Espotting, un nouvel acteur majeur du positionnement publicitaire

Espotting, l'un des acteurs majeurs du positionnement publicitaire au coût par clic (CPC, à opposer au CPM, coût par mille) en Grande-Bretagne (l'autre acteur étant GoTo, bien sûr), débarque en France avant la fin de l'année, normalement début décembre, avec dans ses valises bon nombre de contrats signés avec des sites portails et des outils de recherche francophones. Le CPC (coût au clic) consiste à faire payer les annonceurs pour des liens présentés sur les pages de résultats des outils de recherche, mais uniquement si l'internaute clique sur ce lien. GoTo et Espotting sont aujourd'hui les deux sociétés les mieux implantées en Europe sur ce marché. On l'oppose généralement au CPM (les offres de positionnement publicitaire de Voila et de Nomade.fr, décrites dans notre lettre du mois dernier, sont, par exemple, basées sur le CPM), pour lequel l'annonceur paye le nombre de fois où le lien sera affiché, qu'il soit cliqué ou non.

Encore peu connu en France, Espotting a déjà bon nombre de clients en Grande-Bretagne, comme :

- Lycos et Hotbot UK : 5 résultats payants affichés (3 de côté, sur la droite ou la gauche selon le site, et deux en bas de page dans chaque cas).
- Looksmart UK : 3 premiers résultats affichés.
- Searchengine.com : moteur de recherche par défaut.
- ASK Jeeves UK : 10 résultats payants affichés en bas de page et 3 résultats payants affichés spécifiquement pour les résultats "shopping".
- Dogpile (métamoteur) : 10 résultats payants affichés.
- Kelkoo au niveau européen (en cours).

Plus d'autres portails et outils de recherche, moins connus en France : UK Plus (5 résultats), EasyEverything Cafes, NTL, BTOpenworld, The Sun, UKOnline, ClaraNet Default, FreeUK, Searchgate Web Search, Netimperative.com, Peoplebank.com, Micromart, Musicmart, Justbooks.co.uk, Virginstudent.co.uk, Peoplenews.com, etc.

L'offre peut prendre plusieurs formes : proposition des premiers liens uniquement (les 3 ou 5 premiers, par exemple) jusqu'à l'affichage de tous les résultats proposés.

Espotting prend en compte aujourd'hui 100 millions de recherches par mois et a un taux de pénétration de 66% outre-Manche. Il gère 4 000 annonceurs (contre 1 500 à GoTo UK, à ce qu'il semble) pour 1 000 sites affiliés.

La société a été lancée en septembre 2000 et emploie aujourd'hui 60 personnes. Son expansion se limite, pour l'instant, à l'Europe. Après la Grande-Bretagne, Espotting va donc s'implanter en France et en Allemagne et mettre en ligne ses premières prestations début décembre 2001.

Affichage des liens payés
~~~~~~~~~~~~~~~~~~~~

Ses premiers clients dans l'Hexagone seront, sans surprise, Lycos et HotBot. Les sites français (www.lycos.fr et www.hotbot.fr) verront donc des liens achetés fleurir sur leurs pages de résultats, exactement sous la même forme que sur les sites anglais (www.lycos.co.uk et www.hotbot.co.uk). Pour avoir une idée de ce que celà peut donner, tentez la requêtes "cars" sur Lycos UK :
http://search.lycos.co.uk/cgi-bin/pursuit?matchmode=and&mtemp=main&etemp=error& query=cars&cat=brit

et sur HotBot UK :
http://www.hotbot.lycos.co.uk/result.html?query=cars&bool=all &z=1111121111111111211&languagefield=any&numresult_field =10&description_field=full&uk=ww&hs=s

Dans les deux cas, l'information est affichée en deux "pavés" : le premier contient trois liens (les trois classés premiers aux enchères, voir plus loin) et est appelé "Sponsored Links". Il est affiché à droite de l'écran sur Lycos et à gauche sur Hotbot.
Le deuxième, proposant les deux liens situés en 4ème et 5ème position aux enchères au moment de l'interrogation, sont situés en bas d'écran, sous l'appellation "Sponsored Links to cars:". Les liens affichés sont rigoureusement identiques sur les deux sites. Il en sera de même sur Lycos France et HotBot France à partir du 1er décembre prochain, donc.

Mais ces deux outils de recherche ne seront pas les seuls à être pris en compte par l'offre Espotting. Looksmart France prendra également en compte ces résultats. Sur le portail français de Netscape (http://www.netscape.fr), 75% de la recherche par défaut en seront également issus, ainsi que les trois premiers résultats de la zone "recherche" du site (http://home.netscape.com/fr/escapes/search/netsearch_0.html). Et d'autres partenariats sont en train d'être signés.

Mode d'achat des mots clés aux enchères
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Si vous désirez être annonceur sur Espotting et, donc acheter un mot clé ou une expression donnée, il suffit d'aller sur le site français de la société (http://www.espotting.fr) et d'ouvrir un compte (pris en compte en 48 heures généralement) en déposant au départ 100 euros HT minimum, tarif qui reste accessible à toute taille d'entreprise. Vous pourrez, une fois cette phase administrative réalisée, gérer vos campagnes en ligne, donner un coût au clic pour un mot clé, regarder les enchères monter ou descendre pour vos mots clés favoris, et également (ce que GoTo ne propose pas), avoir la possibilité d'insérer un logo dans les résultats des moteurs de recherche. Pour avoir une idée, tapez le mot clé "voiture" sur le site d'Espotting France, vous verrez apparaître un lien qui propose ce type de mise en exergue graphique. Les résultats sur le site français de la société sont sinon, fournis par Inktomi. Pour information, selon Inktomi, 67% des utilisateurs cliquent sur les 5 premiers liens de la page de résultats.

Le système est, sinon, "classique" dans le sens où si vous avez déjà fait une enchère sur GoTo, vous ne serez pas dépaysé sur Espotting. Mais ce dernier propose quelques fonctionnalités intéressantes, comme l'offre "Premier tout le temps", qui surenchérit automatiquement aux enchères effectuées par des concurrents, pour que vous puissiez être tout le temps en première position. Dangereux (par les dérives possibles) mais intéressant.

Espotting propose également un outil de reporting assez complet (clics, prix, heure, etc.) pour gérer ses campagnes au jour le jour.

A priori, le système semble bien marcher. Pour information, les plus grosses enchères actuelles, notamment dans le domaine de la finance, en Grande-Bretagne, tournent autour de une livre à une livre et demi le clic, mais la moyenne globale est bien en deça de ce chiffre.

Implantation en France
~~~~~~~~~~~~~~~~~~~~

Pour son implantation en France, Espotting a, bien sûr, rencontré les principales agences publicitaires de la place et les acteurs majeurs de la recherche d'information dans l'Hexagone. L'objectif est d'avoir 500 annonceurs dans les trois premiers mois pour atteindre rapidement des parts de marché aussi fortes qu'en Grande-Bretagne. Le lancement des offres en France se fera en même temps qu'en Allemagne.

Liens
~~~~~~

Le site anglais d'Espotting :
http://www.espotting.com/

Le site français :
http://www.espotting.fr/

La FAQ française sur l'offre Espotting :
http://www.espotting.fr/advertisers/ad_faq.asp

Interview de M. Sébastien Bishop, Directeur d'Espotting
~~~~~~

- Pouvez-vous vous présenter en quelques mots, ainsi que la genèse de la société ESpotting ?

Je m'appelle Sébastien Bishop et je suis co-fondateur de la société Espotting. Nous avons commencé à deux personnes en février 2000 pour être 60 aujourd'hui et nous installons des bureaux en Allemagne et en France. Nos clients annonceurs au CPC ont aujourd'hui tous les profils car le coût d'entrée de la solution, 100 euros, est faible et permet à tous les types de structures, depuis le commerçant indépendant jusqu'au grand compte, de prendre en considération nos offres et de "marketer" ses produits à faible coût. Parmi nos annonceurs les plus connus, on peut cependant citer Orange, Vizzavi, etc.

- Quels sont les avantages et les inconvénients du CPC par rapport au CPM ?

L'avantage principal est que l'on paye, dans le cas du CPC, uniquement si une personne entre dans votre magasin, par rapport au CPM où l'on paye plutôt pour être vu. D'autre part, le trafic généré est plus ciblé. En effet, d'après les études que nous avons menées, sur le trafic généré par Espotting sur les sites de nos annonceurs, il apparaît que 14 à 16% des visiteurs achètent un produit, ce qui est un taux de conversion énorme.
De plus, pour l'annonceur, nous avons bréveté un système qui permet de rajouter un logo à côté du lien acheté. Une PME gagne ainsi autant de visibilté qu'un grand groupe, ce qui est souvent très apprécié.
Comme l'annonceur peut associer une url interne de son site à un mot clé acheté, il permet ainsi à l'internaute d'aller directement sur la page présentant un produit donné, et non pas sur la page d'accueil du site. Celà crée une différence non négligeable dans la facilité d'accès d'un internaute à un produit donné.
Du côté des risques, nous avons mis en place des solutions techniques pour qu'une personne ne puisse pas cliquer indéfiniment sur un lien et tenter de frauder ainsi pour générer un trafic artificiel. Ce type de pratique, que nous avons connues au lancement de l'offre en Grande-Bretagne, a été rapidement détecté et combattu.

