INTERNET
Recherche avancée et outils de veille

Support de cours commun

ADBS – Octobre 2002

(version révisée août 2003)

"Trouver l'information est un art, pas une science" Jean-Pierre Lardy

dd/06/yy

SOMMAIRE

PREMIERE PARTIE
Recherche avancée

Points de repère sur l'Internet

Les internautes

- ✓ Estimation à 665 millions d'utilisateurs dans le monde au début 2003 (pour 400 millions début 2001 et 540 début 2002), selon Computer Industry Almanac www.c-i-a.com. A noter que les estimations Nielsen Netratings se situent en dessous www.nielsennetratings.com, mais encore faut-il se mettre d'accord sur le concept d'"utilisateur"!
- ✓ Estimation 2004 : 724,9 Millions d'utilisateurs, chiffres repris par le Journal du Net http://www.journaldunet.com/cc/01_internautes/inter_nbr_mde.shtml
- ✓ En France, environ 18,7 millions de personnes s'étaient connectés durant le mois de janvier 2002 selon Mediamétrie pour moins de 12 millions en début d'année 2001 et 17 millions durant le mois de janvier 2002 (www.mediametrie.fr), 21,4 Millions en juin 2003 (toujours source Mediamétrie, repris par le Journal du Net ici : http://www.journaldunet.com/cc/01_internautes/inter_nbr_fr.shtml)

  Selon une étude du cabinet GfK, en France 24 % des foyers disposaient d'une connexion à Internet fin 2002, contre 22,4% fin 2001, 17% fin 2000 et 11 % fin 99 (étude annuelle réalisée pour le compte du magazine Science et vie Micro).

La Taille du Web

Il est très difficile d'estimer la taille réelle du web. Sa croissance se poursuit à un rythme très rapide (quelque 7 millions de pages supplémentaires par jour), mais de nombreuses pages ont une durée de vie très limitée. La plus grande difficulté provient aujourd'hui du nombre très important de pages dynamiques (cf le chapitre consacré au web invisible), et donc de la définition que l'on donne à une "page web". Cela dit, en toute logique, on doit dépasser actuellement les 4 milliards de pages, sans compter les informations contenues dans les bases de données.

Les études sérieuses sont malheureusement rares :

(voir aussi sur http://c.asselin.free.fr/french/webenchiffre.htm)

Benchmark Group, avril 2001	2,9 milliards de pages
Cyveillance, juillet 2000	2,1 milliards de pages
Inktomi/Nec Research Institute, déc 1999	plus de 1 milliard de pages
Nec Research Institute, février 1999	800 millions de pages
Nec Research Institute,décembre1997	320 millions de pages

Plus de 42 millions de sites web au niveau mondial, pour 1 million en avril 97 et 7 millions en 2000 (selon Netcraft www.netcraft.com). (chiffres juillet 2003 : **42,298,371** http://news.netcraft.com/archives/2003/07/02/july_2003_web_server_survey.html)

A noter : Selon une étude de juin 2001 de l'OCLC (Online Computer Library Center, Inc), le nombre de sites était alors de 8,7 millions, contre 7,4 en 2000. (http://wcp.oclc.org) ; Netcraft donnait à la même époque une estimation de 27 millions. Contrairement aux apparences, ces deux chiffres étaient à peu près compatibles En effet, pour l'OCLC, chaque site correspond à une adresse IP distincte, quant Netcraft tient compte des différents sites coexistant sous une même adresse IP.

La topologie du Web

Selon une étude menée par des chercheurs d'IBM, Compaq et AltaVista, parue en mai 2000, le Web aurait la forme d'un « nœud papillon » comprenant 4 parties. Le nœud ou "cœur" du net, très interconnecté, représentait 30 % des pages. Il est facile d'y accéder depuis de nombreux sites, simplement en suivant les liens. Environ 24 % des pages sont considérées comme "initiatrices". Leurs liens permettent d'accéder au cœur du web, mais la réciproque est fausse. À l'inverse, les pages "destination" (24 % des pages sondées) peuvent être facilement repérées depuis le cœur du web, mais elles n'y renvoient pas. Les 22 % restants sont des pages complètement disjointes du cœur. Elles peuvent être reliées à des pages initiatrices ou destination, voire même constituer des îlots totalement déconnectés. Il peut s'agir des pages perso d'une famille ou d'un groupe d'étudiants, par exemple. Seule solution pour s'y connecter : connaître l'adresse, puisque même les moteurs de recherche ne peuvent les trouver.

Cette étude n'a malheureusement pas été remise à jour récemment.

(http://www.almaden.ibm.com/cs/k53/www9.final/)

Caractéristiques de l'information sur Internet

- ✓ Grande hétérogénéité dans les contenus et dans les publics (grand public et professionnels)
- ✓ Contenus dynamiques et renouvellement continuel
- ✓ Instabilité des localisations (de plus en plus d'erreurs de type "404")
- ✓ Fragmentation plus ou moins importante, selon les disciplines
- ✓ Multilinguisme et couverture géographique mondiale
- ✓ Information gratuite et payante (tendance à plus d'information, plus rapide, moins chère, avec une frange d'information à valeur ajoutée payante).

Les dix règles d'or de la recherche d'information sur Internet

1. **"Affiner"** savoir poser les bonnes questions :sa question (type de recherche, sujet précis et objectif, étude des concepts, recherches préliminaires éventuelles), choisir ses stratégies de recherche. (OA "lorsqu'on a une recherche à faire sur le web, la première chose à faire, c'est de ne pas aller sur le web")
2. **Maîtriser** les outils de navigation et de recherche : gestion des signets, récupération des données, répertoires, moteurs et méta-moteurs. Pour les moteurs, utiliser au moins deux moteurs ayant des approches différentes et complémentaires.
3. **Trouver** de bons points de repère : annuaires et "bons sites" (associations professionnelles, experts, usuels du domaine) dans un domaine :
    - Retrouver les équivalents de ses sources habituelles (d'où l'importance d'avoir une idée, même approximative, de l'offre documentaire dans le domaine recherché).
    - Compléter avec les sources originales
    - Trouver les répertoires et "méta-pages" spécialisées.

    Une adresse fiable qui renvoie directement au sujet d'une recherche constitue un bon point de départ parce que :

    L'administrateur d'un bon site spécialisé est généralement averti de l'existence et la création des autres sites de la spécialité : Il sélectionne les meilleures références et parfois les commente ; Il passe du temps sur le réseau dans son domaine de compétence ; Il met en jeu son expertise.
4. **Toujours analyser** l'information : recouper l'information, faire preuve d'esprit critique, évaluer rapidement
5. **Utiliser** en cours de recherche son carnet d'adresses pour garder trace des sites ou pages intéressants mais momentanément hors sujet, et "noter" rapidement les ressources enregistrées.
6. **Savoir se limiter** dans le temps : ne pas se rendre esclave d'une recherche d'exhaustivité à tout prix, ne pas s'obstiner en vain. Internet contribue souvent à répondre à la question "où trouver" (chercher l'info qui conduira à l'info).
7. **Choisir** les bons mots-clés
8. **Rester clair** sur ses objectifs, sa stratégie et ses critères de choix établis auparavant face à "l'hyper-choix". Rester vigilant sur la trajectoire parcourue et celle qui reste à parcourir. "on ne doit pas rechercher l'info de la même manière suivant que l'on est novice ou expert sur un sujet.

    Le novice recherche les sites web les plus riches et les plus visités. Il n'a pas de temps à perdre et veut éviter le bruit. Il obtient des résultats rapides, après la phase d'acclimatation au problème.

    L'expert n'est pas intéressé par les sites classiques. Il recherche au contraire le bruit afin de trouver le "signal faible" qui lui donnera l'avantage. Il est prêt à y consacrer beaucoup de temps. (il fait beaucoup d'efforts pour des résultats marginaux)
9. **Conjuguer harmonieusement** recherche dans les outils classiques, web invisible, presse et actualité et navigation hypertexte : la recherche d'information sur Internet est un processus itératif qui oblige à passer par différents modes d'accès à l'information.

Etre "agile" : Développer une lecture rapide, lancer plusieurs recherches à la fois, savoir rebondir d'une information à l'autre, d'un outil à l'autre, d'un article à une institution. Se souvenir qu'il n'existe pas de méthode infaillible et que chercher l'information sur Internet, c'est avant tout un état d'esprit. Ainsi, si je cherche le premier producteur de statistiques en Irlande, je peux commencer, sans trop de risques d'erreurs, par faire l'hypothèse que l'INSEE propose des liens vers ses homologues européens.

Faut-il commencer une recherche sur Internet ?

Internet est-il complémentaire à d'autres supports ou se suffit-il à lui-même ? . On trouvera rarement matière à une étude complète d'un sujet via Internet (test : essayez avec un sujet que vous connaissez bien = vous serez toujours très déçu). Par contre, bien (et rationnellement utilisé) le Web sera souvent plus rapide et moins cher que d'autres supports pour des recherches de type "questions-réponses".

Enfin, Internet et ses différents services (mail, newsgroups, mailing lists) se prêtent bien à la pratique de la veille, de part son caractère mouvant, décloisonné, international.

Les répertoires de recherche

Principe des répertoires de recherche

- ✓ "Collections" généralistes ou spécialisées de sites web classées par catégories organisées hiérarchiquement (au niveau mondial, on arrive à des systèmes de catégories très importants : quelque 300.000 pour Looksmart et 460.000 pour le Open Directory ; Nomade ("Tiscali Recherche") annonce quelque 10.000 catégories).
- ✓ Filtrage et classement " manuels " : la sélection peut être plus ou moins rigoureuse, avec une évaluation et une description des sites éventuellement enrichies.
- ✓ Pas d'indexation en texte intégral des pages des sites.
- ✓ Les répertoires généralistes mondiaux intègrent les fiches descriptives de 2 millions de sites web pour Yahoo, "plus de 4 millions" pour Looksmart et près de 3 millions huit cent vingt mille sites (400000 sites de plus en 5 mois) pour le Open Directory.
  Au niveau francophone, quelque 150000 sites sont répertoriés par Nomade et Yahoo (+ 10.000 en 6 mois), 65.000 sur les guides de Voila, de Lycos France ou de MSN, et pour environ 90000 sites francophones gérés par le Open Directory (+42 % en un an). (Nomade "reçoit" quelque 2000 soumissions par semaine et rejette 40 % des soumissions)
- ✓ Outils de première approche : Donnent une vue d'ensemble d'un domaine à l'utilisateur, qui peut ensuite naviguer à l'intérieur des sites indiqués pour aller plus loin.
- ✓ Ne gèrent pas les requêtes complexes, mais permettent généralement de faire une recherche par mot-clé sur une catégorie seule.
- ✓ Problèmes de mise à jour et de " désherbage ".

Modes de recherche

- ✓ Recherche dans le plan de classement : Cette méthode est parfois complexe, aucune norme n'existant pour l'arborescence des répertoires. Les sites sont indiqués par ordre alphabétique.
- ✓ Recherche par mot clé : la recherche se fait sur les champs suivants : intitulés des catégories, titres des sites, résumé des sites, adresses URL des sites. Avec ce mode de recherche, les résultats bénéficient généralement d'un classement de pertinence opéré uniquement sur les fiches descriptives des sites. Le Open Directory ne recherche pas sur les catégories.

Utilisation

Les répertoires sont à réserver pour des recherches plutôt thématiques, ou sur des mots clés assez généralistes ; notons toutefois que les catégories deviennent au fil du temps de plus en plus "pointues" en fonction du sujet.

Si l'on utilise des mots clés trop précis, ou trop de mots clés, la plupart des répertoires passent le relais à des moteurs de recherche partenaires (Google dans le cas de Yahoo) qui effectuent des recherches sur le texte intégral des pages web.

C'est pourquoi la distinction entre annuaires et moteurs est de plus en plus difficile à percevoir (cf "nouveau Yahoo" en .com et en .fr qui ne différencie plus les

résultats pages et sites mais donne des "web matches", qui proviennent de Google, mais reprennent la catégorisation de Yahoo s'il s'agit de sites). Mais elle reste néanmoins fondamentale.

Les répertoires sont aussi utiles :

- ✓ pour se faire une idée du vocabulaire utilisé dans un domaine (même en anglais, via Yahoo)
- ✓ pour retrouver, à partir d'un site web donné, d'autres sites traitant du même sujet
- ✓ pour trouver des sites fédérateurs ou portails spécialisés
- ✓ pour obtenir rapidement tous les sites d'une organisation importante.

LES PRINCIPAUX RÉPERTOIRES FRANCOPHONES ET INTERNATIONAUX GENERALISTES

(ordre alphabétique)

Répertoires	Internationaux	Français
About	www.about.com	
C'est trouvé (ex Eureka)		www.ctrouve.com (moteur inactif en 2003)
Looksmart	www.looksmart.com	www.looksmart.fr (arrêté)
Nomade		www.nomade.fr
Open Directory	http://dmoz.org	http://dmoz.fr
Virtual library	www.vlib.org	
Voila (Guide)		recherche.wanadoo.fr ou guide.voila.fr
Yahoo	www.yahoo.com	www.yahoo.fr

Important : De nombreux autres portails intègrent bien entendu ces répertoires

Disparitions récentes (depuis 2001) :

- NBCI (ex Snap) disparaît en tant qu'annuaire. C'est désormais Overture qui est utilisé par la chaîne américaine.
- Disparition du répertoire sélectif Alpha Search

Actualités :

- Lancement de Looksmart France : Looksmart fournit son annuaire à MSN, Excite, AltaVista, iWon, AOL,etC. Looksmart a Google comme partenaire moteur. 12 mars 2003 : (source Enfin.com) = **12/03/2003 : "Looksmart France : mort et enterré"** (archivé)

Et oui, cette belle aventure de Looksmart France vient de prendre définitivement fin depuis peu. Le site qui restait en ligne comme témoignage de l'échec qu'il représente est désormais une redirection vers la version anglaise. A noter que Looksmart Angleterre ne se porte pas vraiment mieux depuis que le principal partenaire, BT, s'est retiré de l'affaire.

