- **Une question aux outils de recherche : l'indexation de parties de sites dans un annuaire**

Une nouvelle rubrique, qui reviendra périodiquement dans la lettre "Recherche et Référencement" : il s'agit de poser une question unique, mais à tous les outils de recherche majeurs du Web francophone, qu'il s'agisse des annuaires ou des moteurs, selon la question. Si vous avez des interrogations générales qui pourraient s'intégrer dans cette rubrique, n'hésitez pas ;-)

Ce mois-ci : "Lorsqu'un site est disponible à l'adresse www.site.com, est-il possible de soumettre une nouvelle fois dans votre annuaire une partie de ce site, disponible soit à l'adresse www.site.com/sports/, soit à l'adresse sports.site.com ? Ce type de soumission est-il obligatoirement écarté ? Est-il traité au cas par cas ? Dans quel cas est-il éventuellement accepté ?"

Voici les réponses des annuaires :

Yahoo! France (Cheikné Keita, Lead Surfer) :

Cela dépend. On va se poser la question : "Le site est-il une sous-page d'un site existant ? Le site est-il juste une tentative de promotion d'un site principal ?"
En effet, Yahoo! n'est pas un moteur et refuse de référencer les sous-pages qui n'apportent rien de plus que leur site principal. Il est d'usage de ne référencer que la page fédératrice et de ne pas gonfler artificiellement l'annuaire en référençant douze ou quinze sites qui n'en font en fait qu'un seul. Toutefois, quelques exceptions à ce principe de base peuvent être prises en compte lorsque la sous-page a un contenu autonome, complet et riche.
Un exemple : Le tour de France sur France 2.
Un site intéressant pour tous les passionnés de cyclisme. Pourtant, si l'on s'en tenait aux règles strictes de classification, on le refuserait sous prétexte qu'il est une sous-page du site France 2. Et comme il serait dommage de priver les internautes d'une information aussi pertinente, on indexera ce site.
Un autre exemple : Hôtel Tronchet.
Ce site est une sous-page d'un annuaire des hôtels de France qui en compte un bon paquet. Au lieu de référencer individuellement la centaine ou millier d'hôtels de l'annuaire Hôtels de France, nous référencerons l'annuaire lui-même s'il n'est pas déjà présent dans Yahoo! :-)
Pour conclure je dirai que Yahoo! remarquera certainement le contenu de qualité et évitera en tout cas de spammer l'internaute avec de nombreuses sous-pages d'un site principal.

Nomade.fr (Pauline Tourneur, Responsable de l'annuaire) :

Sur Nomade, aucune soumission n'est écartée automatiquement (404 mis à part). En effet, chaque soumission fait l'objet d'un choix documentaire, qui découle d'une politique documentaire en constante évolution. La politique documentaire a pour but d'exclure les sites à contenus illicites, mais aussi et surtout d'intégrer les sites à forte valeur ajoutée.
En théorie, en tant qu'annuaire de recherche, Nomade référence les sites web et non pas les différentes pages d'un site. Mais ce qui fait qu'une soumission sera intégrée dans l'annaire c'est avant toute chose, la qualité de son contenu, le niveau informationel apporté par rapport à ce qui est déjà présent dans l'annuaire.
Ainsi, pour les universités par exemple, il serait dommageable de ne référencer qu'une seule URL pour la simple et bonne raison que l'on passerait à coté d'un contenu informationnel important fourni par les labos, les bibliothèques d'UFR... En effet, une indexation (titre, résumé, nom, nature et origine de l'éditeur + mots clés complémentaires) ne saurait à elle seule rendre compte de la richesse d'information contenu dans le site.
C'est pourquoi, Nomade référence parfois des pages de sites. A titre d'exemple, les sites http://www.paris4.sorbonne.fr, http://www.amaes.paris4.sorbonne.fr/, et http://www.cema.paris4.sorbonne.fr/ sont référencés dans Nomade car ils apportent une réelle valeur ajouté dans leur domaine respectif, et qu'ils ont une durée de vie fiable en comparaison de certains dossiers événementiels que l'on trouve souvent sur certaines pages qui nous sont proposées en soumissions.
Lorsqu'il s'agit de plusieurs soumissions pour un même contenu, du type www.site.com/sports/, et www.sports.site.com, il ne sert absolument à rien de soumettre les deux URLs. Dans le meilleur des cas, une seule sera retenue (cf cas ci-dessus), dans le pire, les deux seront rejetées au profit de www.site.com.