- Quels sont les pays les plus "chauds", les plus "prêts" pour ce type d'offre ? Pourquoi avoir commencé en Grande-Bretagne ?

Nous avons commencé en Grande-Bretagne car, avec mon ami co-fondateur, nous avons vécu à Londres et qu'il nous a semblé que le marché était prêt ici. Mais le marché allemand est également gigantesque. D'autre part, il nous semble qu'il est important de se positionner en France, puisqu'aucune structure de ce type ne s'y trouve encore. C'est pour celà que nous créons des filiales à Hambourg et à Paris. Mais d'autres régions, comme les pays nordiques, nous intéressent, même si le marché potentiel y est moins important.

- D'après votre expérience en Grande-Bretagne, le public réagit-il bien aux "liens achetés" ?

A priori, oui, car les résultats issus de la base de données d'Espotting sont toujours indiqués comme étant des "sponsored links" et il n'y a jamais confusion entre liens achetés et liens issus d'un index "classique". Nous refuserions, d'ailleurs, de signer un contrat avec un outil de recherche ou un site web qui n'identifierait pas clairement nos résultats comme se rapprochant d'une publicité en "noyant" nos résultats parmi d'autres issus d'une base de données de type index de moteur ou d'annuaire.
Mais certains utilisateurs trouvent également, assez paradoxalement, que la pertinence et la précision des résultats est au rendez-vous. En effet, nous vérifions toujours de façon humaine que les mots clés achetés par un annonceur sont en corrélation avec son site. Une entreprise ne peut pas acheter comme mot clé le nom de son concurrent ou des termes ne correspondant pas à son activité. Du coup, certains internautes utilisent même le site d'Espotting pour effectuer leurs recherches sur le Web, car ils trouvent les résultats très pertinents. C'est pour nous la meilleure des récompenses, qui nous pousse à continuer dans cette voie.

- **Nouvelle présentation des résultats sur Yahoo.com et Yahoo.fr**

Le site américain de Yahoo! (http://www.yahoo.com) a inauguré il y a peu de temps une nouvelle version de sa page de résultats. Ces modifications arrivent également en Europe, comme nous allons le voir dans cet article.

Les sites de Yahoo! rénovent et relookent donc leur interface de visualisation des résultats. Cette décision a été prise suite à des études qui ont été menées, aux Etats-Unis et en Europe, auprès des utilisateurs de l'outil de recherche, identifiés selon plusieurs profils différents (internautes ayant l'habitude d'utiliser Yahoo!, utilisateurs débutants, etc.). On peut noter que les modifications portées à l'interface américaine sont différentes de celles qui vont être (ou sont déjà) mises en place en Europe et notamment en France. Enfin, si tous les sites européens ont la nécessité de présenter un visage assez similaire, ils ont cependant une latitude non négligeable pour afficher, selon chaque pays, plus ou moins d'informations (sur le fond et la forme) à l'utilisateur de l'annuaire. Chaque version régionale de Yahoo! peut ainsi personnaliser son interface, tout en restant quand même dans le "ton" européen pour garder une certaine homogénéité de l'outil d'un pays à l'autre.

Yahoo.com : une vision plus globale des résultats
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Si vous avez déjà visualisé la nouvelle présentation de la page de résultats du site américain de Yahoo!, vous n'avez certainement pas pu ne pas faire le rapprochement avec les pages de résultats de Google, notamment pour l'affichage des sites qui présentent plus qu'un air de famille avec ceux du moteur de recherche. Passons-les en revue...

Tout d'abord, les liens "Inside Yahoo! Matches", qui n'ont pas beaucoup changé par rapport à précédemment. Il s'agit de services issus de la "Galaxie Yahoo" et répondant à la demande. On peut estimer qu'il est normal que Yahoo! mette en avant ses propres sites et services...

Puis les catégories. Le grand credo de Yahoo! est d'aider les internautes à rechercher leurs informations dans les catégories plutôt que directement des sites, dans la liste proposée en bas de page. Aussi, les liens "catégories" (Category Matches) sont-ils plus mis en exergue dans cette nouvelle page de résultats. Changement important : Yahoo! ne propose plus une liste parfois interminable de rubriques pour des mots clés trop génériques (exemple : voiture, assurance, etc.), mais pas plus de 5 ou 6 catégories, sur une ou deux colonnes. Un lien "Next 20" permet alors d'aller sur une nouvelle page proposant toutes les autre scatégories. Ouf, rien que pour cela, on peut louer cette nouvelle présentation...

Grande nouveauté également : les catégories ne sont plus proposées sous une forme développée (exemple : Business and Economy > Shopping and Services > Automotive > Rentals > Budget Rent a Car Corporation) mais uniquement sous un intitulé raccourci, appelé "short names" (Auto Rentals > Budget Rent a Car Corporation) qui facilite grandement la lecture. Notons que ces intitulés raccourcis ont été rentrés à la main par les netsurfeurs de Yahoo! pour toutes les catégories de l'annuaire. Chaque rubrique est donc caractérisée à la fois par un intitulé long et par un "raccourci" destiné aux pages de résultats. Imaginez la somme de travail qu'il a fallu développer de la part des 150 netsurfeurs (environ) de Yahoo! pour effectuer ce travail ! Et encore, le projet au départ était plus ambitieux puisqu'il prévoyait une phrase explicative (par exemple "Comment louer une voiture chez Budget ?") pour chaque catégorie. Finalement, c'est la version courte des intitulés qui a été retenue, mais il n'est pas impossible qu'une version "langage naturel" ne soit pas mise en oeuvre par la suite.

Ceci dit, au delà de l'aspect racourci de l'intitulé d'une catégorie, le "short name" est surtout une forme de désignation plus conviviale, lisible et intuitive de la rubrique. Elle restitue le contexte et aide à une meilleure compréhension des catégories correspondantes à la requête de l'utilisateur.

Autre nouveauté : les résultats issus de l'arborescence "Exploration géographique" n'apparaissent plus dans les pages de résultats, sauf lorsque la demande est explicite (France, Paris, cars France, etc.). Un lien "List results by location" peut également apparaître en 6ème position pour avoir accès aux résultats géographiques s'ils existent.

Enfin, comme on l'a vu, les sites eux-mêmes sont présentés de façon très "Googlienne", même si les informations affichées restent inchangées par rapport à la version précédente, si ce n'est la catégorie à laquelle appartient le site qui est présentée en "version courte".

Point important : la forme beaucoup plus compacte de la première page des résultats de la recherche amène une meilleure visibilité des services Yahoo!, des catégories et des sites qui repondent le mieux à la requête de l'internaute. Du fait de la place plus réduite, ou en tout cas optimisée de l'affichage des catégories, les sites sélectionnés dans l'annuaire Yahoo! ont plus de visibilité dès la première page des résultats de la recherche. Pas négligeable...

Notons également que l'algorithme de pertinence est le même que précédemment. La nouvelle version de Yahoo! est basée sur un nouveau "look", mais les résultats proposés auraient été les mêmes il y a un mois de cela, lorsque l'ancienne version était en ligne.

Yahoo.fr : Nouveau look pour la mi-octobre
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

La sortie du nouveau "look" de Yahoo! France devrait coïncider avec la sortie de la lettre "Réacteur" de ce mois-ci. Si ce n'est pas encore le cas, vous pouvez aller voir cette nouvelle présentation sur le site italien (http://www.yahoo.it/) qui propose déjà ces changements. DERNIERE MINUTE : Vendredi 12 octobre, c'est Yahoo! UK (http://www.yahoo.co.uk/), qui prenait en compte également ces changements.

Sur le site italien, vous pouvez voir, en tapant un mot clé en italien (allez, au hasard, "pizza" :-)), que la présentation des résultats est bien différente de ce qu'elle était jusqu'à maintenant. De petits icônes, symbolisant des dossiers, ont fait leur apparition. Ces dossiers sont fermés lorsqu'ils caractérisent la présentation d'une catégorie, de façon générale, et ils s'ouvrent lorsque c'est un site qui est affiché mais qu'il est rappelé la rubrique dans laquelle ce site est référencé. Les logos affichés sur le site italien sont différents de ceux qui ont été choisis pour le site français, mais ils s'en rapprochent graphiquement.

D'autre part, un pictogramme "i" (comme "information") propose quelques données explicatives. C'était déjà le cas, depuis quelques semaines, pour les "sites les plus populaires" à l'intérieur de certaines catégories sur Yahoo! France (rappelons que le classement des "sites les plus populaires" est basé sur l'indice de popularité fourni par Google).

Notons que la version "résumée" des intitulés de catégories ne sera pas (encore) mise en place sur les sites européens, notamment par manque de temps des équipes de netsurfeurs. Le projet n'est cependant pas abandonné et pourrait voir le jour dès l'année prochaine. Rien de décidé pour l'instant...

L'affichage des résultats de type "sites web" est ensuite regroupé par catégories (pictogramme de dossier ouvert). A priori, les résultats seront identiques en France, à la différence près que l'url des sites ne sera pas affichée dans les fiches descriptives. Autre nouveauté : la mention "Commento Y!" (qui devrait se muer en "Description Y!" sur le site français) pour bien indiquer que ce résumé est fourni par Yahoo! et non par le site, les études menées ayant montré que de nombreux utilisateurs ne faisant pas très bien le distingo entre données issues de l'outil de recherche et informations extraites des pages elles-mêmes (comme sur un moteur).