- Nomade modifie sa présentation pour présenter en premier les catégories pertinentes (cf Yahoo). La catégorie n'apparaît plus explicitement sous chaque site trouvé. Un lien "sites similaires" permet d'obtenir les sites classés dans la ou les mêmes catégories. A noter la présence envahissante des liens payés sur les pages de résultats pour les mots populaires (ex : voyage)

  Nomade choisit Fast comme partenaire moteur (après Inktomi, puis Google)

- Ctrouve, basé sur la soumission des éditeurs, référence actuellement plus de 200000 sites francophones. Nouvelles fonctionnalités : recherche par popularité, par visibilité, par visiteurs, recherche moteur/annuaire, Récemment modifiés, par disponibilité, par région, modifiés souvent, par univers. (ne semble plus fonctionnel cet été 2003, même si les adresses répondent)
- Yahoo.com (partenaire Google) ne différencie plus les sites et les pages web mais annonce des "web matches" (octobre 2002), suivi par Yahoo France en janvier 2003.

TYPOLOGIE DES RÉPERTOIRES

Les répertoires généralistes "classiques"

Répertoires ayant vocation à indexer tous les sites et qui n'effectuent une censure que sur la base de principes prédéfinis : sites manifestement illégaux, sites en construction totale ou sans contenu réel, sites personnels trop "personnels", etc. Des équipes dédiées appartenant à la société détentrice du répertoire enrichissent les catégories.

Citons Yahoo, Nomade, , Looksmart. Notons que le nombre de ces répertoires généralistes tend à diminuer (disparition de SNAP)

Les répertoires généralistes"contributifs" ou "ouverts"

Répertoires dont l'enrichissement est effectué par différentes équipes d'internautes, non intégrées à la société gérant le site. La responsabilité d'une ou plusieurs catégories est confiée :

- ✓ Soit à des experts rémunérés pour leur prestation : About.com travaille ainsi avec des spécialistes qui sélectionnent les sites pour leur thématique et sont chargées de l'animation de leur section. Celle-ci peut d'ailleurs être considérée comme une "méta-page" du domaine, voire un répertoire spécialisé. About se présente donc comme un annuaire de guides du web. Voir par exemple http://websearch.about.com qui représente l'un des points de départ incontournables pour la recherche d'information sur le Web. En septembre 2001, About.com supprime 300 des 750 guides de son catalogue et réoriente son activité vers le commerce électronique : "About is going to be much more based on what users need to know, rather than something for everyone" est-il dit à la Direction

- ✓ Soit à des internautes bénévoles dont la compétence dans le domaine couvert pour cette catégorie a été vérifiée. Ces internautes reçoivent alors les demandes de référencement de leur catégorie, décident ou non d'intégrer les sites, et le cas échéant, rédigent eux-mêmes la description du site : Ainsi, le Open Directory "racheté" en 1998 par Netscape qui propose des licences d'utilisation à d'autres acteurs du Web, tels Lycos (plus de 52000 éditeurs issus de 229 pays en 44 langues). Bien entendu, l'inconvénient d'un tel système réside dans une qualité inégale selon les catégories. Le Open Directory signale actuellement environ 100 000 sites francophones.

  A noter que le Open Directory fait des "émules", mais qui se rapprochent plus du modèle ci-dessus, avec une rémunération des éditeurs : exemple wherewithall.com (dont l'outil de recherche se situe aujourd'hui à l'URL www.xoron.com) ou bien Zeal.com, répertoire ouvert proposé par Looksmart et qui sert également à alimenter ses bases

- ✓ Soit à des centres spécialisés (universités, centres techniques, etc.) : Ainsi, la Virtual Library du W3C (World Wide Web Consortium) fut le premier catalogue de ce type du Web. On est renvoyé pour chaque thématique à une section spécifique sur le serveur du centre concerné.

Les répertoires sélectifs

Répertoires dont les gestionnaires mettent en place des critères de qualité précis et intègrent uniquement les sites répondant à ces critères : Exemples www.bonweb.com ou www.britannica.com (encyclopédie Britannica).

Les répertoires spécialisés, ou "méta-pages"
Répertoire dont les sites répertoriés relèvent tous d'un domaine ou d'un secteur particulier (le vin, le tourisme, le sport, les ressources humaines, etc.). Un répertoire spécialisé peut, par exemple, ne prendre en compte que les entreprises d'un secteur, ou les produits d'un domaine. Les répertoires spécialisés sont souvent la base d'un portail thématique ou "vortail" : Ainsi, Indexa intègre les sites web d'entreprises (et par extension, du monde professionnel : fédérations, presse, etc.). Attention à l 'exhaustivité, à la mise à jour et à l'aspect sélectif.

Exemples de méta-pages spécialisées sur nature de documents :

Usuels et référence	http://www.bnf.fr/pages/liens/
Personnes	http://www.nedsite.nl/search/people.htm
Recherche d'images en ligne	http://www.ebsi.umontreal.ca/jetrouve/internet/moteur4.htm
Thèses	www.theses.org
Site universitaires	www.braintrack.com
Statistiques	www.statistics.com
Presse généraliste	www.presseweb.ch
Presse scientifique	www.libs.uga.edu/science/fullalph.html
Bibliothèques	http://sunsite.berkeley.edu/Libweb/index.html (Monde) www.abf.asso.fr/sitebib (France)
Cartes géographiques	www.internets.com/smaps.htm
Administration française	www.service-public.fr

Exemples de méta-pages thématiques :

Médecine	www.cismef.org	CISMEF – CHU Rouen
Juridique	www.legifrance.gouv.fr www.conseil-constitutionnel.fr/signets/autressi.htm	Legifrance Cons. Constitutionnel
Collectivités	http://www.ait.asso.fr/Liens.htm	AIT
Economie	www.ccip.fr/rime	RIME (grandes écoles commerce)
Informatique	www.inria.fr/InfoWeb	Inria
Environnement	www.ulb.ac.be/ceese/meta/cdsfr.html	Université Libre de Bruxelles
Sciences sociales	www.sosig.ac.uk	

Les répertoires d'outils de recherche

Répertoires spécialisés dans le signalement de répertoires généralistes, de répertoires spécialisés, de moteurs de recherche généralistes, de moteurs de recherche spécialisés, de méta-moteurs, voire de portails. Ces répertoires proposent parfois un signalement géographique, comme Indicateur.com, Search Engine Collosus (www.searchenginecolossus.com) ou Ariane6 (www.ariane6.com/moteurs.htm)

Certains répertoires de ce type jouent également le rôle de méta-moteurs (exemple The Big Hub).

7alpha (www.7alpha.com) ; Beaucoup (www.beaucoup.com) ; Enfin (www.enfin.com) ; Finderseeker (www.finderseeker.com) ; Indicateur (www.indicateur.com) ; Metamonster (www.metamonster.com) ; Searchability (www.searchability.com) : Search Engine Guide (www.searchengineguide.com) ; Search Power (www.searchpower.com) ; The Big Hub (www.thebighub.com) ; Strategic Road (www.strategic-road.com) ; "Vite, tous les Outils" (Jean-Pierre Lardy) (www.adbs.fr, rubrique Recherche d'information ou URFIST le Lyon : http://urfist.univ-lyon1.fr/risi/risi.htm) …..ETC ETC…

Signalons les répertoires "académiques" les plus connus de méta-pages (en-dehors de la Virtual Library, déjà citée) :

Strathclyde University, Ecosse : Bubl Link : bubl.ac.uk/link

Internet Public Library (University of Michigan) : www.ipl.org

Université de Göttingen : http://www.sub.uni-goettingen.de/ssgfi/

Library of California : Librarian's index to the Internet : www.lii.org

En français, voir notamment la selection de Sciences-Po Paris : http://www.sciences-po.fr/docum/ebibliotheque/index.htm, et de la BNF

Un répertoire à la loupe : Yahoo

Présentation

- ✓ Plus de 8000 soumissions par jour sur Yahoo US (700 en France). Deuxième (assez loin derrière Google) dans le palmarès des outils utilisés.
- ✓ Recherche possible dans les dépêches d'agences (Reuters, Cyperus, AFP, AP, etc.). On peut chercher directement dans l'actualité via http://news.yahoo.com ou bien en France http://fr.news.yahoo.com Finance : http://fr.finance.yahoo.com
- ✓ Recherches possibles dans une sous-catégorie
- ✓ Les sites ou catégories appartenant également à d'autres catégories sont repérées par un @
- ✓ Yahoo France présente d'abord les catégories concernées par la recherche, puis les sites web. Mais depuis peu, seules les premières catégories concernées apparaissent, pour laisser la place aux sites sur la première page de résultat. Yahoo.com ne différencie plus les résultats sites et pages.
- ✓ Partenaire moteur : Google

Syntaxe

- ✓ Opérateur ET implicite (pour plus d'options, passer en recherche avancée) Utilisation possible du +, du – et des " "
- ✓ Troncature automatique (sauf pour les mots courts), mais possibilité de troncature à droite avec *
- ✓ Limitations de champs à l'URL, taper u :nom à rechercher ex u :danone Limitation au titre, taper t :terme à rechercher ex t :optronique
- ✓ Les majuscules et minuscules ne sont pas distinguées
- ✓ Yahoo gère quelques synonymies dans son système de recherche.

A noter…

- ✓ La catégorie Commerce et Economie / Sociétés qui liste les sociétés par secteur d'activité
- ✓ On trouve pour certains domaines la sous-catégorie "Annuaires et guides web" qui répertorie des sites portails ou répertoires spécialisés.
- ✓ "Saut" possible de Yahoo France à Yahoo US à partir d'une catégorie.
- ✓ Yahoo prend en compte la popularité d'un site lorsque le moteur est utilisé (et ne renvoie donc pas une liste par ordre alphabétique dans ce cas)
- ✓ Fusion avec le site eGroups, (listes de discussion ou "Yahoogroupes") : http://groups.yahoo.com
- ✓ Le rachat de Inktomi par Yahoo devait être finalisé début 2003. On peut donc s'interroger sur la pérennité du partenariat avec Google, même si, à la rentrée 2003, c'est toujours Google qui motorise cette partie du répertoire.

Les moteurs de recherche

Principe des moteurs de recherche

Un moteur de recherche est un outil automatique constitué de plusieurs éléments :

1.Robot d 'exploration (spider) : collecte du contenu de millions de pages web dans une base de données structurées en champs (texte de la page, titre de la page, URL). Ces pages sont stockées dans un index qui se rafraîchit à la vitesse des visites du robot.

2. Indexation automatique : l'index de la base de données contient tous les mots significatifs des pages visitées par le robot.

3. Interrogation de l'index : l'utilisateur rentre un ou plusieurs mots clés. Chaque page contenant au moins une fois l'un de ces mots est considérée comme une réponse pertinente.

Attention : les moteurs indexent rarement toutes les pages des sites visités : par exemple AltaVista a mis en place une "limite de taille" d'environ 400 pages par sites. De plus, toutes les pages ne seront pas prises en compte en même temps.

La mise à jour de l'index est variable et peut prendre de un jour à quatre semaines. Plusieurs moteurs s'orientent actuellement vers une mise à jour "partiale" en travaillant d'abord sur les sites les plus populaires et les plus mouvants. De façon générale, les moteurs travaillent aujourd'hui plus sur la représentativité que sur l'exhaustivité de leur index.

La plupart des outils indexent également les méta-données,

Les principaux moteurs français et internationaux

(ordre alphabétique)

Moteurs	Internationaux	Français
Alta Vista	www.av.com	www.altavista.fr
AOL	www.aol.com	www.aol.fr(technologie Exalead)
Exalead (sur le Open Directory)	www.exalead.com	
Excite		www.excite.fr (Fast)
Fast	www.alltheweb.com	
Google	www.google.com	www.google.fr
Hot Bot (résultats Fast)	www.hotbot.lycos.com	www.hotbot.fr
Lycos (résultats Fast)	www.lycos.com	www.lycos.fr
Mirago	www.mirago.com (UK)	www.mirago.fr
MSN (résultats Inktomi)	search.msn.com	search.msn.fr
Teoma	www.teoma.com	
Voila		www.voila.fr
Wisenut	www.wisenut.com	

Disparitions récentes (depuis 2001) :

- Infoseek, Ecila, Excite (en tant que technologie moteur), Webtop (Dialog), Lokace et Northern Light

Ce 14 juillet 2003, **Yahoo! a annoncé le rachat d'Overture pour 1,63 milliard de $**. La société Overture est leader des liens sponsorisés et promotionnels* et a elle-même racheté en février le moteur Altavista et la division Web Search de FAST, l'éditeur du moteur AlltheWeb. De son côté, Yahoo! a finalisé en mars l'acquisition d'Inktomi afin de posséder ses propres technologies de recherche. En effet, jusqu'ici, le moteur utilisé par Yahoo!, c'est Google, un partenaire encombrant qui finalement lui capte et "vole" de nombreux clients et internautes.