Guide de Voila (Catherine Henaff, Cyberdocumentaliste) :

Les critères de référencement de sites dans le Guide sont les suivants : notoriété, richesses du contenu, fréquence de la mise à jour, qualité graphique et ergonomique, site original/insolite.
A partir de ces critères, nous acceptons de référencer une partie de certains sites. Ce type de soumission n'est pas obligatoirement écarté. Elle peut être traitée comme une soumission à part entière.
Cependant, nous ne référençons pas les sous-parties de répertoires ou d'annuaires tels que les centrales d'hôtels, les boutiques regroupant des commerçants.
Certaines parties de sites sont d'ailleurs parfois référencées sans qu'il y ait eu de soumission de la part du webmaster.
Par exemple pour les sites de médias, il est fréquent que des sujets d'actualité soient indexés.

Lycos France (Florence Lassalle, Responsable annuaire) :

Nous ne "tronçonnons" pas en règle général les sites et n'apprécions pas les soumissions faites dans plusieurs catgéories ; il peut y avoir des exceptions :
- si la partie www.site.com/autretheme/ possède une certaine automonie thématique par rapport au reste du site et si son contenu est suffisament riche ou s'il s'agit un service à part (exemple : un site ouvre une partie e-commerce de qualité : le site poourra rester sous l'arborescence de Économie, Entreprise > Sociétés et la partie e-commerce pourra être ajoutée sous Économie, Entreprise > Sociétés > Commerce, distribution > Boutiques en ligne puisque ces sites sont aussi repris dans notre guide shopping (http://www.fr.lycos.de/webguides/shopping/liste/)
- si www.site.com/autretheme/ nous aide à compléter une catégorie pour laquelle nous n'avons pas beaucoup de sites.

Open Directory (Michelle Carrupt, éditeur ODP France) :

Le "deep-linking" n'a pas très bonne presse parmi les éditeurs de l'Open Directory, qui ont appris à reconnaître les techniques utilisées pour augmenter le nombre de référencements dans l'annuaire Dmoz.
En principe, un site n'est accepté qu'une seule fois, dans une catégorie. (Dans votre exemple: Sports) Cependant, ce type de soumission est traité au cas par cas.
Les directives d'édition ne l'interdisent pas formellement, s'il s'agit de contenu unique et de qualité pour chaque catégorie. Si le contenu en vaut la peine aux yeux de l'éditeur, s'il est unique pour chaque catégorie, un référencement pour chaque subdivision peut être accepté, dans chacune des disciplines sportives dans votre exemple. Par contre, un annuaire de sites sportifs n'aurait pas droit à un référencement par catégorie.
Le bon sens et l'expérience des éditeurs déterminent le choix de l'acceptation ou du refus. Nous sommes assez vigilants face aux techniques de

deep-linking qui se multiplient!

L'Ile des Médias (Jérôme Hérard, Directeur Général Adjoint) :

Les soumissions multiples en sous-domaine sont difficiles à identifier, généralement elles passent le cap du référencement mais peuvent être rapidement déréférencées lors du contrôle de l'arborescence. Ce contrôle a pour objet d'étudier le contenu des rubriques dans le but d'adapter le rubricage de l'annuaire (trop de sites dans une rubrique par exemple) et de nettoyage (doublons). Même si la suppression du site en sous-domaine n'est pas immédiate, il est quasiment sûr qu'elle sera effectuée.
Pour les sous-dossiers sur des noms de domaine, le site ne sera pas retenu au moment du référencement dans 90% des cas (les 10% restant étant de l'ordre de l'inatention du référenceur). Ces 10% disparaissant ensuite au nettoyage de l'arborescence.

Looksmart France (Emmanuelle Siary, Directrice Business Development) :

Oui cela est possible en soumettent plusieurs urls dans notre offre "Soumission Express" (offre payante de Looksmart France, NDLR).