D'autre part, les liens "envoyer cette catégorie à un ami" et "mettre cette catégorie en favori" (voir version amércaine) dans une catégorie donnée seront peut-être rajoutés sur la version française. A l'heure où ces lignes étaient écrites, la décision n'était pas encore prise.

Autre nouveautés à noter : la barre de navigation, en haut de page, qui regroupe les types de résultats identifiés par famille (Risultati principali | Categorie Yahoo! (4) | Siti in Yahoo! (80) | Altre pagine Web) s'enrichit d'une donnée ("Résultats principaux") qui correspond à la page de résultats "par défaut" de l'outil lors de la requête et propose maintenant le nombre de liens identifiés dans chaque zone.

Il est également intéressant de noter que, contrairement à la version américaine, Yahoo! France continuera, dans un premier temps tout du moins, à afficher les catégories issues de la branche "Exploration géographique" dans ses pages de résultats. Cette situation pourrait cependant changer à moyen terme avec un lien spécifique vers les résultats "régionaux" répondant à la demande, comme sur le site américain.

Enfin, si la version US propose ses "Related Searches" (expressions de deux ou trois mots clés contenant le terme demandé, ces expressions étant extraites de la base de données des requêtes les plus souvent demandées par les internautes dans les mois précédents), il n'en est pas encore question sur les versions européennes. Dommage, car ce type de données représente une aide non négligeable à la recherche d'informations. Ce n'est peut-être, là encore, que partie remise...

Rappelons également que, depuis quelques mois, un site n'est plus affiché qu'une seule fois dans la page de résultats pour un mot clé donné, même s'il est présent dans plusieurs catégories de l'annuaire. La situation est donc différente d'il y a quelques temps, lorsque 4 occurrences d'un même site étaient proposées dans la page de résultats si le site était référencé dans 4 catégories différentes. Une seule apparition dans cette page de résultat ne signifie donc pas que votre site a été enlevé des autres rubriques. Important...

Conclusion
~~~~~~~~~

On peut donc conclure que Yahoo.com a mis en oeuvre une version radicalement différente des ses pages de résultats en version américaine. L'avenir dira si cette direction est la bonne, même si, à l'usage, cette nouvelle mouture semble très agréable et bien plus efficace qu'auparavant.

Pour ce qui est de l'Europe et de la France plus particulièrement, les modifications entrevues sont moins radicales (l'effectif n'est pas le même en interne et le travail à réaliser était déjà considérable pour ce "premier jet"), mais on peut parier que de nombreuses autres modifications, dont le concept est déjà présent aux Etats-Unis, seront mises en place à moyen terme. Simple question de temps... et de moyens humains ! Rappelons que Yahoo! met également en place, en Europe et dans les semaines qui viennent, l'offre "Yahoo! Express" de soumission payante qui est une étape importante pour l'outil de recherche et qui risque également de monopoliser grandement ses équipes de netsurfeurs. Les mois qui viennent devraient donc être chauds pour l'annuaire !

- **Antisearch / Antibot : la nouvelle techno "moteur" de Francité et Lokace**

Si vous avez regardé vos logs (la mémoire des connexions de votre serveur) ces derniers mois, vous avez peut-être trouvé trace d'un robot appelé "Antibot", très actif depuis quelques temps. Il s'agit du "spider" de la société Antidot (groupe IXO, http://www.ixo.fr/, ex-Infosources), dont les agents de recherche de leur solution "Antisearch" équipent aujourd'hui les technologies "moteur" des sites Lokace (http://www.lokace.com/) et Francité (http://www.francite.com). Une bonne occasion pour tenter d'en savoir un peu plus sur ce nouvel acteur du domaine.

L'entreprise Antidot, créée en 1999 et qui compte aujourd'hui 4 personnes plus deux consultants travaillant à temps partiel pour la société, a, historiquement, travaillé sur la maintenance du moteur Lokace, dans sa première version (celle créée par l'équipe de Caramail). Puis, les services techniques de l'entreprise se sont vite rendu compte qu'il devenait nécessaire de mettre en place une nouvelle version, plus efficace, de l'outil en se basant sur l'expérience accumulée lors de la maintenance de la première version du logiciel. Le moteur de recherche a donc été entièrement réécrit,

selon un concept plus large : la solution "AntiSearch" qui a demandé plus d'une année et demi de développement de la part de l'équipe technique.

La solution Antisearch est basée sur un concept de "briques", d'agents spécialisés, à qui l'on peut demander d'effectuer un certain nombre de tâches pour aller rechercher de l'information sur des sources d'information très diverses : annuaire de site web, moteur de recherche, annuaire d'e-mail, encyclopédie, publicité, DNS (similaire à un outil de type "Realnames"), catalogue de commerce électronique, etc. La solution "Antibot" représente donc la partie strictement "moteur de recherche" d'une solution logicielle middleware beaucoup plus globale.

Une fois qu'un agent spécialisé est sollicité, le résultat de la recherche est renvoyé au site client au format XML qui peut, de son côté, l'intégrer dans une feuille de style pour affichage sous sa charte graphique. Le client peut donc choisir d'aggréger plusieurs agents pour son site, chacun d'entre eux s'adaptant à un type de donnée précis.

Par exemple, l'agent mis en place pour le site Top Achat (http://www.topachat.com/) travaille sur un export XML de sa base de données, remis à jour très fréquemment. Si vous tapez le mot clé "DVD" sur la page d'accueil du site, la page de résultats (http://www.topachat.com/micro/pages/produits.php?mc=dvd) sera fournie par les agents d'AntiSearch qui auront exploré les bases de données du site afin de retrouver l'information la plus pertinente. La technologie peut non seulement prendre en compte un site isolé, mais également des bases de données réparties sur plusieurs sites.

Si les agents d'Antisearch effectuent plusieurs requêtes, chaque agent étant spécialisé dans une tâche bien définie, les résultats proposés peuvent tout à fait être mixés, comme sur Lokace où les liens émanant de l'annuaire et du moteur sont "mélangés". Ce n'est, en revanche, pas encore le cas sur Francité (seuls les résultats "moteur" sont affichés), mais la situation pourrait évoluer prochainement.

La solution Antibot / moteur de recherche
~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

Le moteur de recherche (solution Antibot) a donc été entièrement réécrit par rapport à l'ancienne version. Il respecte l'état de l'art du domaine, et prend donc les champs suivants en critère de pertinence :

- Titre des pages (balise <TITLE>) : très important.
- Balises Meta : NON, aucune n'est prise en compte, sauf la balise Meta "Robots". Notons que le fichier Robots.txt est également pris en compte.
- Texte visible : Oui, bien sûr, et plus précisémment le premier paragraphe : plus le mot demandé est "haut" dans la page, meilleur sera le classement de celle-ci. Les critères suivants sont également importants : texte en gras, avec une taille de police importante, texte à l'intérieur d'un lien (i.e. en couleur et souligné), la proximité des mots, le nombre d'occurrences (plus un mot est présent, mieux la page est classée), etc.
- Les urls : la présence des mots demandés dans l'url, et notamment le nom de domaine n'est pas prise en compte, sauf si l'agent spécialisé sur le DNS est activé.
- L'indice de popularité est pris en compte et étudié non seulement en tenant compte des liens entrants (combien de pages, sur le web, pointent vers votre site ?) mais également sortants (vers qui pointez-vous ?). Cette analyse est récursive : l'importance (l'indice de popularité) des sites qui ont mis un lien vers vous ou vers qui vous pointez est donc également prise en compte.
- Les pages dynamiques sont prises en compte s'il n'y a pas de point d'interrogation dans les URLs. Mais cette posssibilité d'intégrer des pages incluant des passages de paramètre dans leur adresse est prévue et sera intégrée au fur et à mesure.
- Options ALT des balises IMG : pas pris en compte pour l'instant mais prévu à moyenne échéance.
- Le format Flash n'est pas pris en considération (seul le fichier HTML lançant l'animation Flash est intégré).
- Les pages graphiques (pas de texte visible affiché) ne sont pas prises en compte.
- Les commentaires (balises <!--...-->) ne sont pas lus.
- Les CGI en page d'accueil (détection de plug-in par exemple) ne semblent pas poser de problème majeur.
- Frames : elles sont prises en compte à tous les niveaux (page mère + pages filles).
- Les liens Javscript ne sont, en général, pas suivis dans les codes HTML, mais le contenu des balises "noscript" est pris en compte.

L'index d'Antisearch contient actuellement environ 10 millions de pages statiques.
L'objectif est d'aller à 20 millions pour obtenir une visibilité a peu près exhaustive du Web francophone. Le délai de rafraîchissement actuel de l'index global est d'environ un mois, avec un objectif à court terme de deux à trois semaines.

Syntaxe d'interrogation
~~~~~~~~~~~~~~~~~~~~~

L'espace est pris, par défaut, comme un ET (il existe une possibilité de changer ce mode de fonctionnement en utilisant les préférences).

