition" (http://www.sinequa.com/html-fr/fr-edition.press.html). Le quotidien Le Monde ou encore le Sénat l'ont déployée. Le second distribue sa solution AFS@web (pour *Antidot Finder Suite* - http://www.antidot.net/) auprès d'acteurs comme la chaîne de télévision TF1 ou encore Skyrock.

Enfin, d'autres entreprises, se sont spécialisées sur des marchés de niche. L'éditeur israélien Celebros (http://www.celebros.fr/) n'adresse que les sites marchands. Le franco-suisse Go Albert (http://www.albert.com/) permet quant à lui d'associer une technologie de recherche interne à un outil de veille automatisé sur le Web (AMI Website Access). A voir également, l'outil PG-Recherche Pro (http://www.perl-pro.com/moteur_recherche/index.html), de la société nantaise Raynette (http://www.raynette.fr/).

Google décline le principe de l'appliance sur le marché de la recherche sur sites

Depuis 2004, Google propose aux entreprises une "Google Search Appliance", c'est-à-dire un boîtier packagé - associant hardware et soft - pour mettre en place une technologie de recherche sur réseau intranet ou intra-site. Plusieurs modèles sont proposés en fonction de la taille des réseaux à prendre en compte (sachant qu'il est possible d'empiler les "racks", si besoin).

Le produit est disponible en trois modèles :

- le GB-1001, pour les départements d'entreprises et les sociétés de taille moyenne.
- le GB-5005, développé pour les services de recherche dédiés tels que les sites Web des entreprises qui communiquent principalement avec leurs clients par leur intermédiaire et les applications intranet des grandes entreprises.
- le GB-8008, spécialement conçu pour les déploiements centralisés au sein de plusieurs unités organisationnelles.

Avec cette solution, le moteur reproduit un modèle adopté par la plupart des éditeurs du monde de la sécurité informatique, mettant en avant le côté "pluggable" de sa solution, qui ne nécessite que peu d'intégration tandis que les déploiements sont généralement longs et complexes. Autre avantage : contrairement aux solutions hébergées, ce système est installé dans l'entreprise, donnant l'impression aux DSI "craintifs" qu'ils gardent la main sur la solution.

Le "package", qui comprend matériel, logiciel et deux ans de support, coûte de 32 000 US$ à 175 000 US$.

Déjà perçu comme ayant un peu "cassé les prix" sur ce marché, le moteur offre aussi désormais une solution d'entrée de gamme (Google Mini) qui permet d'indexer jusqu'à 50 000 documents. En novembre 2005, cette version "light" était facturée 3 000 dollars l'unité. Depuis le début du mois de mars, elle est proposée à 1 995 dollars.

A noter, toutefois, que cette solution est souvent critiquée pour deux raisons principales : ses fonctionnalités de personnalisation sont jugées limitées, tout d'abord. Le système de PageRank, qui a fait le succès de Google sur le marché grand public et qui consiste à privilégier dans les résultats les documents vers lesquels pointent le plus de liens "de qualité", n'est pas nécessairement pertinent lorsqu'il est appliqué sur un site d'entreprise.

http://www.google.com/appliance/

Solutions en mode ASP

Avec les solutions en mode ASP, l'indexation du site est opérée par un moteur distant et les requêtes sont effectuées sur la base "miroir" de ce moteur. Les résultats sont restitués sous la forme d'une page dont l'apparence peut, dans la plupart des cas, être adaptée à une charte graphique. On estime que le coût des premières offres en ASP démarre à environ 300 € par mois.

D'origine française mais bien implanté au niveau international, le Français Synomia (http://www.synomia.fr/) est l'un des acteurs les plus représentatifs sur ce segment de l'ASP. Il revendique environ 300 clients, parmi lesquels le quotidien Libération ou le magazine L'Express, la ville d'Oakland (http://www.oaklandnet.com/) et … le site Abondance ! Issu d'un projet du CNRS, la force de sa technologie réside, entre autres, dans ses capacités d'analyse linguistique et syntaxique, ainsi que dans ses fonctionnalités "Smartlinks" permettant de générer automatiquement des liens en rapport avec les contenus visualisés.

L'Américain WebSideStory (qui a racheté Atomz en février 2005) offre une autre solution, entièrement hébergée. Elle porte toujours le nom d'Atomz (http://www.atomz.com/).