MSN Search (Guillaume Le Friant, Responsable outil de recherche) :

L'équipe de MSN Search n'indexe qu'une seule page par site. En fait, nous indexons dans notre annuaire des sites et non pas des pages web. Lorsqu'un site est soumis à l'inscription (http://www.msn.fr/leguide/suggestions.asp), celui-ci fait l'objet d'une évaluation basée sur des critères de sélection qualitatifs (http://www.msn.fr/help/aideguide.asp). Lorsque le site soumis est sélectionné, nos surfeurs rédigent une description reprenant l'intégralité du contenu de ce site. Les éléments d'information livrés sur nos pages de résultats (titre et description) résument l'intégralité du contenu d'un site et non pas celui d'une page ou d'une rubrique.
Dans certains cas, cependant, l'équipe de MSN Search peut indexer une rubrique spécifique d'un site si cette dernière propose un contenu de qualité, riche et parfaitement identifié. Ainsi il nous arrive d'indexer des rubriques de sites qui possèdent un nom de domaine différent ou qui font partie intégrante du site. Prenons l'exemple de TF1, nous avons indexé :
TF1
Présentation des programmes d'information et de divertissement de TF1.
http://www.tf1.fr/
Les News - TF1
Site d'information en ligne de la première chaîne de télévision française, TF1.
http://www.tf1.fr/news/
TF1 Sports
L'actualité sportive par TF1, chaîne de télévision française.
http://sports.tf1.fr/
De plus, MSN Search propose une sélection de sites couvrant les thèmes d'actualité dans la catégorie :
http://search2.msn.fr/infos_meteo/a_la_une/
Dans ce cadre, nous indexons les dossiers journalistiques ou les pages d'articles pertinents par rapport au sujet développé.

CTrouvé (Bernard Gay, Responsable de l'annuaire) :

Oui et... non : Oui si la partie du site en question est spécifique par rapport à l'ensemble du site : par exemple si le site parle de la montagne et propose une partie réservée aux sports en montagne, on enregistrera cette partie. Dans les autre cas la réponse est non. Enfin ceux qui enregistrent leur site sous la forme sport.site.com risque de passer au travers de nos programmes automatiques de validation.

- **Compte rendu de la conférence Search Engine Strategies (15 février 2001)**

La conférence "Search Engine Strategies" s'est tenue à Londres (Intercontinental Mayfair Hotel) le 15 Février 2001. Le thème principal de cette journée était le référencement et plus précisément l'optimisation d'un site dans l'optique d'un bon positionnement dans les résultats d'un moteur de recherche. Environ 250 personnes ont assisté à cette journée. Parmi celles-ci, plusieurs personnes nous ont envoyé leur compte-rendu. La synthèse de ces notes est présentée ici. Merci donc aux personnes nous ayant adressé ces informations !

Les conférenciers représentaient des sociétés spécialisées dans l'un des trois domaines suivants :
- le référencement,
- le design de sites Web,
- les moteurs de recherche.

Les personnes assistant au salon étaient dans leur majorité issus de deux milieux :
- sociétés dont l'activité principale est le référencement,
- sociétés possédant des sites web marchands avec comme unique but l'amélioration de leur position dans les résultats des moteur de recherche.

Environ 250 personnes ont donc assisté à ce salon, dont la majorité était d'origine britannique ainsi que des pays nordiques. Notons la très faible proportion de français parmi l'assistance.

Voici un compte-rendu des différentes conférences.

Advanced Search Engine Optimisation Issues (Advanced Track)

Thème abordé : Le sujet principal de cette conférence était le cloaking et dans une moindre mesure le spam. Cette conférence fut la plus instructive de la journée.
Elle était organisée sous la forme d'une table ronde animée par Detlev Johson, vice-president responsable de la technologie de la société Position Technologies. Les personnes répondant aux questions posées par l'assistance étaient les suivantes :
- Mikkel Demib Svendsen, vice-président responsable développement produit pour ACSys (Scandinavie)
- David Turner, Outrider (Royaume Uni)
- Henrik Anton, responsable Stratégie Marketing Internet et Spécialiste SEO (Search Engine Optimisation) pour B-Seen (France)
- Swen Heiderich, responsable production pour NetBooster (Allemagne).
Il faut souligner la très grande compétence des conférenciers suivants : Detlev Johson, Mikkel Demib Svendsen, Henrik Anton.