Le SAUF est symbolisé par le signe "-".
Le OU peut être expréssément demandé en utilisant l'option "|".
Le "+" sert également comme un ET, notamment si on veut forcer un mot considéré comme "vide" (de, un, la, etc.).

On peut sinon utiliser la recherche avancée qui propose en outre une possibilité de restreindre la recherche à certains périmètres : sites du gouvernement francais, les pages en ".fr", etc. Antidot peut également générer des périmètres thématiques à la demande des clients (sites sur la banque et l'assurance, etc).

Via la page de préférences on peut enfin changer le nombre de réponses par page ou demander à filtrer les contenus "pornographiques".

Procédure de référencement
~~~~~~~~~~~~~~~~~~~~~~

Pour référencer votre site sur la base de données d'Antibot, vous pouvez le faire à l'adresse :
http://www.antisearch.net/URL_SUBMISSION/submit_url.php

ou sur le sites de Lokace :
http://lokace.antisearch.net/URL_SUBMISSION/submit_url.php

Il n'y a pas de différences pour l'instant entre les deux soumissions. Le délai d'apparition effective dans les réponses est de l'ordre d'un à deux mois. Pour accélérer le référencement d'un site on peut soumettre les pages principales en plus du point d'entrée, mais "pas trop" (ne pas dépasser 10 pages par exemple pour un même site, sous peine d'être pris pour un spammeur). Il ne sert à rien en tout cas de soumettre un site sur les 2 services cités ci-dessus.

La procédure de soumission sur Francité semble passer par un choix de catégories et donc plutôt servir à la soumission sur l'annuaire de l'outil de recherche. Mais le référencement sur Francité inclue, en fait, de façon "invisible", un référencement sur AntiSearch. Les sites référencés sur l'annuaire de Francité bénéficient d'ailleurs d'un "bonus" pour les recherches sur ce même outil.

Enfin, notons que Francité et Lokace disposent du même index, mais que les résultats affichés sur les pages de résultats des deux outils sont différents, les deux sites ayant demandé des "réglages" différents de leurs agents respectifs.

Le "business model" d'Antidot est donc basé sur la vente de solutions de recherche d'information clés en mains. La plupart du temps, les clients intégrent un frontal chez eux, mais toute la technologie tourne sur les machines d'Antidot (c'est le cas de Lokace et Francité). Le coût de la solution dépend du volume et est basé sur un paiement au nombre de requêtes (aux environs de 5 euros pour mille requêtes, mais ce prix peut fortement changer en fonction du site pris en compte et du type d'agent mis en place).
~~~~~~~~~~~~~~~~~~~~~~

Terminons en disant que Francité et Lokace ont fait une demande à Antidot en ce qui concerne des outils de référencement payant (de type "Paid inclusion") sans qu'aucune décision ne soit encore prise, a priori, chez ces outils pour mettre en oeuvre ce type de solution. Mais la demande est logique car tout cela est dans l'air du temps...

Bref, un nouvel entrant dans le domaine des moteurs. Les résultats de notre étude comparative des outils de recherche francophones (http://etudes.abondance.com/) indiquent d'ailleurs les progrès qu'amène cette nouvelle technologie sur des sites comme Lokace et Francité, même si elle doit encore se rôder pour arriver au niveau des meilleurs. Au vu de l'aspect encore très récent du produit, il semble qu'il en prenne, peu à peu, le chemin. Bonne route !

Plus d'infos :
~~~~~~~~~~~~~

http://www.antibot.net/
http://www.antisearch.net/
http://www.antidot.fr/

Législation : une jurisprudence dangereuse pour les moteurs ?

Un jugement rendu dernièrement par le Tribunal de Paris nous semble important dans le cadre de l'utilisation faite par les moteurs de recherche des pages du Web, et notamment lorsqu'il s'agit d'indexer des documents issus de sites dynamiques gérant des bases de données. Voici une description des faits, écrite sur la base de trois articles issus de Yahoo! Actualités (et notamment du site Transfert.net) disponibles aux adresses suivantes :
http://fr.news.yahoo.com/010906/166/1lig2.html
http://fr.news.yahoo.com/010910/166/1q6nn.html
http://fr.news.yahoo.com/011001/166/200wb.html

"Le tribunal de Paris a interdit, mercredi 5 septembre 2001, au moteur de recherche d'annonces d'emploi Keljob de référencer les fiches du site Cadremploi. Keljob a été condamné à payer 1 million de francs de dommages et intérêts.

Droit des bases de données, contrefaçon de marque et liens "profonds". Tels étaient les terrains juridiques sur lesquels l'avocate du site d'annonces Cadremploi avait orienté sa plaidoirie contre Keljob.com. La troisième chambre civile du tribunal de Paris ne lui a pas donné raison sur le dernier argument.

Le juge a condamné Keljob.com, moteur de recherche d'annonces d'emploi, pour "atteinte à la base de données" de Cadremploi , "contrefaçon de marque" et "atteinte à la dénomination sociale" de la société. Keljob devra cesser de référencer les annonces de Cadremploi et lui verser 900 000 francs de dommages et intérêts, plus 100 000 francs de frais de justice. C'est la troisième décision dans une affaire qui avait vu Keljob perdre en référé, puis gagner en appel.

Le premier élément du litige concernait la reproduction de la marque. Keljob ayant reproduit le nom de Cadremploi dans ses plaquettes publicitaires et sur son site, le tribunal a estimé que cette mention n'était pas faite dans un but d'information, mais à des fins commerciales, "dans le cadre d'une activité de recensement et de sélection d'offres d'emploi directement concurrente de celle exercée par le plaignant".

La question suivante portait sur la base de données de Cadremploi. Keljob, qui se défendait de l'avoir "téléchargée", interrogeait tous les soirs, et de façon automatisée, le site concurrent et affichait le résumé des annonces sur son site, l'internaute devant aller chez Cadremploi pour avoir l'information complète. L'expertise soumise au tribunal par Cadremploi a établi qu'en moyenne, un tiers des offres se trouvait référencé chez Keljob. Le juge a estimé que les informations reprises (secteur d'activité, intitulé de l'offre, zone géographique et date de publication) constituaient des "éléments essentiels [...] qui font la valeur de la base de Cadremploi". "Si aucune offre ne l'intéresse, l'internaute ne consulte pas le site Cadremploi dont la base de données a néanmoins été utilisée", a diagnostiqué le tribunal. Une base de données qui, rappelle-t-il, bénéficie d'une protection lorsqu'elle a fait l'objet "d'un investissement financier, matériel ou humain substantiel". En l'occurrence, Cadremploi y aurait consacré 53 millions de francs depuis 1991.

Le jugement rappelle enfin que "Cadremploi dénonce, au titre de la concurrence déloyale, la mise en place des liens hypertextes dits "profonds"". L'avocate du plaignant avait estimé que ces liens pointant vers des pages secondaires étaient "prohibés dans la mesure où ils dénaturent et détournent le contenu du site et portent atteinte à son intégrité". Le tribunal a rejeté cette conception, jugeant qu'il n'y avait pas de risque de confusion entre les deux sites, notamment parce qu'une fenêtre intermédiaire indiquait la redirection vers un site extérieur. Et que l'internaute parvenu sur Cadremploi.fr pouvait ensuite y naviguer selon son bon plaisir.

À la suite de cette condamnation, la direction de la société Keljob a indiqué qu'ils "feraient appel et retourneraient pour la deuxième fois devant la Cour d'appel de Paris". Prochain épisode de la saga Keljob/Cadremploi : résultat de l'appel en 2003."

Un jugement dont l'extrapolation peut être préjudiciable aux moteurs
~~~~~~

Cette jurisprudence sera, je pense, d'importance, car elle préfigure des problèmes qui pourraient arriver avec les index de pages web utilisés par les moteurs de recherche. Bien sûr, dans l'article qui précède, l'utilisation du nom de Cadremploi par Keljob (premier point du jugement) ne concerne pas les moteurs. Il s'agit d'un problème purement commercial. En revanche, le référencement des fiches "produits" d'un site distant pose effectivement problème, notamment pour des sites dynamiques, gérant de très nombreux descriptifs issus d'une base de données.

La question est donc : quel type de pages un moteur de recherche peut-il aspirer depuis un site distant ? A priori, au vu du jugement rendu, si ce sont des pages "institutionnelles", celà ne semble pas poser de problèmes (notion de "liens profonds"). Mais si ce sont des documents décrivant des produits d'une entreprise, celà peut être considéré comme une violation de base de données. Et les problèmes commencent alors... Mais comment dire à un moteur qu'une page est institutionnelle ou issue d'une base de données comemrciales, puisque tous les moteurs sont basés sur des procédures automatiques ?

La situation n'est pas aussi simple qu'on pourrait le croire. Bien sûr, il existe des solutions techniques, comme le fichier robots.txt (mais un aménagement spécifique de ce fichier aux restrictions techniques des sites dynamiques exploitant des bases de données, devrait peut-être être envisagé), qui permet d'exclure certains dossiers de l'indexation des moteurs ou la balise Meta "Robots" (était-elle présente sur les pages de Cadremploi ?), pourraient faire l'affaire. Mais ces fonctionnalités ne sont pas prises en compte à 100% par tous les moteurs. Et ont-elles une valeur juridique quelconque ? En tout cas, il semble clair qu'il existe ici un chantier important à mettre en oeuvre pour les moteurs sous peine de problème juridiques importants dans les mois qui viennent...