(source : C. Asselin, Intelligence Center : http://c.asselin.free.fr/french/juillet03/yahooverture.htm)

Quelques chiffres sur les moteurs

- ✓ **Estimation du nombre de pages indexées par chaque moteur**

Sorce : Searchengine Showdown – Greg Notess
http://www.searchengineshowdown.com/stats/sizeest.shtml

Sorce : Search Engine Report Déc 2001 www.searchenginewatch.com

Search Engine	Showdown Estimate (millions)	Claim (millions)
Google	3,033	3,083
AlltheWeb	2,106	2,112
AltaVista	1,689	1,000
WiseNut	1,453	1,500
Hotbot	1,147	3,000
MSN Search	1,018	3,000
Teoma	1,015	500
NLResearch	733	125
Gigablast	275	150

Data from: Dec. 31, 2002

Based on AlltheWeb reported size and percentages from relative size showdown

AlltheWeb: 2,106,156,957 reported

Fast (Alltheweb), Google se livrent une bataille acharnée pour ravir la première place, Alta Vista restant actuellement plus loin derrière.

Le langage de recherche des moteurs : les options "standard" (Rappel).

- ✓ Opérateurs inclusifs et exclusifs (+ et -)
- ✓ Troncature : *
- ✓ Expression : " "
- ✓ Limitation par langue

Les outils disposent aussi d'une interface de recherche guidée ("plus d'options", "recherche avancée" "power search", etc.) qui évite de connaître le langage d'interrogation et permet d'exploiter simplement différentes options.

Attention à la recherche d'avancée de Alta Vista qui exige une syntaxe différente de la recherche simple (cf fiche moteur Alta Vista)

AVANTAGES ET INCONVÉNIENTS DES MOTEURS

- ✓ Gestion de recherches complexes (par opposition aux annuaires)
- ✓ Réponse à des recherches très précises
- ✓ Manque d'exhaustivité
- ✓ Les algorithmes de pertinence développés ne pallient pas les limites d'une indexation souvent "basique en texte intégral" = bruit
- ✓ Pas d'accès au "Web invisible" (voir le chapitre spécifique)
- ✓ Pas très performants en recherche sur autre chose que du texte (images, sons…)
- ✓ Lenteur de rafraîchissement de l'index (environ 4 semaines) donc pas efficaces pour des recherches sur l'actualité.

QUELQUES IDÉES REÇUES SUR LES MOTEURS

- ✓ Il existe des centaines de moteurs… FAUX : Il existe en fait de nombreuses interfaces "opérant" sur les mêmes bases. Une société comme Inktomi propose des licences de ses bases à de multiples outils (Lycos utilise ainsi conjointement Fast et Inktomi)
- ✓ "Je cherche une page que j'ai vue sur le web il y a un an"… Les moteurs de recherche n'archivent pas les documents qui ont été modifiés ou qui ont disparu: ce n'est pas parce que vous avez vu une page un jour sur le web que vous la retrouverez forcément. A noter que Google propose toutefois d'obtenir la page telle qu'elle était lorsqu'elle a été visitée par le robot (environ une fois par mois = option "en cache") (solution de dernier recours = la Wayback Machine de www.archive.org).
- ✓ Quand vous interrogez un moteur, vous scrutez le web en temps réel"… FAUX : vous interrogez l'index d'une base de données.
- ✓ "On ne sait jamais quelles fonctionnalités sont disponibles sur un moteur"… FAUX : les aides en ligne (help, tips) sont généralement bien rédigées.
- ✓ "If you've found it once, you'll find it again"… FAUX : la plupart des moteurs changent, les algorithmes de pertinence varient, et peuvent donner des résultats très différents (voir la notion de "Google Dance" mensuelle). Les pages disparaissent, évoluent. On n'utilise pas exactement la même requête.

PRINCIPAUX CRITÈRES DE COMPARAISON DES MOTEURS DE RECHERCHE

- ✓ Provenance de l'index, taille de l'index, ressources prises en compte
- ✓ Délai moyen de rafraîchissement et conditions de mise à jour
- ✓ Mode d'indexation et traitement éventuel des ressources (linguistique, statistique, parsing : extraction des éléments signifiants)

- ✓ Options de recherche simple et avancée, aide à la reformulation des questions.
- ✓ Critères déterminants pour le classement des résultats
- ✓ Présentation des résultats : informations disponibles, source du résumé, datation des résultats, regroupement des pages d'un même site (cluster), mise en exergue des mots-clés sur la page, archive de la page, cartographie, etc.
- ✓ Critères subjectifs : interface de consultation, adéquation aux types de recherche effectués.

LE TRI DE PERTINENCE DES MOTEURS

Principes

Les moteurs mettent au point des "tris de pertinence" pour classer de façon automatique leurs résultats de recherche, afin de présenter en début de liste ceux qui obtiennent le meilleur score pour une requête donnée. Les algorithmes de tri sont différents en fonction des outils et plus ou moins performants et complexes. Ils ne sont généralement pas connus de façon précise et varient dans le temps pour chaque moteur. Les principaux critères utilisés sont les suivants :

- ✓ ***Par rapport à la requête de l'internaute*** **:**
  - position des mots dans la requête : Ainsi, sur Alta Vista et Google, l'ordre des mots de la question n'est pas neutre.
  - correspondance d'expression : similarité entre l'expression de la requête et l'expression correspondante dans un document
- ✓ ***Par rapport aux pages de résultats***
  - "densité" des mots-clés : nombre d'occurrences du (des) terme(s) demandé(s) / nombre de termes de la page en question, une fois éliminés les mots vides.
  - présence dans le titre ou dans le premier tiers de la page
  - mise en exergue du texte (gras, taille des caractères)
  - présence dans les méta-données* (ce critère tend à perdre de son importance). Des outils comme Google ou Fast n'utilisent pas du tout ce critère, et Voila ne leur donne plus beaucoup d'importance.
  - présence dans l'adresse de la page
  - proximité des mots-clés sur la page
- ✓ ***Par rapport à la base de données du moteur :***
  - rareté des mots (déterminé par le nombre d'occurrences du mot dans l'index) : des mots rares dans une requête ont une pondération plus importante que des mots communs
  - popularité des pages : indice de clic (basé sur l'audience) ou indice de popularité (basé sur le principe de citation).

La popularité comme mesure de pertinence

Depuis deux ans, on a assisté à la naissance, au développement, puis au franc succès de deux nouvelles mesures de pertinence appelées respectivement "indice de clic" et "indice de popularité". Ces mesures s'ajoutent le plus souvent à d'autres "ingrédients" pour classer les résultats des moteurs, mais ils constituent aussi le critère de tri primordial des nouveaux venus inventeurs de ces technologies. Ces nouveautés, issues du "filtrage collaboratif", sont symptomatiques d'un certain désarroi des acteurs et utilisateurs du

réseau face aux multiples difficultés d'un recueil rapide d'informations pertinentes.

- ✓ **L'indice de clic**

Il s'agit ici d'analyser le comportement des internautes posant la même question au moteur et de privilégier dans le classement les pages les plus "cliquées", et sur lesquelles le temps passé est le plus important. Il permet donc de classer les résultats des requêtes les plus populaires, en récupérant le jugement implicite de communautés d'usagers. Fonctionne donc en "tâche de fond" sur un moteur existant, la base s'enrichissant ainsi.

Direct Hit (www.directhit.com), racheté par Ask Jeeves en 2001, puis devenu Teoma, est la référence dans ce domaine et est utilisé par de nombreux moteurs comme Lycos et MSN (plus de 50 sites clients), mais aussi Ask Jeeves. Alta Vista et Inktomi ont développé leur propre système sur un principe similaire.

Un système de positionnement payant "DirectHit Network" permettra d'acheter un positionnement dans les résultats de Direct Hit. A noter que (février 2002), Ask Jeeves envisagerait d'arrêter cette année le site web consacré à Direct Hit pour centraliser ses efforts de développement sur le moteur Teoma. Il conserverait la technologie pour la proposer à ses clients, mais le site serait fusionné avec celui de Teoma.

La technologie Global Brain (www.globalbrain.net) est très proche et fut notamment mise en œuvre sur le répertoire NBCI avant son arrêt récent.

Le défaut de l'indice de clic reste de privilégier fortement les sites "installés" et qui ont des moyens publicitaires importants, au détriment des "petits nouveaux". Il ne faut toutefois pas nier l'ingéniosité du principe, ni les services que ces outils peuvent rendre.

Pour savoir si les pages ramenées par un moteur sont issus de Direct Hit, il faut scruter attentivement le bas de la page de résultats. Si tel est le cas, la ligne "powered by Direct Hit" apparaît.

- ✓ **L'indice de popularité**

On s'intéresse ici aux "backlinks" ou "liens à l'arrivée", c'et à dire au nombre et à la qualité des liens pointant sur une page : on mesure ainsi sa popularité, et donc selon les concepteurs de ces technologies, sa pertinence. Les anglophones disent pour mieux expliquer le principe de l'indice de popularité :"It's not what you know, it's who knows you". En d'autres termes, le plus important n'est pas ce que vous dites ou ce que vous savez, mais qui vous connaît.

Le principe, rendu célèbre par le moteur Google, n'est pas totalement nouveau. Ne mesure-t-on pas la crédibilité d'un auteur scientifique au nombre de citations qui sont faites sur ses articles ?

Google examine la structure des liens sur l'ensemble du web. Quand on fait une recherche, un URL avec un fort "page rank" a plus de chance d'être listeée en premier. Chaque page de l'index de Google est notée : le "page rank" est une propriété de la page en elle-même, indépendante des requêtes effectuées : elle équivaut à la probabilité quel 'internaute aboutisse à cette page sur Internet.

Définition formelle : Soit A une page du web et T1...Tn les n pages citant A. Soit C(X) le nombre de liens pointant en dehors de la page X.Soit d la probabilité qu'a l'internaute virtuel de changer de page au hasard (souvent mis à 0.85). Alors Le PageRank de A est Pr(A)=(1-d)+d(PR(T1)/C(T1)+...+PR(Tn)/C(Tn))
Si A était une page contenant tout le web alors le PageRank de cette page serait de 1. Le PageRank forme une distribution probabiliste sur l'ensemble des pages du web.

Le tri des résultats pour une requête intègre d'autres critères plus classiques, dont bien entendu la présence des termes de la requête dans les pages de résultat, ou identifiée comme pertinente via l'analyse du contexte des liens.

Le grand avantage du système est de donner une meilleure visibilité aux sites incontournables du domaine de recherche. L'inconvénient majeur est là encore, de pénaliser les nouveaux venus peu connus.

A noter

- ✓ Alta Vista n'applique pas son algorithme de pertinence en mode avancé, mais propose de trier les résultats en fonction du ou des mots saisis dans le champ "Trier par" (sort by).

Complément d'information :

- ✓ Sur le ranking de différents moteurs, voir entre autres la page suivante *www.user.cityline.ru/~asona/secret/searchengine5.html*
- ✓ Sur le site Abondance (www.abondance.com), Olivier Andrieu donne pour chaque moteur présenté l'importance relative des différents critères (peu détaillé toutefois).

Les sociétés spécialisées dans **le référencement** cherchent bien entendu à connaître le plus précisément possibles les critères clés de chaque moteur (cf paragraphe précédent). L'objectif est de faire apparaître en bonne position (ranking) les pages web de leurs clients sur les listes de résultats à une requête comportant certains mots-clés.

Ce travail de référencement se fait parfois au mépris de l'éthique et donne lieu à une activité de "spamdexing" ou "spamming". (Création ou modification d'un document avec l'intention de tromper un catalogue ou un système de classement électronique. Toute technique qui a pour objectif d'augmenter la position potentielle d'un site aux dépens de la qualité de la base de données du moteur de recherche. Définition issue du glossaire réalisé par les membres francophones de la liste de diffusion I-Search Digest hébergé par le fournisseur d'hébergement IDF www.idf.net/mdr/glossaire.html).

Cette activité a notamment amené à la baisse d'importance rapide du critère de présence des mots-clés dans les méta-données.

LE REFERENCEMENT PAYANT (SOURCE : ABONDANCE.COM)

le référencement payant devient une norme, au moins pour l'étude des dossiers. L'extrême est bien sur Overture, où les mots-clés sont "mis aux enchères" :

- ✓ **La soumission payante**

  La soumission payante est proposée principalement par les annuaires. Elle permet de voir son site rapidement évalué (en quelques jours) par les netsurfers et d'obtenir une réponse par mail, que celle-ci soit négative ou positive. Il ne s'agit en rien d'une garantie d'inscription dans l'annuaire. Le responsable du site paye uniquement pour être sûr que sa source d'information sera visualisée rapidement et qu'il sera averti dans la foulée de la décision de l'outil de recherche. Vous pouvez donc tout à fait payer pour voir votre site refusé. Cependant, il semblerait que le pourcentage de sites acceptés par ce biais soit assez important outre-Atlantique. Le fait que Yahoo! ait rendu cette procédure obligatoire dans un certain nombre de catégories semble également montrer qu'elle est plutôt bien ressentie par les propriétaires des sites. Ex : Yahoo, Looksmart, NBCI…

- ✓ **Le référencement payant**

Le référencement payant garantit la présence d'un certain nombre de pages d'un site dans la base de données d'un moteur de recherche, ainsi qu'un rafraîchissement de ces documents dans des délais courts et garantis. Il ne s'agit en rien, dans ce cas, d'une garantie de positionnement sur tel ou tel mot clé, mais uniquement du fait que le webmaster saura que sa (ou sa "collection de") page sera joignable par l'internaute, qu'elle sera présente dans l'index. Côté moteur, il s'agit également d'un moyen de lutter contre le spam et de prendre en compte des pages dynamiques (accessibles à l'aide d'une url "exotique", c'est-à-dire contenant un "?" dans leur intitulé) puisque les clients sont, cette fois, dûment enregistrés. Aucun soumission anonyme n'est possible. Ex : Altavista, Inktomi, Fast…

- ✓ **Le positionnement payant**

Le **positionnement payant** permet, pour sa part, d'obtenir une page web de votre site dans les X premières positions de l'annuaire (sur une catégorie ou une saisie de mot clé) ou d'un moteur de recherche (pour un mot clé donné). Ex : Yahoo Sponsored sites. Offres de positionnement publicitaire des outils de recherche grâce à l'achat en ligne pour certains mots-clé ex : les premiers résultats d'Overture sont présentés en tête de la page de résultats sur AOL, Netcape, MSN, Yahoo US ou Lycos US. ("sponsored sites", "feature links"). Et Les résultats de Espotting apparaissent sur les outils français Lycos, Altavista, Hotbot, Nomade depuis janvier 2002 et bientôt Yahoo,. Copernic a signé un accord pour 5 liens maximum, en fonction des mots-clés demandés

Attention, depuis peu, des sites payent pour apparaître en première position lors de recherches sur des marques de notoriété mondiale.