Le cloaking

L'ensemble des conférenciers s'entendaient pour dire que la pertinence est le meilleur moyen d'être correctement référencé. Cependant, le cloaking semble être une technique utilisée par certains référencurs pour améliorer le positionnement de leurs clients. Une première définition du cloaking pourrait être la suivante : une page web "cloakée" est une page dont il existe deux versions : une destinée aux spiders utilisés par les moteurs de recherche pour indexer le web, une destinée à l'utilisateur final (l'utilisateur ne voit, lui, que celle-ci). L'intérêt de cette technique est bien évidemment d'améliorer le positionnement d'une page par rapport à celui que la page vue par l'utilisateur final aurait obtenu en passant directement dans les "mains" d'un spider. Il existe évidemment un problème d'éthique lié au cloaking : les différents intervenants n'ayant pas la même opinion à

ce sujet. Le seul point d'accord entre les participants est le suivant : pour être honnête , un référenceur doit lors de la soumission d'une page indiquer si elle cloakée ou non.
Au coeur du débat sur le cloaking, les remarques suivantes furent émises par les intervenants :
- une mauvaise utilisation du cloaking peut être considérée par les moteurs comme une sorte de spam,
- les moteurs ne considèrent pas le cloaking comme du spam et donc ne bannissent pas les sites l'utilisant correctement, du moins pas encore...

Pour en savoir plus :
Qu'est-ce que le cloaking ?
http://www.ability-referencement.com/068bis-reponses-referencement-cloaking.html
Est-ce que le cloaking est autorise ?
http://www.ability-referencement.com/068bis-reponses-referencement-cloaking-autorise.html
Qu'est-ce que le page hacking ?
http://www.ability-referencement.com/068ter-reponses-referencement-page-hacking.html

Le "page jacking"

Il s'agit d'une technique de cloaking dont la légalité est discutable. Les moteurs bannissent les sites la pratiquant. La solution est simple : pour être positionné en numéro 1 pour une requête donnée sur un moteur, il suffit de suivre les étapes suivantes :
- identifier la page numéro 1 pour la requête donnée,
- télécharger la page sur son site,
- attendre patiemment le spider ou soumettre la page,
- lorsque la page est référencée (probablement en numéro 1), changer le contenu en celui de son site !

Indexation du DHTML

Les outils de recherche indexent le DHTML.

Les liens

Le poids des liens (indice de popularité) devenant de plus en plus important dans les algorithmes de positionnement des moteurs, il est conseillé d'avoir le maximum de sites pointant sur son site (sur la home page de préférence).

Les pages dynamiques

L'indexation des pages dynamiques par les moteurs est un problème complexe car il peut amener les spiders à un bouclage infini conduisant à un plantage du serveur sur lequel ils tournent. Google est l'un des pionniers dans ce domaine.
Ainsi, il est aujourd'hui conseillé au webmasters de générer dynamiquement des pages statiques à partir de leur base de données. Il faut éviter les " ? " dans les URL.
Pour cela, il faut par exemple se servir du "rewrite module" d'Apache. Voir :
http://www.apache.org/docs/mod/mod_rewrite.html
Et :
http://www.engelschall.com/pw/apache/rewriteguide/
Avec Coldfusion, il est possible de transformer des URL de type
http://www.cvfm.com/page.cfm?ID=A130
en
http://www.cvfm.com/page.cfm/A130
Une discussion a déjà eu lieu sur le forum Coldfusion sur la transformation des " ? " :
http://forums.allaire.com/DevConf/index.cfm
Pour des discussions sur le referencement des site site Coldfusion :
http://forums.allaire.com/devconf/Index.cfm?Message_ID=18401
Un passage du compte rendu de la conférence détail les méthodes d'indexation de pages dynamiques.

Deep submit

Pour augmenter les chances que les robots indexent en profondeur les pages du site. Je n'ai pas assisté à la conférence et je ne sais pas le niveau d'efficacité de ces outils mais il est bon d'en tester quelques uns :
- PositionPro (http://www.Positionpro.com)
- Eluminator (http://www.mediadna.com)
- Searchenginecommando (http://www.Searchenginecommando.com)
- Ebrand management (http://www.ebrandmanagement.com/searchmechanics/home)

International Issues (Advanced Track)

Thème abordé : Le sujet principal de cette conférence était les problèmes liés à l'internationalisation des sites.

Cette conférence était organisée sous la forme d'une table ronde animée par Detlev Johson, vice-president responsable de la technologie de la société Position Technologies. Les personnes répondant aux questions posées par l'assistance étaient les suivantes :
- Ted Willmarsgard, spécialiste SEO pour B-Seen (Suède),
- Graham Hansell, Directeur pour SiteLynx (Royaume Uni),
- Raphaël Richard gérant de la société CVFM (France),
- Lars Rabe, Directeur chez NetBooster (Allemagne).

Le contenu de cette table ronde fut très varié sans véritable fil conducteur. Les informations les plus pertinentes furent les suivantes :
- il est préférable d'avoir un site local (dans la langue locale) que d'avoir une traduction de sa page sur le site de référence,
- si une entité possède des sites dans différents pays, il est préférable d'utiliser des sous-domaines que des sous-répertoires.