Nous avons enfin posé la question à Maître Murielle Cahen, avocate à Paris et spécialiste des problèmes juridiques liés à l'Internet, qui nous répond sur ce point précis :

"Le tribunal a jugé qu'il n'y avait pas de risques de confusion entre les deux sites, notamment parce qu'une fenêtre intermédiaire indiquait la redirection vers un site extérieur. Les liens profonds n'ont pas été sanctionnés en tant que tels, mais plutôt l'extraction de la base de donnée...

Il faudrait donc que les moteurs de recherches revoient, en fonction de ce jugement, la façon dont ils extraient les informations de sites pour les référencer. Liens profonds, oui(pour le moment), extraction de bases de données, non ! Pas si simple...

Cet arrêt est en fait assez difficile à interprêter par rapport à ce dont les moteurs de recherche devraient se prémunir. Le TGI de Paris s'est en effet prononcé sur le cas spécifique de sites spécialisés où le moteur de recherche incriminé finit par concurrencer l'éditeur. Pour un moteur comme Google ou Altavista, leur statut de "généralistes" les protège mieux contre ce type d'attaque, car ils ne concurrencent pas de façon frontale les sites web qu'ils indexent. D'autre part, la plupart de ces moteurs n'indexent pas 100% des pages d'un site web, donc pas la totalité, pour reprendre cet exemple, de la base de données produits d'un site. Cela peut jouer dans un jugement. Par exemple, pour citer un précédent, dans une affaire qui avait opposé France Telecom au 36 15 ANNU, le juge a estimé qu'il y avait problème car c'était la totalité de l'annuaire téléphonique, propriété de France Telecom, qui avait été copiée à l 'envers. Ce point de l'exhaustivité avait, à l'époque, été considéré comme très important.

D'autre part, pour revenir aux moteur de recherche, on peut estimer que le fichier robots.txt et les balise meta "Robots" peuvent être considérées comme une protection assez importante. Si un moteur indexe une ou plusieurs pages pourtant "protégées" logiquement par ces fonctionnalités, le site indexé pourra se retourner contre le moteur en cas de préjudice. Mais il devra prouver qu'au moment de l'indexation, les balises meta "robots" ou le fichier robots.txt étaient bien en ligne sur son site. Cela peut se faire, par exemple, en déposant chaque mois un CD-rom contenant les pages du site chez un notaire ou un huissier, à titre d'archives, ou en mettant ce disque dans une enveloppe cachetée et scellée que le webmaster s'envoie en recommandé avec accusé de réception, par exemple."

- **Soumissions gratuites sur les annuaires : quelles solutions ?**

Vous le savez certainement, sauf si vous vivez en hermite dans une grotte au fin fond du désert de Gobi (et encore), le monde du référencement s'oriente de plus en plus vers un modèle payant, que ce soit sur les annuaires comme sur les moteurs.

Il nous semble important de revenir sur le système de soumission payante mis en place par les annuaires. Cette prestation propose donc, contre paiement, une prise en compte rapide (de 2 à 5 jours ouvrés), donc une évaluation - dans un délai garanti - du site soumis, et une réponse par mail notifiant l'acceptation ou le refus de la source d'information par les netsurfeurs de l'outil de recherche.

Tous les annuaires majeurs proposent ou proposeront une telle offre d'ici à la fin de l'année 2001. C'est un point aujourd'hui acquis. Mais si certains webmasters ou propriétaires de sites sont prêts à payer en échange de ce service, qu'en est-il des autres ? De tous ces "petits" sites, qui contiennent parfois des trésors de contenu mais qui ne peuvent en aucun cas débourser entre 100 et 300 euros par annuaire pour une soumission ? Gros problème, il est vrai. En fait, cette question est avant tout, aurions-nous envie de dire, une question économique. Les annuaires possèdent un "staff" limité de netsurfeurs pour prendre en compte les soumissions reçues. Et ce staff sera, de façon logique, occupé prioritairement à traiter les soumissions payantes puisque, dans ce cas, il y a une garantie de délai qu'il faut absolument tenir. Les soumissions gratuites (qui restent encore possibles, pour l'instant, sur la majorité des annuaires, et on peut penser que cette situation va perdurer, en tout cas en dehors de la branche "entreprises" de l'outil de recherche) seront donc traitées dans le (peu de) temps qu'il reste au staff, une fois les soumissions payantes prises en compte. Et là, on va quasi obligatoirement se heurter à un problème d'ordre quantitatif, à savoir que, jusqu'à preuve du contraire, une journée ne contient que 24 heures.

En effet, prenons un annuaire comme Yahoo! France qui reçoit, grosso modo, un millier de soumissions par jour. Imaginons que l'offre "Yahoo! express", qui sera prochainement mise en place, fonctionne bien et que 100 soumissions payantes soient enregistrées quotidiennement (c'est un exemple). Restent 900 soumissions gratuites qui doivent être prises en compte, sachant que les 100 payantes doivent obligatoirement être traitées rapidement et donc de façon prioritaire. Rappelons que l'équipe de netsurfeurs (quelque soit l'outil de recherche) avait déjà beaucoup de mal à "tenir la distance" jusqu'à maintenant, alors que les soumissions étaient gratuites.

Le problème est en fait à la fois quantitatif et qualitatif, sachant que dans les 900 soumissions gratuites, il y a peut-être des trésors largement dignes de figurer dans l'annuaire, mais également des sites "moi-ma femme-mon chien sur la plage" qui, s'ils ont permis aux internautes qui les ont créés de se faire plaisir, n'apporteront pas grand chose au visiteur de l'outil de recherche.

Comment, donc, séparer, dans la masse des soumissions gratuites, les sites à fort contenu de ceux qui ne proposent que des informations, disons "à potentiel moyen" pour employer un doux euphémisme. Rappelons que les annuaires n'ont pas vocation à être exhaustifs, mais à proposer ce qui se fait de mieux sur le Web et, surtout, à répondre aux questions que se posent l'internaute. De ce tri des sites dans les soumissions gratuites dépendra certainement la crédibilité des annuaires dans les mois qui viennent. Rester des outils servant à fournir des réponses pertinentes aux internautes et ne pas risquer de générer une image de "pompes à fric" auprès de ses utilisateurs. Le tout sachant que la prise en considération de TOUS les sites soumis gratuitement semblent matériellement impossible, tout simplement.

Voici quelques pistes que pourront éventuellement prendre en compte les annuaires pour trier ces soumissions gratuites. Certaines sont, d'ailleurs, déjà mises en place par certains d'entre eux. Aucune ne représente, certainement, une solution miracle mais, prises en considération de façon globale, elles peuvent apporter quelques débuts de réponses :

* Nom de domaine : un site disponible au travers d'un nom de domaine propre pourrait être pris en considération de façon préférentielle par rapport à un site présent chez un hébergeur gratuit comme Multimania, Chez.com, perso.wanadoo.fr, etc. Ceci dit, au vu des nombreuses possibilités de redirection d'un nom de domaine vers un site "perso" qui existent à l'heure actuelle, ceci n'est pas obligatoirement un critère valable à 100% (quoique, avec un crawl préalable du site pour déceler la redirection, voir ci-dessous...), mais au moins, le webmaster a-t-il fait l'effort d'acheter un nom de domaine. C'est déjà ça. Comme pour tous les autres critères présentés ici, celui-ci doit être pris en compte en conjonction avec les autres et non pas de façon isolée.

* Indice de popularité : un site ayant un bon indice de popularité (de nombreux liens pointant vers lui) bénéficie alors d'un "bonus". Inconvénient : les sites soumis sur un annuaire sont souvent en début de vie et ont la plupart du temps un indice de popularité faible.

* Qualité de la soumission : tous les netsurfeurs des annuaires vous le diront : certaines fiches descriptives, remplies lors de la soumission par les webmasters ou les référenceurs, sont mieux faites que d'autres. Parfois il manque des champs, les descriptifs sont trop longs ou trop courts, les urls génèrent des erreurs 404 (eh oui), les adresses e-mail sont inopérantes, etc. Tous ces critères peuvent être testés automatiquement immédiatement après la soumission et devenir des critères plus ou moins restrictifs pour une meilleure prise en compte ultérieure.

* La situation actuelle de l'annuaire : certains outils de recherche sont parfois "pauvres" en sites (au niveau quantitatif ou qualitatif) dans certaines catégories ou en réponses à certains mots clés souvent demandés. On peut ainsi imaginer qu'une soumission d'un site dans certaines rubriques soient mises en avant et constitue une priorité par rapport à d'autres catégories qui seraient déjà bien "pleines". De la même façon, une soumission dont le titre ou le résumé contiendrait un mot clé "prioritaire" aurait, là aussi, un "bonus" dans la file d'attente.