Les campagnes de référencement payant peuvent atteindre des sommes très importantes. Ainsi, Kelkoo pour son lancement en Grande-Bretagne a investi 1 million d'euros en achat de mots-clés auprès d'Espotting. Si le montant est comparable à celui d'une campagne publicitaire, son efficacité est en revanche mesurable. Les connexions sur le site de Kelkoo en Angleterre ont été multipliées par 8 durant le premier mois de cette campagne de référencement payant. (01 net)

Selon Merrill Lynch, le marché de la vente de mots-clés atteindra 400 millions de dollars aux États-Unis en 2002.

.

. Autre solution

Google propose le système "Adwords" : le client crée un encart qui apparaîtra en haut et à droite des pages de recherche de Google en fonction de mots clés choisis par le client. Google propose également l'offre Premium Sponsorship (résultats sponsorisés) : le client apparaît en haut de page et avant les résultats de la recherche. Les formats et design sont gérés par Google

Pour le détail sur des tarifs, voir :

www.edantza.com/ressources-referencement/offres_referencement_payant.html

www.referencement.ch/referencement-payant.html.

LES MOTEURS SPÉCIALISÉS

Ils sont encore peu nombreux, car font rapidement appel à des technologies complexes.

Certains font une indexation en texte intégral des pages d'une sélection manuelle de sites web (exemples Medical World Search en médecine www.mwsearch.com, ou encore LawCrawler du site Findlaw dans le domaine juridique http://lawcrawler.findlaw.com)

D'autres catégorisent automatiquement des pages web, tel Voila avec sa recherche thématique (à tester par exemple avec le mot "bilan" dans la thématique "comptabilité").

Exemple : Polymer search on the Internet (plastiques, caoutchouc, polymères) www.polymer-search.com moteur de recherche en texte intégral sur le contenu web de nombreux sites. L'éditeur Rapra sélectionne les nouveaux sites à indexer selon des critères précis (quantité et qualité du contenu).

www.netsearcher.com : recherche sur sélection de sites internet.com

Scirus : www.scirus.com : moteur spécialisé et méta-moteur spéciaslisé : utilise des sites web d'accès libres indexés en profondeur par le moteur de recherche Fast. Seuls des sites à contenu scientifique validé sont sélectionnés et intégrés. + bases de données.

www.newenergy2b.com : recherche sur 200 sites.

Moteur Etat-Partenaires de l'ADIT sur 300 sites publics. www.adit.fr

Moteur sur les sites éducatifs du CNDP : www.cndp.fr/spinoo/

Les moteurs principaux à la loupe

Google a la loupe

Google est la star actuelle des moteurs de recherche (50 % du trafic de la recherche francophone d'après le baromètre Xiti/1ère Position du mois d'avril 2002, loin devant Yahoo France avec 16,4 %). Google utilise aujourd'hui 8000 serveurs Linux connectés en réseau. Il aurait un index de 3 milliards de documents (pages web, archives des forums de discussion et images) dont les ¾ sont réellement disponibles en tant que docs web indexés en texte intégral.

- ✓ Google utilise des algorithmes analogues à ceux des autres moteurs, mais donne davantage d'importance à un critère très intéressant : la «popularité» des pages web. Google calcule en effet l'importance d'une page en fonction du nombre de liens qui, à partir d'autres sites, pointent vers cette page. L'importance probable des sites où se trouvent ces liens est également prise en considération, et elle est évaluée de la même manière. (cf page)
- ✓ **Syntaxe**: L'opérateur par défaut est ET. Google accepte les guillemets. On peut éventuellement utiliser l'opérateur OU. L'opérateur + est néanmoins parfois nécessaire pour forcer le moteur à prendre en compte un mot très courant ("mot vide", ou "stop word" en anglais). L'opérateur – fonctionne. Il n'y a pas de troncature, et la recherche se fait sur la chaîne de caractère indiquée (au singulier si indiqué au singulier), mais sans tenir compte des accents.

  Recherche dans le titre des pages (fonction intitle: ainsi que allintitle:) et dans l'url (inurl: ainsi que allinurl:)

  link:www.monsite.com visualise les pages "pointant" vers "monsite".

  related: www.monsite.com/page.html visualise les pages liées.

  particulier site:www.monsite.com cherche le mot clé "particulier" sur le site Mon site.

  filetype:pdf retrouve les fichiers de type pdf.

  filetype:ppt –site: gihweihtghwhg : 257000 docs

  filetype:asp retrouve les fichiers ASP (Active server pages de Microsoft),

  Fonctionnement identique depuis novembre 2001 pour les documents word (doc) excel (xls), powerpoint (ppt) et RTF (rtf) désormais indexés.
- ✓ ne permet pas la troncature comme dans la majorité des outils, mais permet de remplacer un mot : exemple "trois * chats" va ramener des phrases comportant "trois petits chats", "trois gros chats", etc…
- ✓ **La recherche avancée** permet de retrouver ces fonctions via des menus déroulants et permet également d'inclure seulement (ou d'exclure) les pages provenant d'un site ou d'un domaine. En anglais, une recherche est possible sur le titre ou l'URL.
- ✓ Recherche proposée dans les résultats (revient à rajouter des mots clés à la première équation).
- ✓ Dans les résultats, La ligne de "description" des pages met en situation les mots-clés (habituellement, c'est la première ligne de la page ou la méta-donnée description qui est utilisée)

- ✓ Google conserve une copie (lien "cached" ou « archivé en mémoire » dans les réponses) des pages qu'il a indexées. Ainsi, si la page a été modifiée, a disparu ou si elle a changé d'adresse, il est tout de même possible de la consulter.
- ✓ Google indexe 22 millions de fichiers pdf (voir"trucs et astuces : comment identifier des pages pdf sur le web").
- ✓ Le moteur utilise le répertoire Open Directory, mais reclasse dans chaque rubrique par popularité via son système.
- ✓ Google propose comme beaucoup une barre d'outils à télécharger et qui s'intègre au navigateur. Les + de l'outil : le "page rank" ou taux de popularité de la page de Google en direct ; la recherche sur le site dont provient la page web visitée.
- ✓ Google a racheté en février 2001 l'outil de recherche sur les forums Deja, et propose aujourd'hui les archives des forums depuis 1995, soit 650 millions de messages.
- ✓ Depuis juillet 2001, le moteur permet la recherche sur les dates en recherche avancée : il n'y a pas, comme dans Alta Vista, moyen de configurer précisément sa requête, mais on peut néanmoins choisir d'effectuer une recherche sur les trois derniers mois, les six derniers mois ou l'année précédente.
- ✓ Google est partenaire de Yahoo et de Nomade (quand la requête sur les annuaires ne donnent rien).
- ✓ Recherches d'images (http://images.google.com) : 250 millions de fichiers sont aujourd'hui proposés, ce qui fait de Google un redoutable challenger pour les autres moteurs de recherche d'images.
- ✓ Google a racheté Outride, spécialisée dans la mise en place d'algorithmes de pertinence dans le domaine de tri de l'info disponible en ligne et émanation du Xerox Palo Alto Research Center.
- ✓ Google a lancé une page "News and resources" (news.google.com) sur les dernières nouvelles et dépêches d'actualité dans 7 catégories distinctes. Les infos sont issues de 100 sites d'info en langue anglaise (New York Times, CENT, News.com, Reuters, etc.). les dépêches sont mises à jour toutes les heures, et pour l'instant, aucune recherche dans les archives n'est disponible.
- ✓ Février 2002 : mise en place d'un nouveau système pour les liens achetés (à côté du système classique base sur le CPM): "adwords select" où l'annonceur ne paye que si le lien est cliqué
- ✓ Avril 2002 : Google lance les "Google web APIs" boîte à outils pour les programmeurs qui peuvent ainsi utiliser gratuitement (pour usage non commercial) l'index de Google pour leurs applicatifs. Mapstan a utilisé cette fonctionnalité pour cartographier les résultats fournis par Google sur une requête par mots-clés : search.mapstan.net
- ✓ Voir aussi www.touchgraph.com/TGGoogleBrowser.html pour voir la représentation graphique des liens pointant vers un site.
- ✓ Recherche sur les tendances et faits marquants : www.google.fr/press/zeitgeist.html
- ✓ A noter début 2003 650570 sites visibles sur le web francophone sur Google.

All The Web a la loupe

- ✓ Lancé en 1999 par la société norvégienne Fast suite à un accord avec Dell Computer Corporation : Moteur très rapide avec l'un des plus gros index de pages web actuellement (625 millions de pages).Il est utilisé également par Lycos et Spray
- ✓ Choix d'une langue de recherche
- ✓ Clustering pour les résultats. Un clic sur "more hits from" relance une recherche de All The Web sur le seul site concerné.
- ✓ En recherche simple, à noter les parenthèses pour signifier le OU (un seul niveau de parenthèses possible) :+sécurité +(voiture automobile).

  url.tld:nomdedomaine trouve les pages à l'intérieur d'un domaine spécifique (url.tld:fr trouve les pages du domaine France)

  url.host:urld'unsite) trouve les pages d'un site spécifique (url.host:www.adbs.fr)

  link.all:urld'unsite trouve les pages ayant un lien avec l'adresse indiquée (link.all:www.adbs.fr trouve les pages pointant vers www.adbs.fr)

  normal.title:texte trouve les pages contenant le mot ou la phrase dans le titre

  url.all:texte : trouve les pages avec un mot ou une phrase dans l'URL

  url.domain:text : trouve les pages avec le mot ou la phrase spécifiée dans le nom de domaine

  Pas de troncature
- ✓ En recherche avancée, la plupart des fonctionnalités sont disponibles via une interface guidée
- ✓ Recherches d'images, de clips video, de fichiers MP3 ou FTP (par défaut, recherche sur les sites web, mais on peut naviguer dans les résultats des différentes sources.
- ✓ Est rentré dans le capital de la société Albert (langage naturel)
- ✓ Rafraîchissement annoncé de l'index de 9 à 12 jours, mais l'ensemble du web n'est pas crawlé en un cycle.
- ✓ Service de news (3000 sources, avec mise à jour toutes les deux heures) ; Classement dynamique des 200 premiers résultats (utilisation des catégorie du Open Directory si l'une ou plusieurs correspond, sinon, création d'une nouvelle : voir en haut de la page dans une fenêtre "beta fast topics") ; Outil de pré-analyse qui permet de voir comment le moteur a traduit la requête ("your query was rewritten into").
- ✓ Mars 2002 : Fast va signer avec le site FirstGov (www.FirstGov.gov), le portail Internet du gouvernement américain, pour l'équiper de ses technologies de recherche d'information en remplacement d'Inktomi (51 millions de docs disponibles avec formats hétérogènes (bases de données, html, pdf..)
- ✓ Recherche dans les documents pdf et flash annoncée en septembre 2002 (voir recherche avancée File format), dans les docs word en décembre
- ✓ Décembre 2002 : Amélioration de l'algorithme de pertinence, notamment sur les requêtes de deux mots clés et plus (prend notamment en compte la proximité des mots et classe en premier l'expresssion
- ✓ Janvier 2003 : Interface recherche avancée revue et prise en compte de nombreux opérateurs (page aide très claire).