Designing Search Engine Friendly Sites

Thème abordé : Le sujet principal de cette conférence était "Comment construire un site Web que les moteurs de recherche positionnent correctement dans leur classement".

Cette conférence était dirigée par Shari Turrow, Webmaster et responsable marketing chez GrantasticDesigns.com

Les principales clefs pour le design d'un site "apprécié" par un moteur de recherche sont les suivantes :

- privilégier le texte (le flash, les images ne sont pas indexés),
- simplifier au maximum le code html (en n'hésitant pas à modifier le code html généré par le logiciel utilisé),
- utiliser à bon escient les boutons de navigations et les "image map" (le plus judicieux étant d'y ajouter des liens sous forme de texte),
- éviter les pages dynamiques,
- éviter les pages de redirection (doorway page) du style "cliquer ici pour entrer sur le site",
- limiter au maximum les frames (dans le cas d'utilisation de frames, mettre le texte dans un "frameset"),
- limiter l'utilisation de Javascript.

Certains designers utilisent aussi des techniques de "cascading style sheet" (texte invisible, texte de la même couleur que le fond, ……) mais la plupart des moteurs les détectent avec le risque d'être alors catalogué comme "spammeur".

Leveraging Your Content

Thème abordé : Cette table ronde était une séance de questions/réponses sur la dernière conférence de la matinée dédiée au design des sites webs (cf. paragraphe précédent).

Cette séance de questions réponses avec Shari Turrow, Webmaster et responsable marketing chez GrantasticDesigns.com était animée par Detlev Johson, vice-president responsable de la technologie de la société Position Technologies.

Les principales informations récoltées lors de cette séance furent les suivantes :
- limiter l'utilisation du xml que seuls Google et Inktomi référencent à l'heure actuelle. D'après Detlev Johson, le xml n'atteindra sa maturité que dans 5 ans,
- il est préférable d'utiliser le xhtml que le xml,
- il est souhaitable d'avoir des sites références poitant sur son domaine (Yahoo, ODP, ...).

Dealing with Directories

Thème abordé : Le but de cette conférence était de mieux connaître et comprendre les annuaires.

Cette conférence, animée par Danny Sullivan, se déroula en deux temps :
- présentation de la société par chacun des responsables des annuaires,
- questions/réponses avec l'assistance.

Les annuaires suivants étaient représentés :
- Lycos UK : Alex Kovach,
- ODP : Bob Keating (Resposable Edition),
- LookSmart : Johny Cussen,
- AskJeeves.

Cette séance fut très riche. Les nouvelles informations glanées lors de cette séances furent les suivantes :
- ODP est le seul annuaire complètement gratuit en terme de soumission et d'utilisation et le restera,
- LookSmart UK est détenu à 50% par LookSmart et à 50% par British Telecom.
- GOTO = "highly relevant results for consumers"
- Les outils de recherche aux USA drainent en tout de 180 à 200 millions de requêtes par jour, rien qu'aux USA. Selon Nick Hynes, 40 à 60% seraient d'ordre commercial.
- Le prix de la position des sites est mis au enchères par Goto. Ces prix peuvent être consultés et modifiés en temps réels par les annonceurs. Inktomi est leur "backfield partner" : il donne les résultats secondaires lorsque Goto ne peut proposer de réponses. Une première inscription de 50 $ est nécessaire. Les annonceurs ne payent que lorsqu'un internaute a cliqué sur la bannière ou le lien correspondant. Selon Goto : le coût moyen d'une publicité est de $0.17, contre 5$ pour la bannières (£20 CMP, 4% de CTR). Goto estime avoir 75% du "pay for search engine" au Royaume-Uni. Goto n'est pas un site de "destination", il n'y a donc pas forcément un gros trafic sur leur site.

Meet the Search Engines

Thème abordé : Le but de cette conférence était de mieux connaître et comprendre les moteurs de recherche.

Cette conférence, animée par Danny Sullivan, se déroula en deux temps :
- présentation de la société par chacun des responsables des moteurs,
- questions/réponses avec l'assistance.

Les moteurs suivants étaient représentés :
- Inktomi : Joe Frost (Directeur du Marketing),
- Altavista : Jean Hagerty - (technical architect),
- Google : Craig Silverstein (Directeur de la Technologie).