* Veille : on voit bien (Looksmart France en est un exemple frappant) que, même si une soumission gratuite n'est pas toujours possible dans une catégorie, un netsurfeur sera "obligé" d'inscrire un site incontournable dans la "zone" dont il s'occupe s'il veut rester crédible. Tous les netsurfeurs ont des obligations de veille à l'heure actuelle, et il n'y a aucune raison pour que cela change. La meilleure façon , donc, d'être intégré dans un annuaire sans avoir à payer est de créer un site à très fort contenu, intéressant, original, bref de devenir un incontournable du domaine. Dans ce cas, le netsurfeur en entendra obligatoirement parler dans le cadre de sa veille et l'indexera de lui-même. Si, si, ça arrive, bien sûr ;-) Ceci dit, ce n'est ni la solution la plus simple, ni la plus rapide ;-))

* Crawl préalable : certains annuaires testent actuellement la possibilité d'effectuer un crawl préalable du site soumis avant prise en compte par les netsurfeurs. Ce crawl aurait pour objet de détecter d'éventuelles erreurs 404, mais il pourrait également être utilisé pour fournir des informations intéressantes : présence de liens cassés, absence de contenu en langue française, redirection d'un nom de domaine (voir précédemment), analyse des liens sortants, détection de méthodes de spam, adéquation plus ou moins bonne du contenu du site par rapport à la fiche descriptive fournie lors de la soumission, nombre de pages accessibles sur le site, détection d'un éventuel accès par mot de passe nécessaire, voire évaluation qualitative automatique du contenu (plus complexe), etc.

Ces six pistes sont bien entendu données à titre indicatif. Il en existe certainement d'autres. Comme indiqué auparavant, certaines sont déjà prises en compte par les annuaires. Cependant, on peut penser qu'aucune n'est réellement efficace de façon isolée, mais que la conjonction de certaines d'entre elles peut peut-être donner des résultats intéressants et significatifs. En tout cas, il semble quasi indispensable pour les annuaires de se pencher sur ce problème car, au vu de la situation actuelle, il semble peu probable que TOUTES les soumissions gratuites puissent être prises en compte à l'avenir. Et il serait dommage que les meilleurs sites soumis gratuitement patissent de cette situation. l'avenir dira ce qu'il en est...

- **Interview de M. Olivier Nérot, Human Links**

La société Amoweba (http://www.amoweba.com/) a dernièrement lancé une nouvelle version du logiciel Human-Links (http://www.human-links.com/), système de recherche d'informations basé sur le concept du peer to peer (P2P), prenant en compte les premières remarques des beta-testeurs. Les premiers retours ont été plutôt bons mais, comme pour tout lancement d'une version bêta, les premiers connectés se sont confrontés aux épreuves de l'installation et aux bugs de la solution. Plusieurs dizaines de milliers de pages ont cependant été recensées, les premières requêtes circulent et les contacts se nouent. Une bonne occasion pour faire un point avec Olivier Nérot, l'un des créateurs du système, en lui posant quelques questions...

- Monsieur Olivier Nérot, pouvez-vous vous présenter en quelques mots ?

J'ai 32 ans. Je suis ingénieur, docteur en sciences cognitives et passionné depuis longtemps par la façon dont la vie s'organise et fonctionne, et la frontière entre le naturel et l'artificiel. J'ai étudié dans ma thèse les liens entre le chaos, les réseaux neuronaux et la mémoire humaine. Je souhaite pouvoir appliquer ces théories à des problèmes concrets, et adopter un nouveau point de vue : celui de programmes qui apprennent, qui s'organisent, et permettent d'imaginer une nouvelle façon de traiter l'information. Bref, tout un programme... ;-)
A part cela, je peins, sculpte, répare une maison, et essaie d'écrire un livre. Mais je manque un peu de temps en ce moment...

- Pourriez-vous présenter le projet Human Links ?

Human Links (HL) est né d'une utopie : faire qu'Internet s'organise de la même façon qu'il s'est construit. Comme le langage HTML a permis à chacun d'ajouter des pages, nous cherchons à ce que chacun puisse participer à l'organisation de cette masse d'information. Il est trop rageant de voir à quel point, face à l'un des plus fabuleux moyens de communication, nous sommes seuls derrière un écran, à effectuer toujours les mêmes tâches.

J'ai donc essayé d'imaginer un système où chaque recherche organise un peu plus les information d'Internet, en conservant le point de vue de chacun. Il me semble en effet impossible d'organiser l'information de la même façon pour tous. Il fallait donc concevoir une nouvelle approche qui permette à chacun de trier les données selon son propre point de vue, et rendre l'organisation de chacun accessible à tous. Ceci, dans le respect de la vie privée de chacun.
Cette réflexion a mené au concept de Human Links : un logiciel, installé sur chaque ordinateur, qui apprend à organiser l'information pour son utilisateur, et permet de faire diffuser des requêtes "enrichies" (c'est-à-dire contenant un contexte, et pas uniquement des mots-clés), afin de réutiliser la façon dont les autres ont organisés leurs documents. Ainsi, chacun a accès au savoir de tous, mais sans avoir à espionner les documents des autres, juste en utilisant la façon dont ils les ont organisés.
Aujourd'hui, nous voulons porter HL au sein de l'entreprise. En effet, si les entreprises ont à disposition aujourd'hui des outils logiciels pour bien gérer leur information "structurée", elles n'ont pas d'outils logiciels pour maitriser, à un instant donné, quels sont les domaines sur lesquels les équipes travaillent, ni pour gérer l'evolution de cette connaissance dans l'entreprise. Aucune information dite circulante n'est prise en compte dans les systèmes centralisés de Knowledge Management. Or, notre approche correspond à un besoin : il arrive souvent que deux personnes travaillent sans le savoir à la même choses, qu'une personne refasse un travail déjà fait, qu'il existe des compétences mal révélées, etc.

Au sein d'une même entreprise, HL permet donc d'organiser automatiquement toute l'information traitée à un instant donné, sans ressource informatique supplémentaire, et devient ainsi complémentaire d'un intranet. De plus, l'intégration en entreprise est simplifiée. En effet, la collaboration est plus "naturelle", et fonctionnant en intranet, nous ne sommes plus confrontés aux problématiques de firewalls.

- Quels ont été les principaux enseignements de la première période de test du produit ? Combien avez-vous de béta-testeurs à l'heure actuelle ? Avez-vous atteint vos objectifs ?

Plusieurs enseignements : l'approche séduit mais surprend et change beaucoup les habitudes des gens. Les personnes ont dans l'ensemble pris l'habitude de recherches "jetables", et le "nombre de pages indexées" est souvent un critère de qualité... nous nous sommes habitués à perdre du temps à chercher, à chercher plusieurs fois la même information, et à débroussailler nous même les milliers de pages trouvées, sans conserver la moindre trace des recherches effectuées (les favoris sont tellement difficiles à organiser...).

Bref, si les outils de recherche d'information ne sont pas pleinement satisfaisant, nous nous y sommes tous habitués, et avons appris à travailler ainsi. Notre approche ne fonctionne que si nous savons aider les gens à dépasser ces habitudes : conserver une trace de leurs recherches, les organiser, et accepter que le temps véritable d'une recherche ne soit pas uniquement le temps de réponse du moteur...

Ainsi, au delà des bugs de la toute première version qui ont limités la diffusion du logiciel, notre proposition est peut-être trop en rupture avec les habitudes prises. Nous livrerons donc d'ici peu une nouvelle version, plus "classique", s'approchant de la façon de fonctionner d'un explorateur. La nouvelle version est donc moins en rupture avec les habitudes acquises, et nous espérons qu'elle répondra aux attentes.

En effet, sans utilisation massive du système, il n'est que de peu d'utilité... puisque chacun participe à son enrichissement.
Aujourd'hui, seuls 5 000 utilisateurs ont installé HL, ce n'est pas encore suffisant pour le rendre pertinent. Les premiers utilisateurs font donc les frais d'un manque crucial d'information. Mais peu à peu, le système se nourrit, et nous pouvons espérer le voir se déployer de façon satisfaisante, avec la nouvelle version.

Nous n'avons donc pas atteint l'objectif de valider la puissance de cette approche. Néanmoins, ce béta-test, en situation réelle, est un véritable baptème par le feu, et accélère nos développpements, par les nombreux retours que nous avons.
Les utilisateurs souhaitent voir ce principe marcher et nous les en remercions. Grâce à leur aide, je pense que nous y arriverons, pas à pas. Un tel projet ne peut être totalement fonctionnel qu'en plusieurs mois de travail intensif.

- Quels sont les avantages du P2P dans la recherche d'information ? Les inconvénients ? Les freins éventuels ?

Le P2P a démontré la puissance d'indexation d'un système distribué. Napster par exemple, au delà des problèmes de copyright, a montré à quelle vitesse le réseau pouvait indexer de l'information. De même, par l'intervention de chacun, Internet a engendré plus d'information que durant toute l'histoire de l'humanité.

Je suis donc persuadé de la puissance de cette approche, par la puissance de calcul dégagée, et l'utilisation intensive de la mise en réseau des ressources et des informations. En effet, quand on envoie une requête avec HL, plusieurs centaines d'ordinateurs peuvent chercher simultanément. Dans le cadre de l'entreprise, l'avantage est clair : notre système utilise et s'adapte à l'infrastructure existante. L'adoption de notre approche ne nécessite donc pas l'achat de nouvelles machines spécialisées, ni la création d'une équipe spécialisée pour son administration. De plus, l'installation est similaire que l'on installe le système sur 10, 1 000 ou 10 000 postes. Le P2P s'adapte à la configuration du réseau.