ALTA VISTA A LA LOUPE

- ✓ Alta Vista appartient à la société CMGI depuis 1999 (après avoir appartenu à Digital depuis 1995, puis à Compaq depuis 98)
- ✓ Mise en ligne récente de la nouvelle version de son moteur de recherche : les changements touchent à la fois le look (plus épuré), les fonctionnalités et le tri de pertinence. L'index est d'environ 1,3 Milliards de page (après crawling de 4 milliards), et reste donc en-deça de Google et Alltheweb. On compte aussi quelque 400 millions d'objets multimedia avec une présentation "à onglets" qui ressemble à Google. Altavista annonce un rafraîchissement beaucoup plus performant avec une mise à jour quotidienne de la moitié des liens" ce qui semble très exagéré. On apprécie beaucoup toutefois la notice associée aux résultats " mis à jour dans les dernières 24 h (ou 48 h)" qui s'intéressent non pas à la rentrée de la page dans l'index mais à la mise à jour réelle de la page (plus performant que Google à ce niveau là).
- ✓ L'annuaire partenaire reste Looksmart, et les index nationaux et mondiaux regroupés dans un même index.
- ✓ Les majuscules et minuscules ne sont plus prises en compte.
- ✓ Accents pris en compte. La recherche d'un mot accentué ne ramènera que les mots avec accent.
- ✓ Pour le résumé, Alta Vista privilégie désormais un extrait textuel du document autour du (des) mot(s) demandés (cf Google et Voila)
- ✓ Indexation des fichiers pdf : Alta vista indexe le début des fichiers (jusqu'à 120 pages) comme Google. On peut utiliser la syntaxe de recherche filetype:pdf
- ✓ Traduction automatique
  Est proposée grâce au traducteur automatique Systran, et est opérationnelle de l'anglais vers le français, l'espagnol, l'italien, l'allemand, le portugais, et vice versa… Accès direct à babelfish.altavista.com
- ✓ Pour dépasser 200 réponses, remplacer la valeur qui suit "stq=" à la fin de l'URL
- ✓ Recherche de l'actualité sur 3000 sources, avec des sources telles Moreover, NY Times, BBC, Forbes (meilleure antériorité souvent que sur Google): news.altavista.com
- ✓ Remplacement de l'outil Worldpages.com par Superpages.com en tant que fournisseur de résultats pour les pages jaunes et les pages blanches
- ✓ Liens payants en provenance de Overture : "featured links" et "sponsored links".
- ✓ Programme "trusted feed' : référencement payant pour sites de + de 500 pages (cf web invisible car prise en compte de l'ensemble du site).

Syntaxe recherche simple :

- ✓ Recherche implicite :
- ✓ utilisation du +, du -, des " "., de la troncature avec * (seulement après 3 caractères)
- ✓ Limitations aux champs :

Titre	title:terme recherche
URL	url:terme recherché
Serveur	host:terme recherché
Domaine	Domain:domaine recherche

Fichier image	Image:terme recherché

- ✓ Recherche des pages liées à mon site link:monsite.com
- ✓ PRISMA : remplace l'ancienne "Related searches" Alta Vista garde en mémoire les requêtes saisies pour pouvoir vérifier la répétition d'un terme dans toutes les résponses à une requête. Propose ainsi des expressions contenant le mot demandé : Exemple : on tape « energy » et on se voit proposer « solar energy », « free energy », « alternative energy », etc.

  On peut, soit ajouter ces mots-clés à la requête initiale (en cliquant), soit la remplacer entièrement (en cliquant sur >>).
- ✓ Limitation par la langue : Cf menu des langues. Si on ne précisez rien, la recherche se fera sur des documents écrits en toutes langues (in any language).
- ✓ Cluster pour les résultats : "More pages from this site" pour voir les autres pages d'un même site, pertinentes par rapport à une recherche. En recherche avancée, on peut choisir de voir l'ensemble des résultats d'un même site à une recherche, ce qui est pratique.
- ✓ Raccourcis proposés pour certaines requêtes populaires et issus souvent du web invisible (cf partenariats avec bases de données majeures).

Syntaxe Recherche avancée

- ✓ Utilisation des opérateurs booléens AND OR AND NOT NEAR (moins de 10 mots d'écart entre les mots clés) ainsi que des parenthèses
- ✓ Pas d'application de l'algorithme de pertinence pour les résultats de la recherche en mode avancée. On peut appliquer le "sort by" pour classer les résultats obtenus.
- ✓ Limitation possible par la date (Du: Au:) C'est la date de dernière mise à jour des documents au moment de l'aspiration des pages par AltaVista qui sert de référence (ne garantit pas forcément la fraîcheur des infos). Si rien n'est indiqué dans le champ "AU", c'est la date du jour qui est prise en compte

Altavista France : Ne pas confondre le choix "web français" avec le menu déroulant des langues. Si on choisit "web français", la recherche par mots-clés va porter sur des pages d'origines française, sans tri par le domaine (pages proposées par des sociétés qui détiennent une adresse postale en France, mais sont peut-être en .com). A noter les index nationaux et mondiaux sont regroupés dans un seul et même index.

Nouveaux moteurs (2001–2002)

AOL.FR

- ✓ Lancé en avril 2002, à partir de la technologie Exalead et avec un index de 50 millions de pages pour le web francophone et la base d'Inktomi pour le web anglophone (environ un milliard de documents).
- ✓ Rapide compte tenu de la technologie statistique utilisée
- ✓ Proposition de mots ou groupes de mots les plus récurrents dans les documents trouvés
- ✓ Résultats classés en dossiers et sous-rubriques selon leur popularité et leur pertinence en rapport avec la requête de l'internaute.

EXALEAD (né en août 2001)

- ✓ La technologie Exalead (plate-forme complète d'acquisition, de traitement et de recherche) s'appuie sur une analyse statistique de l'ensemble des documents du corpus et des résultats d'une requête : Les rubriques et expressions les plus significatives sont présentées avec les premières pages de résultat à l'utilisateur. Celui-ci peut donc d'un clic sélectionner une option et relancer sa recherche en la précisant. Les rubriques ne sont pas générées automatiquement par l'outil, mais incorporées en tant que données structurelles de catégorisation du corpus (il peut s'agir d'un annuaire de sites web, de catégories d'un portail ou autre).
- ✓ La technologie de la société française Exalead est "cousine" de l'ancienne fonction "Refine" présente autrefois sur AltaVista.
- ✓ La démonstration de Exalead sur le Web se fait actuellement à partir de documents catégorisés par le Open Directory : 20 millions de documents pour une interrogation du web francophone, 100 millions sur le web mondial anglophone.
- ✓ A noter une fonction intéressante pour privilégier les documents contenant un mot optionnel sans pour autant éliminer ceux qui ne le contiennent pas. Exemple : "vache folle" ?crise
- ✓ Possibilité d'équations complexes : téléphone mobile/portable équivaut à téléphone ET (mobile OU portable)
- ✓ "crise de la vache folle" : reconnaissance des mots composés sans guillemets mais risque de ne chercher plus que ça (voir chemin d'accès).
- ✓ Troncature implicite dès qu'il y a deux mots de la requête, mais si plus de mots, pas de troncature implicite * OK
- ✓ Mars 2002 : Exalead propose une soluion pour les entreprises "Exalead Corporate" : moteur de recherche pour les sources d'infos de différents formats + prise en compte des méta-données + génération de mots-clés pour catégorisation et navigation dynamique. (à chaque étape d'une recherche, Exalead Corporate fournit à l'utilisateur les catégories statiques et dynamiques pertinentes lui permettant de naviguer dans les résultats de sa requête).
- ✓ Avril 2002 : lancement de aol.fr équipé de la technologie Exalead (voir ci-dessus)

- ✓ Utilisation de aol.fr sur Netscape (qui appartient à AOL)

TEOMA

- ✓ Ce nouveau moteur de la société Hawk Holdings , issu d'un projet né en 1998 à Rutgers University aux Etats-Unis, est actuellement en test sur le web. Teoma été racheté par Ask Jeeves très rapidement après sa sortie et fournit depuis janvier 2002 une alternative aux résultats fournis par le système de questions-réponses Ask Jeeves, remplaçant ainsi les résultats de Direct Hit, une autre propriété de Ask Jeeves. Il semble que Teoma devienne un élément majeur dans la stratégie d'Ask Jeeves, au détriment de Direct Hit, racheté en 2001 et qui pourrait être arreté cette année.
- ✓ Il annonce un index de 500 millions d'URL en janvier 2003 (200 millions de pages en avril 2002, puis de 350 millions en novembre 2002.)
- ✓ Teoma propose une page de résultats très riche et innovatrice, qui permet d'avoir des vues complémentaires de l'information réponse :
    - La partie gauche de l'écran renvoie "classiquement" des pages web répondant à la requête de l'utilisateur
    - Le haut de l'écran ("web pages grouped by topic) présente les grands sujets extraits dynamiquement des pages résultats : chaque catégorie peut être explorée en détail d'un clic. Cette fonction n'est toutefois guère exploitable pour les pages francophones.
    - La partie droite de l'écran ("expert's links") est dédiée aux "méta-pages" ou sites fédérateurs riches en liens si le moteur en trouve
- ✓ Pour répondre à une requête, l'outil commence par chercher classiquement dans son index les pages contenant les termes de recherche (ou considérées comme pertinente suite à l'analyse des liens). Puis, Teoma classe ces pages dans des ensembles cohérents grâce à l'analyse des liens (regroupements des pages pointant les unes sur les autres et choix des mots les plus communs). Enfin, un algorithme proche de celui de Google permet pour chaque set de documents, de retrouver les pages les plus populaires.

  Notons qu'à la différence de Google, qui attribue des "page-rank" généraux indépendants des recherches, le score attribué par Teoma est spécifique à chaque catégorie créée. Par ailleurs, contrairement à Northern Light, c'est l'analyse des liens qui permet d'établir des classifications.
- ✓ Opérateur ET par défaut, ou choix de "exact phrase". On peut aussi utiliser les + et – et les guillemets pour l'expression. Néanmoins, dans ce cas, l'interprétation de la phrase sera moins stricte que ci-dessus.
- ✓ Avril 2002 : Nouvelle version : propose des options supplémentatires : sites web similaires, liens sélectionnés par des experts (à partir des communautés identifiées automatiquement)
- ✓ Janvier 2003 : Version Teoma 2.0 Améliorations avec un vérificateur d'orthographe, les extraits pertinents pour les résultats, une recherche avancée en béta.

ILOR

Le moteur Ilor (www.ilor.com) utilise la technologie TEOMA avec des fonctionnalités intéressantes : Le passage de la souris sur un lien ouvre l'affichage d'un menu ("LORLinks Menu") permettant de sauvegarder les paramètres de la recherche, de le mettre en favoris, etc.

WISENUT

- ✓ Wisenut apparaît comme un challenger de Google en calculant la pertinence à partir des "backlinks"(analyse du texte des liens, des termes qui entourent ces liens et du contenu des pages contenant ces liens) et à partir de l'analyse du texte de la page. Racheté récemment par Looksmart (mars 2002, le moteur annonce 1,5 milliard de pages indexées, et Search Engine Showdown en donne 579 millions en février 2002).
- ✓ Autre fonction le rapprochant de Google : "Sneak a peek" pour voir une "archive" de la page, mais sans quitter la page de résultats
- ✓ Le moteur effectue une catégorisation automatique des résultats de la recherche dans des dossiers ("wiseguides") via des liens sémantiques avec les mots de la requête (cf Northern Light). On peut ouvrir la catégorie ou relancer une nouvelle recherche en utilisant la catégorie comme requête.
- ✓ Wisenut propose des liens pour des requêtes similaires
- ✓ Le moteur groupe les résultats par site et liste le nombre exact de pages d'un site définies comme pertinentes.

MIRAGO

- ✓ Ce nouveau moteur, à la technologie propriétaire, ne s'intéresse qu'aux pages françaises.
- ✓ Il permet de faire une recherche régionale : sélection possible d'une ville ou d'une région à partir de laquelle démarrer une recherche
- ✓ Possibilité de classer les résultats selon le nombre de liens pointant vers une page (popularité des pages) ou le nombre de liens contenus sur une page (page riche en liens) ou selon la date (documents les plus récents d'abord) ou selon que la page est riche en images ou non.
- ✓ Ramène les pluriels en singulier, formes verbales à l'infinitif et abréviations aux mots entiers, adverbes et synonymes à la racine du mot auquel ils se rapportent : "enfant adoptif" = "enfant adopté" ; "problèmes d'ado" = "problème d'adolescence"
- ✓ Supporte le langage naturel : option "meilleur résultat" en recherche avancée
- ✓ Donne en premier les noms de domaine contenant les mots de la recherche (résultats non numérotés).
- ✓ Option proche du "near" : choisir "mots liés" dans la recherche avancée. A noter aussi la possibilité de rechercher sur "la plupart des mots".
- ✓ Recherche par dates
- ✓ Recherche sectorielle proposée

Les méta-moteurs "on-line"

Présentation

(voir aussi www.metasearchguide.com)

Les méta-moteurs (parfois appelés méta-chercheurs) interrogent simultanément plusieurs moteurs de recherche et/ou répertoires et compilent les résultats avant de les présenter (élimination des doublons, parfois nouveau tri de pertinence).

Ils ne maintiennent donc pas eux-mêmes de base de données, et se contentent de transmettre la requête aux outils utilisés.

Avantages : ils sont efficaces et rapides pour une recherche du type "Question-Réponse" ou une recherche précise. Ils permettent par ailleurs de se faire rapidement une idée du "répondant" des moteurs à partir d'un ou deux termes de recherche ou citations exactes. Ils innovent beaucoup actuellement. ***A ne pas confondre avec les méta-moteurs clients ("off-line") du type Copernic.***

Les méta-moteurs "on-line" commencent pour certains d'entre eux à proposer un accès au Web invisible (Profusion, Search.com).

Inconvénients : Ils ne traduisent pas toujours les langages d'interrogations. Les recherches complexes génèrent beaucoup de bruit avec les méta-moteurs. Par ailleurs, ils ne sélectionnent souvent que les dix premières réponses fournies par les différents moteurs qu'ils mettent en œuvre. Pour être réellement efficaces, les utilisateurs des méta-moteurs devraient les paramétrer et dépasser la première page de résultats.