De nombreux moteurs de recherche proposent quant à eux la possibilité d'intégrer facilement un cartouche de recherche sur votre site. La recherche s'effectue alors strictement dans les pages de votre site qui sont déjà présentes dans leurs index respectifs. C'est le cas de KartOO Site Box (http://www.kartoo.net/e/fra/sitebox.html) ou de Google. Ce dernier offre un outil gratuit via son programme de liens sponsorisés AdSense : AdSense pour les recherches (https://www.google.com/adsense/ws-overview?hl=fr).

Solutions open source

"Les solutions open source ne coûtent rien en frais de licence. En revanche, leur intégration peut s'avérer relativement onéreuse. Les fonctionnalités offertes sont aussi assez rudimentaires. Mais ces solutions peuvent répondre aux besoins de sites dont le budget de développement est inférieur à 50 000 euros", nous expliquait récemment un représentant d'un éditeur.

Ces solutions sont nombreuses (voir la Lettre R&R de juillet 2005). Les plus répandues sont ht://Dig (http://www.htdig.org/), mnoGoSearch (http://mnogosearch.org/), Swish-e (http://swish-e.org/) et Lucene (http://lucene.apache.org/). Certains projets, comme Lucene, sont soutenus par d'importantes communautés de développeurs (Lucene est lié au projet Nutch).

Perspectives

Plusieurs grandes tendances d'évolution se dessinent aujourd'hui sur ce marché, avec en particulier une évolution vers des outils intégrant des technologies syntaxiques, sémantiques et statistiques.

La catégorisation et la "clusterisation" sont un autre axe d'évolution. *"Je pense que nous allons vers un maillage de l'information plus important*, estime Christelle Ott, directrice marketing d'Antidot. Selon elle, ce maillage consiste à *ne plus simplement présenter LA bonne réponse à la recherche de l'internaute, mais à présenter un maillage de l'information permettant d'organiser les contenus connexes à la recherche de l'internaute et ce de manière totalement automatique.*

De nombreux acteurs comme Fast, Sinequa, Exalead ou Antidot ont également enrichi leurs solutions de recherche Web de nouvelles fonctionnalités permettant d'adjoindre à l'indexation des données non structurées des sites Web un nouveau type de recherche dans les données structurées des bases de données relationnelles. Cette démarche les place en concurrence directe avec de grands éditeurs d'outils de GED et de bases de données comme IBM, Oracle ou OpenText.

Notons cependant qu'à l'instar d'IBM, les grands acteurs de la GED font eux le chemin inverse et investissent de plus en plus massivement les outils de recherche sur les données Web (non structurées). Le paysage concurrentiel risque donc d'évoluer très prochainement…

Trois questions à Christelle Ott
(directrice marketing d'Antidot)

Pouvez vous nous décrire Antidot ?
Antidot a été créé en 1999. Nous concevons des solutions d'accès à l'information : moteurs de recherche, solutions de veille, catégorisation et consolidation de données multi-sources. Nos solutions sont fournies en mode ASP ou en licence.
L'originalité de la solution Antidot réside dans son architecture distribuée autorisant une évolutivité maximale. Elle se distingue également par les fonctionnalités de recherche évoluées qui vont de la suggestion orthographique, à la suggestion de recherche, au filtrage dynamique des réponses, au maillage fin de l'information.
Antidot dispose d'un programme de R&D très important associé à des laboratoires privés et publics, dont les projets sont validés par l'ANR (Agence Nationale de Recherche).

Quels sont selon vous les principaux besoins des sites de e-commerce ?
Les sites de ce secteur ont besoin d'indexation à fréquence élevée (voir très élevée et très proche du temps réel) sur des bases de données contenant de quelques milliers à plusieurs millions de références. Ils nécessitent également des outils d'aide à la vente : le moteur de recherche est utilisé par plus de 50% des internautes comme moyen de navigation principal. C'est un formidable outil de récolte de l'expression du client qui va s'exprimer "naturellement". L'étude qualitative des requêtes est donc primordiale pour d'une part mieux connaître sa cible, et mieux lui répondre, voire anticiper ses besoins. Ils ont également besoin d'influer le flux de réponse, soit par la voix éditoriale dans le cadre de l'animation du site, soit de manière automatique pour la mise en avant.

Quelles sont les principales préoccupations des sites de contenus ?
Les sites de contenus (presse, médias…) ont avant tout besoin de permettre un accès le plus profond possible aux articles afin de rentabiliser la production de contenu. Actuellement la différence principale d'une solution à l'autre réside dans sa capacité à indexer de très gros volumes de documents et à répondre potentiellement à une charge très importante.