De plus, le logiciel étant sur la machine client, les données propres à un utilisateur restent sur sa machine : nous n'avons pas de base de données centralisée, indexant le profil de chacun ! Ceci est très important pour que l'on puisse aller plus loin dans la personnalisation, sans avoir à tout savoir sur chacun individuellement.

Il y a évidemment des freins et des limites. Le premier frein, sur lequel butent tous les systèmes P2P, est celui des barrières logicielles. Dans

certains cas, il est impossible de faire communiquer deux ordinateurs de part et d'autre d'un firewall. Heureusement de nombreuses personnes dans le monde travaillent sur ce sujet, et j'espère que des solutions apparaîtront bientôt. Autre élément, l'approche collaborative n'est pas naturelle, et il faut du temps pour faire comprendre que collaborer ne signifie pas nécessairement partager, et que chacun peut y gagner.

La perception d'un risque de faille de sécurité ne peut pas être négligé non plus... Enfin, le fait que les concept semblent très innovants alors que c'est une simple transposition de mécanismes naturels éprouvés.

- Les internautes sont-ils prêts, aujourd'hui, à partager l'information de cette façon ?

Nous avons veillé à ce que l'outil ne soit pas intrusif : il est hors de question d'aller fouiller les document de quiconque. L'idée centrale est d'organiser le Web de façon collaborative, pas de voir ce que chacun a sur sa machine. D'après les retours que nous avons, ceci a été bien perçu. HL ne sert pas à aller voir les bookmarks d'untel, mais simplement à optimiser les requêtes pour chacun. Quelques débats sur les newsgroups montrent que cette question est importante, mais que la réponse n'est pas nécessairement non.

- Estimez-vous que le premier bilan est encourageant ? Comment voyez-vous l'avenir du produit ? Quels sont les nouveaux projets dans ce domaine ?

Le projet en est à ses débuts... Après 9 mois de développements, il faut prendre en compte les points abordés précédemment : résoudre les problèmatiques de réseau, faciliter et encourager la collaboration, faciliter l'intégration de contenu.

Le premier bilan a permis de faire connaître le concept, de préparer les esprits à une nouvelle approche de la recherche d'information. Il n'est pas toujours facile, pour une structure comme la nôtre, d'être les premiers à concevoir, développer, et tester un tel système. Nous défrichons un terrain inconnu et nouveau, et nous ne pouvons profiter de l'expérience de personne. Nous tatonnons donc parfois, mais nous améliorons notre offre, et le système deviendra peu à peu très pertinent. Nous l'avons validé de façon théorique, reste maintenant à le prouver, avec l'aide de tous.

Nous prolongeons donc le béta-test jusqu'à la fin de l'année, sur deux produits : HL community, qui est une barre possédant les fonctionnalités de HL, intégrée au browser, et HL expert, qui sera la version professionnelle de recherche d'information. HL expert est aussi la première brique de notre offre pour les entreprises, que nous commençons à commercialiser.

Ainsi, pour citer Paul Valery, "le futur n'est plus ce qu'il était", mais nous nous y sommes préparés, et avons adapté notre offre, en nous centrant nécessairement sur une approche professionnelle, immédiatement fonctionnelle.

Mais l'utopie d'un moteur de recherche 'humanisé' permettant une meilleure indexation du web, plus complete et donnant accès à des informations plus personnalisées est une vision que nous souhaitons concrétiser...

Exalead : un nouvel outil de recherche innovant

Nota : cet article est une "version longue" de l'article écrit pour le site "le Journal du Net" et disponible aux adresses :
http://www.journaldunet.com/moteurs/moteurs27.shtml
et
http://trucs-et-astuces.abondance.com/outils27.html

Cet article a pour ambition de décrire un nouvel outil de recherche, ou plutôt une nouvelle technologie de recherche d'information française, créée par la société Exalead, qui se positionne, dès sa création, comme l'un des outsiders les plus créatifs de sa génération.

Les concepteurs de la technologie Exalead connaissent bien le domaine, puisque François Bourdoncle, l'un des dirigeants de la société, a travaillé pour Altavista (il avait créé la technologie Cow9, autrement appelée "Fonction Refine" sur le moteur de recherche) il y a quelques années de cela, en collaboration étroite avec Louis Monier, l'un des créateurs du moteur. Exalead ne partait donc pas dans l'inconnu, loin de là...

L'histoire de cette fonction "Refine" mérite d'ailleurs que l'on revienne quelques temps dessus. Peut-être pourrez-vous lire utilement la page d'Abondance qui parle de la genèse d'Altavista avant toute chose :
http://outils.abondance.com/av_historique.html

La fonction "Refine" : historique

François Bourdoncle travaillait à l'époque à l'Ecole des Mines de Paris. Au printemps 1996, le voici dans l'avion vers les Etats-Unis, invité par Louis Monier (avec qui il avait déjà travaillé dans une "vie antérieure") chez Altavista, donc chez Digital (propriétaire du moteur Altavista à cette époque-là) en "prospection scientifique" pour l'Ecole des Mines. L'école payait l'avion et l'hôtel sur place, mais il fallait que François travaille sur un vrai projet sur place. Du coup, dans l'avion, il cogite et pense qu'il serait intéressant d'imaginer un système qui permettrait d'affiner sa requête sur un moteur de recherche. La fonction "Refine" était née. Nom de code : Cow 9. Pourquoi ? Parce que François Bourdoncle tapait souvent les requêtes "cloud 9" (qui signifie "septième ciel" aux Etats-Unis) et "mad cow" (vache folle) pour faire ses tests de pertinence sur les moteurs de recherche. D'où un amalgame des deux en "Cow 9". C'était pour l'anecdote ;-)

Si certains parmi vous ne se souviennent pas de cette fonction "Refine" (également appelée "Live Topics") sur Altavista, voici un lien qui vous rafraichira la mémoire :
http://www.exalead.com/Francois.Bourdoncle/ina.html

Après avoir travaillé six mois sur ce projet aux Etats-Unis, il revend la technologie Cow9 à Altavista. La fonction reste en production pendant deux ans sur le moteur de recherche Altavista avant d'être enlevée après le départ de nombreux ingénieurs, et notamment de Louis Monier. Grosso modo, personne ne savait plus trop comment fonctionnait la technologie et comment la faire évoluer, et comme à cette époque-là, le moteur de recherche était considéré comme la 5ème roue du carosse (l'heure était alors à la "portalisation" à outrance), le projet a été abandonné. Dommage... Car la fonction prenait quand même en considération de 3 à 7% du trafic et des requêtes sur le moteur, ce qui n'était pas rien.

Bref, après ces péripéties altavistienne, François Bourdoncle se remet au travail au printemps 1999 et entame le développement d'une nouvelle technologie, qu'il désire plus grand public et plus pertinente encore. Les prémices de l'outil Exalead se dessinent...

Une solution basée sur l'analyse statistiques des groupes nominaux

L'outil de recherche a nécessité 250 000 lignes de code et 2 ans de travail acharné par une équipe de 2 à 6 personnes. Aujourd'hui, l'entreprise emploie 15 personnes. La technologie proposée, qui est vendue et intégrée en direct ou en OEM, a pour vocation de s'intégrer dans des portails existants grâce à un langage de développement appelé ExaScript, que la société définit comme un langage de middleware effectuant une synthèse de Java, de XML et des langages de formatage de documents de type ASP/PHP/JSP. Exalead n'a pas pour ambition de devenir un nouveau Google, mais plutôt de proposer ses outils à des sites qui désirent fournir à leurs visiteurs des fonctionnalités de recherche "intelligentes", voire qui désireraient, eux, concurrencer Google.. ;-). De même, Exalead développe actuellement lui-même ses applicatifs, mais, à terme, ce seront éventuellement les intégrateurs, voire les clients eux-mêmes qui en auront la possibilité s'ils le désirent.

La plate-forme Exalead permet de déployer des solutions de recherche dans des bases de données comprenant plusieurs centaines de millions de documents en analysant en fait statistiquement le contenu de ces résultats pour en extraire les éléments caractéristiques, groupes nominaux significatifs (en partant du principe que le sens est porté par les groupes nominaux, pas par les mots eux-mêmes) ou rubriques pertinentes (Exalead

permet en effet d'intégrer un annuaire web ou n'importe quelle classification structurée de données), et permettre ainsi à l'utilisateur d'affiner sa requête d'un clic sur celui ou celle qui correspond le mieux à son idée. Selon des tests réalisés en interne, la technologie serait 5 à 10 fois plus rapide que celles disponibles à l'heure actuelle sur le marché. La cible ? Les portails d'entreprise, les grands intranets (unification, pour la recherche d'information, de l'information proposés sur plusieurs sites distincts), les catalogues en ligne, etc.

L'idée est d'extraire des groupes nominaux des corpus sans dictionnaire mais grâce à des méthodes statistiques. La plupart des technologies imaginées par Exalead font l'objet de brevets déposés ou en cours de dépôt. Une fois le développement en bonne voie de concrétisation, l'équipe se met à la recherche de fonds. Apprenant qu'à cette époque-là, patience et longueur de temps font plus que force ni que rage, la situation se décante enfin au printemps 2000 avec une holding financière (SCA Qualis) qui engage 20 millions de francs dans l'aventure.