Avec la vague des liens payants, l'"indépendance" des méta-moteurs risque d'être sérieusement remise en cause, d'autant qu'il est souvent plus difficile de reconnaître ces liens dans leurs résultats que dans les outils d'origine. D'après une étude de mai 2001 de SearchEngineWatch (http://searchenginewatch.com/sereport/01/05-metasearch.html) pour certains outils de ce type, la moitié des résultats s'avéraient être payés. Voici les pourcentages de liens payés dans la source pour les méta-moteurs choisis pour l'étude

Dogpile 60 %, Mamma 56 %, Meta-Crawler 36 %, Search.com 33 %,

Ixquick 25 %, ProFusion 14 %, Vivisimo 0 %

Critères de choix des méta-moteurs : Outils et sources interrogeables, options de paramétrage, tri et présentation des résultats.

A noter : La plupart des grands outils de recherche de l'Internet se comportent aujourd'hui en fait comme des méta-moteurs, en interrogeant simultanément différentes bases de données (base répertoire, base pages web, base articles de presse ou dépêches, bases d'information sur les entreprises, etc.)

Parmi les plus puissants méta-moteurs du web..

(par ordre alphabétique) :

Dogpile : attention, beaucoup de liens achetés. (Appartient à Infospace, qui a racheté Excite)

Gogettem www.gogettem.com
Lance et ouvre en même temps les outils sélectionnés. Cette particularité peut être utile…

Kartoo www.kartoo.com

Un nouveau méta-moteur qui innove avec une interface graphique censée s'adapter à tout utilisateur, même novice : Les sites sont placés sur une carte thématique, et reliés par les termes les plus fréquents (analyse statistique sur le corpus de résultats). Au passage du curseur sur un site, sa description s'affiche. Si l'on passe sur un thème, deux boutons + et – s'affichent qui permettent d'ajouter le terme à la recherche ou de l'éliminer.

Avril 2002 : sorite de la version 2 : la taille des boules qui représentent les sites sont variables en fonction de la pertinence

Mapstan search.mapstan.net

"Méta-moteur de recherche et de capitalisation de connaissances". A partir d'une technologie propriétaire de cartographie de l'information personnalisée, de filtrage collaboratif, et analyse de corrélation (brevet déposé en décembre 2000).. Présentation des résultats sur un "plan de quartier" où les pages sont regroupées par sites qui sont reliés par des "rues" indiquant leur similarité. Il indique également les pages les plus pertinentes des recherches similaires (en bleu). La société propose des solutions entreprise.

Metacrawler www.metacrawler.com
A voir la recherche avancée : Options de recherche par grande région ou pays, choix de la vitesse de recherche, du nombre de résultats par source, et choix du classement des résultats (par pertinence, par source ou par site)

Profusion www.profusion.com
Racheté par la Intelliseek, qui développe le méta-moteur "off-line" Bulls-Eye, Profusion propose aujourd'hui, après une version beta pendant quelques mois, le nouveau ProFusion : Dépouillé de toutes bannières publicitaires, l'outil a la particularité de permettre une recherche sur des groupes de sources (1000 sources dans plus de 200 groupes), y compris 500 bases de données (web invisible). Il peut "recommander" à son utilisateur des sources d'information additionnelles. A noter également le système d'envoi des résultats par mail à des tiers et d'alerte par mail.

Search.com (ex Savvy Search) www.search.com
Le méta-moteur appartient désormais au réseau de sites CNet "The source for computing and technology". Il propose comme Profusion une recherche sur des groupes d'outils spécialisés, en plus des outils généralistes (annonce 1000 sources thématiques).

Ixquick http://www.ixquick.com
Se présente comme le "métachercheur le plus puissant du monde", et dispose entr'autres d'une interface en français. Il a récemment acheté le méta-moteur Debriefing. Ixquick a l'avantage de traduire les requêtes même complexes à base de parenthèses, de recherche sur champs (si une fonctionnalité spécifique est indiquée, elle sera prise en compte seulement par les moteurs qui la supportent). Le classement se fait via le classement des outils utilisés (par rapport au nombre de moteurs qui ont choisi les pages dans leur "top 10").

Vivisimo www.vivisimo.com

Ce nouveau méta-moteur, issu de Carnegie Mellon University a pour particularité de classer les résultats des recherches dans des dossiers, à la manière de Northern Light.. L'interface propose à gauche un menu hiérarchique de sujets et sous-sujets et à droite le

groupe de résultats choisis. Notons que le méta-moteur n'utilise pour la classification que les titres et les brèves descriptions ramenées par chaque moteur.

Traduit la requête AND + OR NOT NEAR

La pertinence n'est pas recalculée mais est fonction des algorithmes des moteurs utilisés

Vivisimo traduit les requêtes dans le langage des moteurs

Zworks www.zworks.com

Ce méta-moteur récent a l'avantage d'avoir une interface guidée très conviviale et formate les requêtes selon l'outil de recherche utilisé (comme Ixquick). Il propose également un filtre (sexe, etc.)

Les méta-moteurs spécialisés

Il interrogent simultanément sur le texte intégral de plusieurs bases de données, moteurs, répertoires ou sites dans un domaine particulier

A voir notamment les nouvelles fonctions des méta-moteurs "on-line" des méta-moteurs Profusion et Search.com vus plus haut.

Voir également les thématiques de recherche proposées par les méta-moteurs "off-line" tels Copernic, Strategic Finder ou BullsEye.

Autres exemples sur le Web :

Domaine	Exemple	Adresse
Images	ImageWolf	www.trellian.com/iwolf
Adresses e-mail	Mesa	http://mesa.rrzn.uni-hannover.de
Emploi	Keljob	www.keljob.com
Presse	Newstrawler (généraliste) Findarticles (business)	www.newstrawler.com www.findarticles.com
Médecine	Citeline	www.citeline.com
Information financière	Big Charts	www.bigcharts.com
Sciences	Scirus	www.scirus.com

Le web invisible

Il s'agit de l'ensemble des pages non localisables et/ou non indexables par les outils. Le web invisible correspond à plusieurs types de ressources :

- ✓ Pages dont les caractéristiques techniques rendent difficiles, sinon impossible l'indexation par les moteurs : frames, javascrips modifiant le contenu, technologies propriétaires (par exemple flash, active X, java)
- ✓ Pages situées à l'intérieur d'une frame (cadre)
- ✓ Pages qui n'ont fait l'objet ni d'un référencement direct, ni d'aucun lien d'une autre page.
- ✓ Pages nécessitant une identification de la part de l'internaute
- ✓ Pages dont le contenu indique aux moteurs qu'ils ne doivent pas l'indexer

- ✓ Page produite à partir de bases de données ou d'applications, et dont l'URL comporte des paramètres non exploitables par la plupart des moteurs
- ✓ Page produite à partir de données saisies par l'utilisateur via un formulaire html. Exemple : les résultats de l'interrogation d'une base de données avec des critères de recherche entrés par l'utilisateur.

(définition mise au point par les formateurs internet ADBS)

Identification des ressources du Web invisible.

- ✓ Elle passe en bonne partie par une culture significative du web dans son domaine Connaître les portails thématiques, se tenir au courant, être inscrit à des lettres de diffusion thématiques, se prévoir des journées spécifiques découvertes, … et mettre en bookmarks les pages utiles.
- ✓ Pour les bases de données accessibles via l'Internet, utiliser des répertoires spécifiques, tels :

  - en français, le répertoire de Jean-Pierre Lardy "DADI" : http://urfist.univ-lyon1.fr

  - en anglais, www.invisibleweb.com ou www.completeplanet.com

  Selon une étude de la société BrightPlanet (Completeplanet) parue en juillet 2000, il y aurait 100 000 bases de données disponibles, riches en contenu, représentant 550 milliards de pages Web (7 500 Tera Octets d'information) qui serait gratuitement accessibles pour 95% d'entre elles et sont caractéristiques du "Deep web" (expression choisie par Bright Planet.

  Les résultats de l'étude :
  http://www.completeplanet.com/Tutorials/DeepWeb/index.asp
- ✓ Pour les pages produites dynamiquement, utiliser des moteurs d'actualité tels www.net2one.com ou des outils du type www.dailydiffs.com (vérifie tous les jours les changements sur des dizaines de milliers de pages sélectionnées manuellement)
- ✓ Utiliser des méta-moteurs spécialisés, si toutefois il en existe dans son domaine. Le méta-moteur "on-line" Profusion, racheté par la société Intelliseek (qui développe le méta-moteur "off-line" Bulls-Eye), propose une recherche sur des groupes de bases de données par thèmes : http://beta.profusion.com

Les listes et les forums

Listes de discussion

Elles utilisent le protocole du courrier électronique. Les personnes intéressées doivent s'abonner à la liste choisie et reçoivent alors dans leur boîte aux lettres les messages postés. Le serveur de listes gère les échanges en recevant les contributions à son adresse ("l'adresse de la liste") et en les renvoyant à tous les abonnées.

Les serveurs de listes travaillent donc de façon individuelle, ce qui explique la difficulté à pénétrer dans les archives de certaines listes à moins d'y être abonné. Il n'existe pas de site permettant l'interrogation immédiate de l'ensemble des messages parus sur toutes les listes du monde.

On assiste aujourd'hui, d'un part à un mouvement de fusion chez les serveurs de listes hors secteur universitaire / recherche, d'autre part à une multiplication de listes privées, et enfin à une tendance à la gratuité de l'hébergement des listes, au prix d'un peu de publicité.

Au niveau francophone, Francopholistes (www.francopholistes.com) reste le répertoire incontournable avec plus de 3200 listes indexées. La société propose depuis peu une recherche centralisée sur les archives récentes de l'ensemble des listes francophones (recherche parmi 13000 messages)

Citons aussi Kitalettre (http://www.buongiorno.com/fr/) qui s'est spécialisé dans les "périodiques" envoyés par mail et acheté en mars 2001 par le groupe italien Buongiomo, spécialisé dans l'e-mail marketing (et a depuis changé de nom).

A l'international, les deux principaux hébergeurs de listes de discussion sont Topica et Egroups / Yahoogroupes

- ✓ La société Topica (www.topica.com) est née à la fin 98, et a racheté le très connu répertoire de mailing-lists Liszt en avril 99, et n'a cessé depuis de progresser en notoriété et en audience : le serveur héberge actuellement plus de 200.000 listes pour 40.000 en mai 99.
- ✓ Yahoo est devenu avec Yahoo!Groups (http://groups.yahoo.com) l'un des plus importants hébergeurs de listes depuis la reprise de E-groups, qui avait lui-même racheté son concurrent OneList.

Directory of Scholarly and Professional E-Conferences : Diane Kovacs a remis en ligne son répertoire des mailing-lists intéressant étudiants, chercheurs et professionnels http://www.kovacs.com/directory . La recherche se fait par mot-clé ou catégorie.

Forums de discussion

Les forums de discussion rentrent dans deux catégories distinctes :

- ✓ **Les forums "classiques" (ou newsgroups ou forums usenet)** se sont développés dans les années 80. Ils sont organisés selon une arborescence précise, et fonctionnent grâce à un réseau spécifique de serveurs. Deux modes de consultation sont envisageables :
  - avec le logiciel de news intégré à son navigateur, ou via un autre logiciel spécialisé : on consulte alors les messages postés dans leur format d'origine, et on est tributaire du choix de forums proposé par son fournisseur d'accès ou son entreprise. En France, il est rare d'avoir ainsi accès à plus de 12000 news internationaux
  - sur le Web : Via le site web de sociétés qui archivent sur des serveurs web les messages échangés sur le réseau Usenet, qui sont alors consultables avec un simple navigateur. Le choix de forums est alors souvent beaucoup plus large que dans le premier cas, et on peut répondre directement sur le Web.

  Deja a longtemps été la référence en donnant accès à plus de 45000 forums et aux archives depuis 1995 (plus de 500 millions de messages). En février 2001, Deja a été racheté par Google qui donne accès aujourd'hui à 650 millions de messages depuis 1995. Après une période de transition, les fonctionnalités de recherche sont assez complètes : par newsgroup, par sujet, par auteur, par langue et par date. http://groups.google.com

  A voir : http://newssearch.pilum.net

  Citons aussi, pour la France, le serveur mis en place par Voila http://news.voila.fr et le challenger Foorum (www.foorum.fr) qui suit 13000 forums de discussion francophones.

- ✓ Les "web forums" (ou message boards ou bulletin boards) apparus beaucoup plus récemment : il s'agit d'espaces sur le Web, créés à l'intérieur d'un site sous forme de pages html où l'on peut poster et consulter les messages. Il est donc nécessaire de se connecter d'abord au site hébergeant le forum pour y participer.

  De nouveaux outils permettent de faire une recherche sur le texte intégral des messages postés sur de nombreux forums :

  www.boardreader.com (fondé en mai 2000 par des ingénieurs et étudiants de l'Université du Michigan)

  www.messageking.com classe les résultats par catégories

Trucs et astuces

QUAND UTILISER QUELS OUTILS ?

La réponse à cette question ne peut pas être définitive. Rappelons que la recherche d'information sur Internet n'est pas une science, et tout dépend aussi de son expérience de la recherche et du Web, et de sa façon de travailler.

Disons en simplifiant beaucoup…

En fonction du type de recherches

- ✓ Recherches larges ou première approche : ⌦annuaires généralistes
- ✓ Recherche d 'information ponctuelle (tous secteurs) : ⌦moteurs généralistes
- ✓ Recherche sur des données de nature bien définie (statistiques, pays, presse, indicateurs…) : ⌦annuaires et outils spécialisés sur ce type de recherche
- ✓ Recherches récurrentes sur un sujet: ⌦identification de sites via pages de liens ou annuaires spécialisés, puis recherche par navigation / ⌦méta-moteur off-line
- ✓ Recherches précises sur noms ou chaînes de caractères (sans booléens) : ⌦méta-moteurs.