Les nouvelles informations glanées lors de cette séances furent les suivantes :
- Inktomi estime la taille actuelle du Web à près de 3 milliards d'URLs. L'index d'Inktomi serait rafraichi tous les 20 jours. Ils traitent 60 millions de requêtes par jour à destination de 125 portails, en particulier celui d'AOL.
- l'algorithme de pertinence d'Altavista utilise l'indice de popularité. D'après Altavista, le spam existe mais il est facilement détectable. Les pages nord américaines représentent 5/8 de l'index total d'Altavista. AltaVista index 75 Millions de nouvelles pages chaque semaine.
- Il y a eu de nombreuses difficultés initiales dans l'algorithme "PageRank" de Google (problèmes de références circulaires). Google gère les pages dynamiques malgré la difficulté technique, ainsi que le PDF. Google garde en cache l'intégralité du web sur leurs serveurs (mise à jour tous les mois). Il existe une fonctionnalité de détection automatique de la langue de la requête. Une révision de l'index est effectuée tous les mois chez les partenaires de Google, par téléchargement complet du web sur leurs disques durs. Sur Google, le ranking par rapport à une requête se fait à partir du page rank (liens) et du contenu du texte. Le nombre de liens pointant sur une page ne compte pas : c'est le poids des liens qui compte. Google a commencé à crawler les URL dynamiques, depuis quelques mois (ainsi que les formats WML, XML, PDF). Google ne conseille pas de référencer son site : leurs crawlers le font très bien. Google : "no cookie, please !"
- Google et Northern Light interdisent le cloaking tandis que certains outils européens l'autorisent au même titre que les pages satellites.

Quelques infos glanées sur les outils de recherche

ODP (Open Directory)
---

34 000 éditeurs
2,3 millions de sites, 250 demandes de référencement à l'heure +5% par mois
229 pays, 60 langues
Contenu utilisé par des portails verticaux qui reprennent des parties du contenu de Dmoz (exemple : humania2.org)

77 000 sites allemands, 55 600 sites polonais, 48 700 sites espagnols, 33 000 français, 24 900 suédois, 19 600 sites hollandais, 18 900 sites italiens. 100% gratuite pour les Internautes : selon Bob Keating, il ne sera jamais question de référencement payant.

Looksmart

Contrôlent 100% de leurs filiales chinoises, australiennes et américaines.
Joint-Venture en Europe et en Asie qui s'appelle BT Looksmart (50%-50%). Pénétration (à travers les accords de distributions comme avec Club Internet) : 30% en France, 50% au Royaume Uni, 40% au japon, 35% en Hollande, 35% au Danemark, 30% en Suède.
La pénétration correspond au % de la population locale qui effectue au mois une recherche par mois sur la version locale de Looksmart ou les sites qui utilisent sa base.
Apparement, ils travaillent sur les technologies wap, wml, cHTML, genie (Japon). Au Royaume Uni, Looksmart alimente Excite.co.uk et msn.co.uk (renvoie des résultats différents de Excite.co.uk et de Looksmart.co.uk)

Distribution de Goto Uk dans les outils UK

L'achat mots-clés dans Goto Uk permet d'arriver en premier sur Freeserve (plus précisément les 10 premiers résultats ont des résultats de Goto UK : testez sur "Cars" ou "Search Engine Optimisation"). Distribué également sur Dogpile (métamoteur apparemment un peu utilisée au UK). Par ailleurs, je pense qu'ils seront bientôt sur altavista.co.uk qui risque de ne plus s'alimenter a la base général Goto US. Plus de 90% de trafic de Goto provient du trafic sur des moteurs partenaires. Le CA visé pour GOTO Europe en 2001 : 200 M$: "we treat search as a market place"

Lycos UK
---

24 éditeurs
114 000 liens

Inktomi

Au total, Inktomi possède 4 index :
- l'indexation des données multimédia (Inktomi Media Index)
- Gen3
- Webbestindex
- Private value.
Inktomi traite 60 millions de requêtes par jour sur les 125 portails qu'ils ont en ASP.

Nouveautés chez Altavista

Le site américain Altavista.com a fait peau neuve il y a peu. Notons que, le mois prochain, la lettre "Réacteur" ici présente, consacrera un dossier à la nouvelle mouture d'Altavista France, elle aussi totalement rénovée depuis quelques jours. Altavista.com a profité de cette nouvelle version pour apporter quelques modifications à son outil de recherche.