Le mieux, pour se faire une idée des possibilités de l'outil, est de le tester directement sur le site d'Exalead ou sur les sites sur lesquels il est déjà implémenté comme Scoot France ou 6ème Sens, le portail de Bouygues Télécom (voir adresses en fin d'article).

Sur le site d'Exalead, par exemple, tapez le mot clé "vache folle". Sur la gauche apparaîssent les catégories en relation avec la requête : Santé/Nutrition, Société, Régional/France, Actualité, Loisirs/Humour, Sciences/Agriculture, etc. Et en dessous s'affichent des mots clés connexes proposés pour affiner la recherche : Maladie de la vache folle, Épidémie de la vache folle, Encéphalopathie spongiforme bovine, Dossier de la vache folle, Abattage sélectif, Jean Glavany, etc.

Enfin, à droite de la page de résultats apparaissent les pages Web qui répondent à la question, classés par pertinence et popularité décroissante. Vous pouvez alors affiner votre recherche soit en cliquant sur un nom de catégorie, soit sur un des mots clés proposés pour effectuer une nouvelle requête prenant en compte ces données. Ce processus est itératif, permettant ainsi de converger en une ou plusieur étapes. A l'usage, on s'aperçoit qu'on arrive très rapidement à cerner les documents pertinents recherchés. S'ils existent, bien sûr.

Syntaxe d'interrogation
~~~~~~~~~~~~~~~~~

La syntaxe d'interrogation d'Exalead est assez puissante, mais également assez complexe et inhabituelle. Elle vous sera proposée si vous tapez "n'importe quoi" (en tout cas, une requête qui n'a pas de réponse) dans le formulaire de saisie de mot clé du site de démo de la société. Par exemple :

http://www.exalead.com/cgi/exalead?v=1&q=JHGjhgekgf%2Cvh%2Cbfv&p=Demo

A lire avant toute chose si vous désirez taper (et comprendre !) des requêtes comme celle-ci :

téléphone mobile/portable +nokia -siemens/ericsson ?7110/6210 ;-)

Sachez, par exemple, que l'espace est égal à un ET et que le OU se demande par un "?" (terme optionnel) ou un "/" qui permet de séparer des synonymes.

La base de données utilisée sur le site de démo d'Exalead (mais en place également chez la majeure partie de ses clients qui désirent proposer à leurs visiteurs un annuaire de sites web et un moteur de recherche de pages web) reprend les données de l'Open Directory pour la partie "annuaire" et un index de pages web (20 millions en français, 100 millions en anglais) constuit sur la base d'un "crawl" réalisé en interne, par ses propres soins. Cela le place quand même au deuxième rang des index francophones, derrière Voila (37 millions de pages).

Exalead est intéressant car c'est un outil qui semble avoir compris que c'était à lui de s'adapter à l'internaute et non le contraire. Il mène donc le visiteur par la main en lui proposant successivement de nouvelles voies pour affiner au fur et à mesure sa recherche. Une technologie qui paraît simple, rapide et... Efficace ! A tester sans tarder, assurément...

Adresses :
~~~~~~~~~~

Exalead (site de démo de la technologie)
http://www.exalead.com/

Scoot (option "Préciser la recherche" sur la page de résultats)
http://www.scoot.fr/

6ème Sens (Bouygues Télécom : module de recherche)
http://www.6sens.com/

- **Bruits et chuchotements**

Une rubrique qui regroupe tous les bruits et rumeurs dans le (petit) monde des outils de recherche mondiaux et francophones. Rien n'est obligatoirement vérifié, mais toutes les infos sont données... de source sûre ;-)

-> Le coût de l'offre "Yahoo! Express" de soumission payante sur Yahoo! France devrait se situer dans une moyenne honnête entre 200 et 300 euros. A vous de calculer ce que peut être une moyenne honnête entre 200 et 300. Non, m'sieur Yahoo!, j'l'ai pas dit ! ;-) Ceci dit, l'offre pourrait être disponible sur le site de Yahoo! France courant novembre, pour les professionnels du domaine et le grand public.
L'équipe de Yahoo! oeuvre dans ce sens, en tout cas. Courage !

-> GoTo s'installerait en Allemagne en février 2002. Il n'y aurait pas de date encore fixée pour une installation en France, mais il semblerait que cela puisse se faire dans le courant de l'année 2002 également. D'autant plus qu'Espotting, son concurrent le plus sérieux en Europe (voir l'article qui lui est consacré ci-dessus) sera en France d'ici fin 2001...

-> Altavista France attendait toujours, le 15 octobre, la mise en place de son nouvel index "touf neuf". Rappelons qu'Altavista a pris la décision de "rapatrier" tous ses index régionaux et de ne plus utiliser qu'un seul index, géré depuis Palo Alto, pour ses sites américains et hors-USA. Un nouvel index, complètement rafraîchi, devrait alors être disponible, par la même occasion, pour Altavista France. La mise en place de cette nouvelle structure était prévue pour le 3 octobre 2001, puis le 10, mais le projet semble connaître quelques retards. Ceci dit, c'était imminent au moment où ces lignes étaient écrites.

-> Altavista France (et de façon plus générale Altavista Europe) serait en train de travailler sur trois produits de référencement payant et positionnement publicitaire :
* Express Inclusion, qui existe déjà sur le sité américain (http://www.altavista.com/sites/search/express_incl) et qui devrait bientôt être disponible en France au tarif de 39$ pour 6 mois avec un tarif dégressif en fonction du nombre d'Urls à indexer. Cette offre proposerait un référencement sous 2 semaines environ et la garantie de faire partie de l'index, bien sûr. L'offre serait plutôt ciblée vers les "petits" sites (moins de 1000 pages).
* Trusted Feed , qui existe également sur le site américain (http://www.altavista.com/sites/search/trustedfeed). Il permet d'augmenter le nombre d'URLs présents dans l'index et d'améliorer la pertinence des URLs soumises lors de la saisie d'un mot clé donné. Il favorise également un meilleur positionnement dans la page de résultats. Cette offre s'adresse aux sites d'au moins 1 000 Urls à indexer, avec un maximum de 100 000 Urls (ce qui laisse quand même une certaine marge...). Le webmaster envoie toutes les semaines un fichier XML fourni au préalable par Altavista afin de donner des paramètres de crawl pour le moteur. Un refresh est donc effectué chaque semaine en fonction des paramètres déterminés dans le fichier XML envoyé par le référenceur au moteur. Le prix serait de 1$ par Url minimum et par mois (à 1000 urls minimum, faites vos comptes...).

* Le lien partenaire, offre de positionnement publicitaire, qui est déjà disponible. Le lien se situe au dessus du 1er résultat de recherche. Ce produit est commercialisé au CPM, 640 FHT en exclusivité, 520 FHT en non exclusivité. Vendredi 12 octobre, par exemple, un lien de ce type était présent pour le mot clé "assurance" à l'adresse :
http://fr.altavista.com/q?pg=q&q=assurance&kl=XX&what=fr&mm=1

- **En Bref**

-> En octobre, Yahoo! France fête ses 5 ans d'existence.

-> Looksmart France lancera ses "centres" (zones co-brandées par un annonceur, avec un lien vers celle-ci affichée sur saisie d'un mot clé) le 15 octobre si tout va bien. Après le référencement de sous-pages produits, proposé depuis quelques semaines, le produit "centres d'informations" apparaîtra en résultat de recherche d'un mot-clé et se présentera sous la forme d'un lien, "Visiter le Centre Hi-Tech" par exemple. Le lien sera positionné en haut de la page de résultats parmi les Sponsors et en première position dans la liste des résultats de recherche de l'annuaire. Après avoir cliqué sur le lien, l'internaute arrivera sur une page unique qui synthétise le site de l'annonceur : toutes les offres seront résumées et apparaîtront sous forme de liens hypertextes (image, boite de recherche...). En cliquant dessus, l'internaute sera renvoyé directement dans le site d'origine de l'annonceur où se trouve l'offre qui l'intéresse. Les Centres LookSmart seront thématiques. Trois Centres LookSmart vont donc être lancés sur le réseau français : le Centre Immobilier avec LeSiteImmobilier.com, le Centre Hi-Tech avec Teknopolis.fr et le Centre Voyage et Loisirs avec Lastminute.com.
En Grande-Bretagne, où le produit existe depuis six mois, plusieurs acteurs de l'Internet ont déjà opté pour cette solution : eBay pour le Centre Enchère, Reed.co.uk pour le Centre Emploi, Orange pour le Centre Mobile ou encore Flipside pour le Centre Jeux.

-> Le site Metaccess (http://www.metaccess.com) propose en ligne et chaque mois un "award" du meilleur référencement (choix "Metaccess Awards" dans le menu déroulant en bas de page).

-> Le site Laurion (qui propose d'effectuer une veille sur le Web afin d'y détecter des documents, on va dire, "piratés" sur votre site), a mis en place un site de démonstration de sa technologie : http://demo.laurion.com/, ce qui n'était pas le cas lors de son lancement.

-> Google a adopté un nouveau look avec des "tabs" permettant d'aller directement vers ses sites de recherche d'images, dans les forums, etc. Mais vous l'aviez tous remarqué... :-)

-> Ca y est : le site Ecila n'existe plus. Il est redirigé vers Nomade automatiquement depuis quelques jours.

abondance