En fonction de sa connaissance du sujet :

	Faible connaissance du sujet	Bonne connaissance du sujet
"Question-réponse"	.Recherche sur les moteurs ou méta-moteurs .Remonter à un concept plus généraliste et utiliser les annuaires	"Sites de référence" (Sites spécialisés sur le sujet, repérés au préalable)
"Tout savoir sur"	.Annuaires pour identifier les bons sites et les bons mots clés .Recherche sur " sites de référence" .Recherche sur moteurs	" Sites de référence" complétés par recherches sur moteurs ou méta-moteurs

COMMENT TROUVER DES SITES SIMILAIRES À UNE SOURCE DÉJÀ CONNUE ?

Cette stratégie de recherche est souvent payante, et permet de compléter une information ou même d'identifier des concurrents d'une société.

Plusieurs solutions sont envisageables :

- ✓ Utilisation des répertoires : Le nom du site devient le mot-clé à utiliser. Il suffit alors de cliquer sur la rubrique concernée par le site, ou le cas échéant de choisir la catégorie la plus adéquate. Par exemple, en tapant "adbs" dans Nomade, on peut se diriger ensuite sur la rubrique : Sciences humaines et sociales > Sciences

de l'information et de la communication > Documentation > Associations, organismes professionnels > France.

Sur Yahoo, on peut aussi utiliser l'adresse du site connu (ou des mots de l'URL) comme clé de recherche. On pourra ainsi écrire u:adbs.fr.

- ✓ Utilisation des moteurs classiques : On choisira alors, à partir de la page de résultats, la fonction appelée "Related pages" ou "Sites similaires". Cette option est disponible dans plusieurs moteurs, notamment Alta Vista et Google.

  Sur Alta Vista, on peut directement utiliser le mode de requête like:www.adbs.fr
- ✓ Utilisation des moteurs linguistiques : Un moteur comme Webtop, qui procède par analyse de contenu et extraction de concepts, permet également de trouver des sites similaires : il faut alors utiliser la fonction "copy and paste" et faire un "copier-coller" de la page la plus pertinente du site connu.
- ✓ Utilisation de la fonction "Apparenté" de Internet Explorer ou "Infos connexes" sur Netscape, c'est à dire en fait de l'utilitaire Alexa (aujourd'hui propriété de Amazon.com) qui peut aussi être téléchargé rapidement :Alexa propose notamment des pages similaires en suivant les liens des pages et en utilisant le filtrage collaboratif. Alexa donne aussi des informations sur les pages visitées
- ✓ De nouveaux outils dits "contextuels" ont fait leur apparition ces derniers mois. Ainsi, Kenjin (www.kenjin.com) est un logiciel diffusé gratuitement par Autonomy : S'intégrant à la barre des tâches de Windows, Kenjin réalise une analyse sémantique contextuelle à partir d'une page web ou word (ou d'une sélection), et propose des sources complémentaires. Le nombre de résultats est limité à six.

  Voir aussi dans la même famille : Flyswat (www.flyswat.com), Gurunet (www.gurunet.com), Nano (www.nano.com), Zapper (www.zapper.com), etc. (très anglosaxons dans le choix des sources).

Qu'est-ce que le "peer-to-peer" ?

Le "peer to peer" ou "p2p infosharing" ("pair" à "pair"), mis en lumière par Napster, qui permet d'échanger des fichiers musicaux au format MP3 directement entre ordinateurs particuliers, fait beaucoup parler de lui. Napster distribue gratuitement un petit logiciel qui permet aux internautes de s'échanger des fichiers sans intermédiaire. Grâce à un répertoire constamment mis à jour (fichiers disponibles et adresses de leur propriétaire sur un serveur central), Napster aiguille chaque internaute vers un pair qui détient sur son disque dur les morceaux de musique convoités.

Il est ainsi possible de lancer une requête pour des fichiers vers les ordinateurs des autres internautes adhérents au système et partageant le logiciel, et de créer un système distribué pour la recherche d'information.

L'objectif est même de créer un système totalement distribué de partage de fichiers entre utilisateurs, sans les intermédiaires serveurs classiques. Les autres intérêts sont d'avoir une information toujours à jour, et de pouvoir récupérer différents types de fichiers. L'inconvénient majeur réside dans le danger d'une mauvaise configuration des logiciels laissant certains dossiers ouverts à tous…

- ✓ Gnutella permet d'effectuer des recherches sur des bases de données internes à un site. Chaque ordinateur dialogue avec ses voisins pour être informé à chaque instant, des fichiers disponibles et de leur localisation. La technologie est utilisée

par des jeunes développeurs qui ont développé le moteur de recherche infrasearch , acquis par Sun en mars 2001dans le cadre de son projet de rechercherche ("jxta") pour développer les techniques de recherche, de partage et de stockage de l'information à travers: (http://search.jxta.org)

- ✓ Freenet (www.freenetproject.com) : A revoir un espace des disques durs des ordinateurs connecté est réservé aux échanges de fichiers par Freenet. On y trouve les fichiers destinés à être partagés mais aussi des fichiers très demandés que Freenet recopie sur certains "nœuds" de ce réseau virtuel.
- ✓ Pointera (www.pointera.com) : Le "Pointera search engine" travaille "classiquement", mais est aussi capable de poursuivre ses investigations sur les disques durs des utilisateurs volontaires et connectés. Ceux-ci ont préalablement défini un espace ouvert à Pointera. Pointera revendique la possibilité de travailler avec 500 millions de PC et cible les portails et sites "de contenu vertical"
- ✓ Human Links (www.human-links.com) : La société française Amoweba teste auprès de plus de 100.000 volontaires son outil Human Links exploitant le peer to peer. Le logiciel présente une interface utilisateur originale représentant les centres d'intérêt , les contacts et les pages web en relation les uns avec les autres sur une carte. Le nombre de cartes créées est illimité.Il s'agit en fait d'échange entre internautes des URL indexées dans leurs favoris. En février2002, la société a proposé la plate-forme Human-Links Organization, une version intranet/extranet de son moteur appliquée à la gestion de la connaissance, développée en collaboration avec l'éditeur L2T.
- ✓ A suivre : le projet Pandango (www.pandango.com)
- ✓ Alta Vista propose enfin depuis peu d'exploiter ce concept dans l'environnement des entreprises. L'outil permet de faire ses recherches d'une part sur un ensemble de sites web, et d'autre part sur les ordinateurs des salariés, et ce dans 200 formats et 30 langages différents. Ce système peut toutefois apparaître comme très indiscret aux salariés de l'entreprise.
- ✓ Fév 2002 : Nouvel outil gratuit en anglais widesource (www.widesource.com) ; il s'agit d'un système de partage de "carnet d'adresses internet"

Complément d'information

- ✓ sur Red Herring, article de décembre 2000 :"Can peer-to-peer grow up ? : http://www.redherring.com/mag/issue86/mag-grow-86.html
- ✓ Sur le site de O'Reilly and associates, article issu de "The O'Reilly P2P Conference, tenue en février 2001 à San Francisco "Gnutella and Freenet represent true technological innovation" www.oreillynet.com/pub/a/network/2000/05/12/magazine/gnutella.html
- ✓ NB : Napster a été déclaré coupable par une cour américaine en février 2001 de complicité d'enfreinte au droit d'auteur. La société vit sous la menace d'une interdiction définitive ou, au mieux, d'une transformation en canal de distribution payant du catalogue musical de son nouvel actionnaire Bertelsmann.
- ✓ Voir aussi les liens de la page : http://www.cyberpolitik.org/informatique/peertopeer.html

PEUT-ON UTILISER LE LANGAGE NATUREL SUR LES OUTILS DE RECHERCHE ?

"Everyone's trying to get away from keyword" (Paul Hagen, analyste chez Forrester Research)

Sur le Web, la plupart de ceux annonçant "comprendre" le langage naturel se contentent le plus souvent de supprimer les mots parasites (où, quoi, pourquoi, qui, est,…) de la question pour ne conserver que les mots signifiants et lancer alors une requête classique "full text".

Le traitement du langage dit "naturel" fait appel à des analyses syntaxiques et sémantiques complexes et coûteuses. Rares sont donc les outils de l'Internet proposant ce type de recherche. Citons le moteur Oingo lancé fin 99.

- ✓ Oingo, développé par la société Applied Semantics, travaille à partir d'un base de termes compilés par une équipe de linguistes reliant les mots à des synonymes, expressions, termes familiers et concepts. Il propose sur le web (www.oingo.com) une démonstration de sa technologie utilisant le répertoire Open Directory. Le moteur présente avec chaque liste de résultats des mots ou concepts reliés aux termes de recherche. Peut aussi être utile lorsque l'on cherche des suggestions de mots-clés

- ✓ En français, Albert (www.albert-inc.com). propose une démonstration de sa technologie sur une version démo basée sur l'index du moteur Fast.). Le groupe vient d'annoncer la disponibilité sur le marché français de son offre d'accès à l'information. Voir le site de l'ONU sur l'assistance humanitaire www.reliefweb.int (+ de 150000 documents) un des premiers clients d'Albert. Le moteur précise les interrogations des utilisateurs, les reformule et les interprète ; fonctionnant sur le principe de la logique floue, il prévoie les erreurs de syntaxe, d'orthographe ou les questions ambiguës et formule plusieurs requêtes en tenant compte de ces biais au système de recherche. Albert stocke et analyse l'historique des requêtes dans une base de connaissances de façon à pouvoir s'adapter. Pas de dictionnaire intégré. Signature d'un accord mondial avec l'américain Verity, éditeur de solutions pour les portails d'entreprise et d'indexations de contenus.

- ✓ Le moteur américain Ask Jeeves (www.ask.com) a innové dans ce domaine en travaillant depuis 1997 sur une base de données de plus de dix millions de questions / réponses, une page web et une seule étant sélectionnée pour chaque question. L'utilisateur choisit donc, dans la liste de questions proposées après analyse de sa requête par l'outil, celle se rapprochant le plus de son besoin. Une équipe d'une trentaine de personnes est chargée d'alimenter la base. Ask Jeeves intègre désormais un répertoire avec les données de l'Open Directory. Mais comme Google, il reclasse les sites dans chaque rubrique par ordre de popularité et non plus par ordre alphabétique. Ask Jeeves prend en compte les données de Direct Hit, racheté debut 2000.

  Ask jeeves commercialise depuis peu un logiciel permettant à toute entité de mettre en place soi-même son propre outil d'interrogation, Premier prix 100000 $. Ask Jeeves a acheté l'outil Teoma très rapidement après sa sortie et propose désormais en alternative les résultats. Il a racheté également en janvier 2002 la société Octopus Inc., société conceptrice de logiciels permettant d'interroger des bases de données, donc le "Web invisible". Ces nouveaux modules de recherche devraient être intégrés aux plates-formes de recherche d'information de Ask Jeeves (et notamment le logiciel JeevesOne, permettant de gérer des applicatifs de type SAV par exemple) d'ici au milieu de l'année 2002. (Abondance)

  Infoclic qui était le "clone" français de Ask Jeeves, vient malheureusement de cesser son activité(www.infoclic.fr

- ✓ On peut également citer les nouveaux outils tels Subjex en anglais (www.subjex.com) qui mettent l'accent sur le dialogue avec l'utilisateur pour tenter de mieux affiner son besoin, d'obtenir de nouveaux mots-clés

Complément d'information *:* article de avril 2000 sur le site "The Standart" : "The language barrier" donnant notamment les références de sociétés américaines spécialisées dans la recherche en langage naturel.
http://www.thestandard.com/article/display/0,1151,14040,00.html

Comment identifier des fichiers pdf sur le Web ?

Le format PDF (Portable Document Format) est créé à l'aide du logiciel Acrobat de Adobe.. Il permet d'avoir une visualisation fidèle du document original, sans avoir besoin du logiciel de création (Word, Xpress), ni des polices de caractères utilisées. Les fichiers pdf sont lisibles et imprimables grâce à un logiciel gratuit et téléchargeable, Adobe Acrobat Reader.

Il est assez répandu pour la diffusion de documents professionnels sur Internet (très peu grand public, d'où son intérêt). Les documents au format pdf font partie du Web invisible (cf page) dans la mesure où ils ne sont généralement pas pris en compte par les moteurs traditionnels.

- ✓ Adobe a développé un moteur de recherche pour les documents pdf http://searchpdf.adobe.com Le moteur travaille à partir de l'indexation en texte intégral de plus d'un million de résumés (générés automatiquement) associés au document original. L'outil permet de visualiser le résumé avant de décider de voir celui-ci.

  Le moteur utilisé est celui de Alta Vista et utilise donc les mêmes fonctionnalités (recherche simple).

- ✓ Google permet depuis février 2001 d'identifier des documents pdf répondant aux requêtes de ses utilisateurs : les fichiers sont indexés et proposés en version texte (mots clés choisis en couleurs), et Google leur applique son calcul de pertinence comme aux autres pages de son index.

  Dans les pages de résultats, les fichiers pdf sont clairement identifiés et l'option "Texte" remplace le "Cache" des documents en html.

  Prévoir une requête du type : "mot-clé pdf" ou mots-clés inurl:pdf

  Pour ne récupérer que les fichiers en format pdf (et non également toutes les pages html qui "parlent de" fichiers pdf), effectuer une requête du type : mot-clé filetype:pdf

  Alltheweb fait aujourd'hui de même

- ✓ Les fichiers au format pdf accessibles à partir de documents web sont toujours clairement identifiés comme tels. On peut donc utiliser "pdf" comme mot-clé pour repérer ces fichiers. Mais cette stratégie ne permet qu'une recherche très large, puisqu'elle ne porte ni sur le résumé, ni bien sûr sur le texte intégral du document pdf.