Recherche d'information

Tout d'abord, l'index d'Altavista est passé de 350 à 550 millions de pages web. Bien entendu, comme vous l'avez certainement constaté, le portail est devenu outil de recherche, se recentrant autour de son activité originelle : la recherche d'information. Ceci dit, cela ne semble pas remettre en cause l'existence du site Raging Search (http://www.ragingsearch.com/), qui avait également cette même approche à son lancement.
Un outil de recherche spécifique sur le commerce électronique (http://shopping.altavista.com/) a également été lancé sous une nouvelle forme, car, selon Altavista, 20 à 25% des requêtes ont trait à ce domaine sur le Web, notamment américain. D'autres outils de recherche verticaux, comme sur l'actualité (http://news.altavista.com/) sont également disponibles.

Expressions, ET et OU

Sur l'outil de recherche, Altavista a mis en place un nouvel algorithme de recherche. Pour les expressions saisies comprenant 2 à 4 mots clés, il effectue, au départ, une recherche dans un dictionnaire qui lui est propre et qui contient quelques 500 000 expressions constituées de plusieurs termes. Si l'expression demandée est identifiée dans ce dictionnaire, c'est celle-ci qui sera recherchée, comme si l'internaute avait utilisé les guillemets. Les pages les contenant "telles quelles" seront donc affichées en premier dans les résultats. Puis, à la suite, les documents contenant obligatoirement tous les mots demandés (opérateur ET entre les différents mots) sont affichés. C'est donc maintenant un ET qui est appliqué par défaut, à la place du OU, comme avant. Pour les expressions de 5 mots et plus, la détection de phrases dans le dictionnaire est encore active, mais c'est, ici, un OU qui sera appliqué entre les termes une fois les occurrences de type "expressions" affichées. Mais il est vrai que peu de gens tapent plus de 4 termes dans le formulaire.

Pour résumer :

- Mot clé unique : l'algorithme "classique" est pris en compte.
- Expressions de 2 à 4 mots : recherche en premier de l'expression exacte (équivalent aux guillemets) si elle est trouvée dans le dictionnaire d'Altavista. Recherche avec un ET entre les mots après affichage des pages contenant l'expression identifiée dans le dictionnaire. Si l'expression demandée n'est pas trouvée dans le dictionnaire, un opérateur ET est appliqué entre les différents termes dès le départ.
- Expressions de 5 mots et plus : processus identique au précédent, mais avec un OU entre les différents mots clés à la place du ET.
Bien entendu, les autres critères de classement (notamment l'indice de popularité) d'Altavista doivent également être pris en compte dans cette réflexion. Tout n'est pas si simple chez les outils de recherche... ;-) Enfin, la fonction "Search Within Results" a également vu le jour dans les pages de résultats du moteur. Elle permet, comme sur de nombreux autres outils de recherche, d'affiner une recherche sur les pages identifiées sur une première requête.

Nouvelle procédure de référencement

De plus, un nouveau système de soumission de sites devrait être mis en place d'ici peu. Partant du principe que 90% des sites qui leur sont soumis contiennent du spam (Inktomi, de son côté, avance le chiffre de 95%), Altavista désire mettre en place un système de filtre humain à la soumission.

Selon Altavista, ce procédé va affecter à la fois la procédure de soumission et les délais de prise en compte (jusqu'à maintenant, un site soumis était indexé approximativement dans la semaine).
Lorsqu'un webmaster voudra référencer un site, une suite aléatoire de chiffres et de lettres (appelée Submission Code) apparaîtra. Pour référencer le site, il faudra rentrer à la main cette suite de caractères dans un formulaire adéquat. Il sera ensuite possible de soumettre jusqu'à 5 adresses différentes avant qu'une nouvelle demande de saisie d'un nouveau code soit effectuée. Cette fonctionnalité est, d'ailleurs, déjà utilisée pour la

soumission d'un site sur le "Live Directory", l'annuaire de NBCI (http://www.nbci.com/). Il s'agit clairement, ici, de bloquer la soumission de pages de façon automatique, qui est l'une des pratiques préférées des spammeurs. Plus d'infos bientôt...

- **Délais actuels des moteurs de recherche entre une soumission manuelle et la prise en compte effective d'une page**

Au tableau d'honneur : les moteurs qui indexent de nouveaux sites assez rapidement (moins de 15 jours) : Altavista France (il semble y avoir cependant un petit problème de ce côté là depuis le lancement de la nouvelle version), Google, Fast (donc Lycos et Lycos France).

Les moteurs qui sont "moyens" (certains sites indexés de façon plus ou moins rapide, d'autres non pris en compte, bref des délais assez variables mais qui restent raisonnables) : Inktomi (donc Hotbot), Ecila.

Les moteurs à la peine ce mois-ci : Altavista.com, Voila, Excite.

- **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 ;-)

-> Une toute nouvelle version de l'index et du système d'indexation de pages du moteur Voila serait en cours d'élaboration. Mise en place imminente (dans les semaines qui viennent)...

-> L'indexation des sites dans l'annuaire de Voila (le Guide) semble également en panne. Plus d'infos le mois prochain, normalement. Pas mal de bouleversements à prévoir également chez Nomade, d'après certaines rumeurs, dans les semaines qui viennent...

-> Voila n'indexerait que les pages d'accueil des sites en JSP.

-> Ecila indexait 3,5 Millions de pages en janvier et février dernier. Ils sont passé à 7,5 Millions il y a quelques jours. L'index d'Ecila devrait atteindre les 10 Millions de documents référencés d'ici peu.

-> Ecila n'accepte plus de soumission de site depuis d'autres pages que la sienne (l'"officielle"). Les autres soumissions (logiciels, sites web) sont automatiquement refusées.

-> Prochaines améliorations en vue sur Ecila :
- Une seul URL par site dans la réponse proposée (clustering).
- Optimisation des résultats prenant en compte de nouveaux facteurs : nom de domaine, URL , taux de clics, classification des websurfers.
- A venir dans un avenir moins proche : utilisation du graphe (nombre de referers vers un site), ajout d'outils complémentaires de lutte contre le spamming/spamdexing, rajout d'une fonction d'analyse linguistique de la page sélectionnée.
- Déjà réalisé : un gros nettoyage de la base d'origine qui a supprimé pret de 40% des entrées (inutiles, sites fermés, sites adultes, URL inclues dans d'autres etc).
- A venir en parallèle : le croisement avec un annuaire afin de fournir des réponses catégorisées. Les développements (mineurs) et les adaptations (importantes) sont en cours : -)

- **En Bref**

-> Direct Hit n'est plus utilisé dans l'affichage des résultats de Lycos France. Les premiers sites affichés sont donc ceux de l'annuaire (lorsque le nom de la catégorie apparaît), puis ceux de Fast. Le mois prochain, la lettre "Réacteur" proposera un article complet sur la façon dont est intégré Fast dans Lycos France.

-> Le nouveau Copernic (version 5.0, appelé Copernic 2001) sera disponible dans la semaine du 20 mars.

-> L'annuaire Ctrouvé (http://www.ctrouve.com/) prépare une nouvelle version de son annuaire/moteur. Parmi les nouvelles fonctionnalités, on trouvera notamment :
- Un découpage en une douzaine d'univers (thèmes).
- Une indexation immédiate de la modification d'un site déjà référencé.
- Un service payant par abonnement : Référencement 24h, positionnement payant, achat de pub sur le site.
- Des critères de recherche habituels + popularité + audience + "le plus visité" + disponibilité, ...
- Une recherche annuaire + recherche contenu 1ère page (avec prise en compte des méta tag).
- Des fiches d'infos par site (avec possibilité de rajouter des infos payantes).
- Un parcours par mots clés "multi dimensionnel" amélioré. - Des outils webmaster gratuits/payants (audience, liste de diffusion, contrôle de la disponibilité d'un serveur).
- Une reprise de l'envoi gratuit par mail des nouveautés avec filtre par mots clés/thèmes.

Afin de qualifier au mieux la base (+ de 155.000 url inscrites), un mail à chaque webmaster et/ou un contrôle par moteur sera effectué. Toutes les autres rubriques éditoriales (partenaires) seront abandonnées dans un premiers temps. La volonté de CTrouvé étant de se recentrer sur la recherche d'informations et sur l'aide aux webmasters.
CTrouvé souhaite également que la publicité ne soit plus donnée en excusivité à une régie mais proposée en achat direct aux sites inscrits (paiement par CB) à un prix très compétitif. Des bandeaux gratuits pour les sites non commerciaux seront également prévus.
Leur objectif en terme de planning est de faire coïncider ces améliorations avec leur 5ème anniversaire : le 1er mai 2001, bien que certaines fonctions ne seront mise en place qu'après cette date.

abondance