Comment identifier des sites fédérateurs (portail vertical ou vortal) ?

Les sites fédérateurs ou portails (cf page) sont des outils de recherche incontournables dans de nombreux domaines. Les répertoires thématiques proposés et les autres ressources peuvent faire gagner beaucoup de temps lors d'une recherche.

Il convient toutefois d'être prudent et d'évaluer sérieusement leur valeur ajoutée et les objectifs de l'éditeur. : la mode est aux portails et des sites de ce type se construisent tous les jours ; certains ont la quête de notoriété pour seul objectif.

Plusieurs voies d'approche sont possibles :

- ✓ Utiliser les répertoires généralistes de type Yahoo ou Open Directory. Pour certaines thématiques, une sous-rubrique "annuaires" ou "directory" sera disponible, pour d'autres, une exploration sera nécessaire, à partir des résultats pour une requête la plus large possible. Le répertoire About.com en anglais est souvent intéressant.
- ✓ Exploiter les répertoires d'outils de recherche et de portails verticaux, et les répertoires professionnels (dont ADBS / Guide thématique du Web). Attention, ils ne sont jamais exhaustifs, et peu critiques pour la plupart.
- ✓ On pourra aussi utiliser un outil comme Argus Clearinghouse (www.clearinghouse.net), répertoire sélectif et commenté de quelque 400 "méta-pages" réalisés par des indépendants et couvrant les grands domaines de la connaissance.
- ✓ S'appuyer sur les sites des associations professionnelles, donnant le plus souvent les liens clés du secteur (sur les répertoires généralistes ou bien sur des répertoires spécialisés entreprises comme Indexa en France www.indexa.fr)
- ✓ Si on en connaît déjà un, aller sur un site de référence sur le sujet, et suivre les liens le plus souvent indiqués
- ✓ Repérer un ou deux "sites de référence" et chercher les "backlinks" (liens pointant vers ces sites). On peut travailler avec les syntaxes spécifiques des moteurs (Google, Alta Vista, Hot bot) ou bien utiliser un méta-moteur spécifique comme link popularity www.linkpopularity.com
- ✓ Passer par un moteur : Exemple sur Alta Vista, pour trouver une "méta-page" en géographie, on peut essayer une équation du type (links* OR liens) NEAR (Internet OR web) NEAR geograph*
- ✓ Le "bouche à oreille" (y compris sous forme électronique avec les listes de discussion et forums) : l'information sur les bons sites circule…

Comment choisir ses mots-clés ?

Quand ? La sélection des mots-clés s'effectue après le choix d'une stratégie de recherche. En effet, le choix sera fondamentalement différent si l'on cherche un portail thématique, ou une source susceptible de fournir l'information ou l'information précise immédiatement. Pour simplifier, disons que dans le premier cas, les mots-clés seront "le plus large possible", dans le second cas, ils seront "le plus précis possible".

Un ou plusieurs ? On procédera par étape pour affiner éventuellement sa recherche à l'aide de plusieurs mots-clés. Si le nombre de résultats est faible avec un seul mot-clé précis (exemple : 100 résultats sur un moteur), inutile de préciser davantage. Donc, utiliser d'abord un seul mot clé (ou expression) quant la terminologie ou l'association terminologique est très spécifique.

Pour ou contre le SAUF ? On peut aussi isoler les mots-clés à exclure absolument car générateurs de bruit (opérateur SAUF ou signe -). Attention toutefois à ne pas aller trop vite, de peur de passer à côté de documents pertinents : Ainsi, si je cherche des informations sur les énergies alternatives autres que solaires, je peux être tenté d'"envoyer" au moteur une équation du type +"énergies alternatives" –solaires. Mais je n'aurai pas alors les ressources qui abordent successivement **toutes** les énergies alternatives.

Majuscules, minuscules ? En général la saisie en minuscules non accentuées donne par défaut toutes les occurrences. De nombreux moteurs (Google, Voila) ne font aucune différence entre les deux saisies. D'autres (Alta Vista, Hot Bot) comprennent différemment la saisie en majuscule et en minuscule : Dans ce cas "Python" renverra les pages contenant Python, alors que "python" donnera des pages comportant python, PYTHON, ou Python.

Même principe pour l'accentuation ou non des mots-clés.

Troncatures ? Sur un outil comme Yahoo ou Voila, les mots de la recherche sont "déclinés" au pluriel. Sur d'autres, la chaîne de caractères exacte est recherchée. La troncature, le plus souvent applicable avec le caractère * permet d'étendre la recherche en remplaçant plusieurs caractères : avec le mot-clé "ménage", la recherche s'effectuera aussi sur "ménages" ou "ménagères".

La troncature sur le Web permet généralement de remplacer plusieurs caractères sur la fin des mots. Sur Hot Bot, le caractère ? remplace un seul caractère (* est OK pour plusieurs caractères). AllTheWeb ne propose toujours pas de troncature.

Et les synonymes ? Il est important d'explorer la terminologie du domaine de recherche, pour repérer les synonymes (très rares sont les moteurs travaillant sur les concepts). De façon générale, les premiers documents intéressants récupérés permettent de valider, compléter ou revoir ses mots-clés.

Astuces pour identifier des synonymes et/ou mots associés

- ✓ Utiliser un dictionnaire de synonymes tel celui du laboratoire de linguistique du CNRS pour les termes en français http://elsap1.unicaen.fr/dicosyn.html
- ✓ Utiliser un moteur de recherche travaillant à partir de dictionnaires, encyclopédies, thesaurus, tel pour les termes en anglais www.xrefer.com. Voir aussi www.thesaurus.com.
- ✓ Utiliser un répertoire de dictionnaires comme Xlation www.xlation.com ou OneLook Dictionaries www.onelook.com
- ✓ Utiliser pour l'anglais le méta-moteur Surfwax (www.surfwax.com) en cliquant sur la petite flèche suivant la ligne "focus:mot-clé choisi" au dessus des résultats à gauche.
- ✓ Utiliser un thésaurus de son domaine (en ligne gratuit, ou acheté comme par exemple celui de la base INSPEC (www.iee.org.uk)
- ✓ Utiliser le "générateur de mots-clés" du site français Abondance : Donne les mots-clés le plus souvent présents dans les pages web contenant le mot-clé demandé (à partir des résultats de Google ou de Alltheweb) www.abondance.com/audit/motscles.html
- ✓ Explorer les balises méta (keywords) de quelques documents pertinents
- ✓ A partir d'un document pertinent, chercher des sites / pages similaires (cf astuce page)
- ✓ Pour passer du français à l'anglais, utiliser à partir d'une catégorie donnée, le "passage direct" de yahoo.fr à yahoo.com : "Poursuite de la recherche sur Yahoo US"
- ✓ Utiliser un moteur travaillant en langage naturel comme oingo.

COMMENT GÉRER LES PROBLÈMES FRÉQUENTS AVEC LES OUTILS ?

- ✓ **Erreurs 404, liens non valables** : remonter dans la hiérarchie du site. Si l'adresse de l'host est bonne, revenir à cette adresse et "tatonner" à l'intérieur du site pour retrouver la page cherchée et sa nouvelle URL. On peut aussi utiliser le lien "cached" sur Google (cf page) ou les archives de Alexa (cf page)

- ✓ **Signification des principaux messages d'erreurs** :

Erreur	Message	Signification
400	Bad Request	Erreur dans l'adresse
401	Access Denied	La consultation nécessite un nom d'utilisateur et un mot de passe
403	Forbidden	L'accès est réservé et vous n'avez pas les privilèges correspondants
404	Not found	La page correspondant à cette URL n'a pas été trouvée sur le serveur
500	Internal	Problème de serveur. Contacter l'administrateur du site
503	Read time out	Le temps alloué à la connexion est écoulé

- ✓ **Réponses hors sujet** : reformuler sa question, rajouter des mots clés…

- ✓ **La page proposée ne contient pas votre terme de recherche** .

Il peut y avoir plusieurs explications, mais la plus vraisemblable est que ce mot se trouvait dans la page lorsque celle-ci a été sauvegardée par le robot du moteur. Puis elle a été modifiée et le mot a disparu de la page. Mais par contre il est resté dans l'index de la base de données. Il se peut aussi que votre terme apparaisse dans un formulaire déroulant, ou enfin en méta-données.

Une solution pour être certain d'obtenir des résultats contenant les mots-clés de votre question consiste à utiliser un méta-moteur "off-line" avec la fonction "raffiner" ou "filtrer".

- ✓ **Non élimination des doublons** : les moteurs utilisent maintenant à peu près tous les techniques de clustering pour la présentation des résultats (une réponse = un site et non une réponse = une page) ou le proposent en option. Mais cela n'empêche pas toujours les doublons.
- ✓ **Problème d'accès à de l'information très récente** : attention, un moteur peut mettre plusieurs jours ou mêmes semaines avant d'indexer un nouveau site… Voir du côté des serveurs d'actualité, par exemple.

PEUT-ON FAIRE UNE RECHERCHE DANS LES BALISES "META KEYWORDS" ?

Les méta-données : Définition

Ces balises du langage html permettent de donner des informations (description, mots-clés) sur le contenu d'une page web.

Elles se trouvent dans l'en-tête HTML de la page Web, (le "HEAD") et fournissent des informations qui ne sont pas visibles par les navigateurs. Les méta-tags les plus courants (et les plus utiles pour les moteurs de recherche) sont KEYWORDS (mots-clés) et DESCRIPTION.

Pour visualiser les méta-tags : Affichage Source (Explorer)
CTRL U (Netscape)

Le méta-tag KEYWORD permet à l'auteur de souligner l'importance de certains mots et phrases utilisés ou non dans sa page. Certains moteurs de recherche tiendront compte de cette information - d'autres l'ignoreront. Certains moteurs donneront en plus un « coup de pouce » dans le classement pour certains documents au cas où le mot clé de requête se trouve dans les méta-tags, mais ils peuvent pénaliser une page où un terme est répété plusieurs fois dans la balise meta keyword..

Le méta-tag DESCRIPTION permet à l'auteur de contrôler le texte affiché quand la page paraît au niveau des résultats d'une recherche. Certains moteurs de recherche peuvent ignorer cette information. Contrairement à KEYWORDS , DESCRIPTION est en langage naturel.

```
<meta http-equiv="content-type" content="text/html;charset=iso-8859-1">
<title>Conseil constitutionnel - R&eacute;publique Fran&ccedil;aise</title>
<meta name="Description" content="Conseil constitutionnel - France - Pouvoir public constitutionr
<meta name="keywords" content="constitution france conseil constitutionnel jurisprudence loi &eac
<meta name="Author" content="J&eacute;r&ocirc;me Rabenou - Conseil constitutionnel">
<meta name="keywords" content="constitution france conseil constitutionnel jurisprudence loi &eac
<x-sas-window top="0" bottom="481" left="0" right="788">
```

En savoir plus sur les méta-tags

: http://searchenginewatch.com/webmasters/meta.html
ou www.abondance.com/docs/meta_1.html

Voila est l'un des seuls moteurs actuellement à proposer cette fonction en recherche avancée :

keywords:mot-clé

description:mot-clé

COMMENT EFFECTUER UNE RECHERCHE PAR NAVIGATION ?

Il s'agit là d'une démarche un peu inverse à celle utilisée via les moteurs de recherche. Au lieu de « lancer » des mots clés et de consulter les pages retrouvées par les moteurs de recherche, il s'agit de se poser au préalable la question : **« Qui pourrait détenir l'information que je recherche ? »** et aller la rechercher à sa source. La difficulté ici réside dans la multiplicité des chemins possibles et la profondeur des emboîtements sur le web.

La démarche la plus simple consiste à exploiter les sites connus, de préférence répertoriés au préalable dans les signets ou favoris. Mais ce n'est généralement pas suffisant.

Il s'agit donc d'identifier des sites « de référence » par rapport à un sujet :

- soit en utilisant des annuaires généralistes (Yahoo, Nomade, Dmoz) ou éventuellement des annuaires spécialisés géographiquement
- soit en utilisant un portail spécialisé. Pour identifier les portails spécialisés, voir plus bas, chapitre « Trucs et astuces ».
- soit en recherchant des sites similaires à un site connu (voir autre "trucs et astuces")
- soit en explorant les liens à partir d'un site connu dans ce domaine.

Cette méthode est particulièrement efficace dans le cas où l'information recherchée se situerait dans le « web invisible ». Elle demande à la fois une grande rigueur (pour éviter de se perdre dans l'exploration) et une certaine intuition…

Quelques conseils pour ne pas se perdre dans la navigation :

- face à un site inconnu : visualiser le plan du site. Utiliser les moteurs de recherche internes aux sites.
- maîtriser ses clics de souris : essayer de mémoriser le cheminement.
- mettre en signet ou favori et classer dans un dossier les sites potentiellement intéressants.

La recherche sur sites de presse

Même si le support numérique n'a pas encore supplanté le support imprimé, la presse d'actualité est de plus en plus présente sur Internet. Le support web permet par ailleurs d'introduire de nouveaux services et plus d'interactivité : forums, newsletters, diffusion d'information continue, signets spécialisés, dossiers d'actualité, voire filtrage et diffusion d'information ciblée (push)…

Comment identifier les sites de presse ?

- à partir d'un répertoire généraliste (**Yahoo, Dmoz**…)
- à partir d'un répertoire de sites de presse (**Presseweb www.presseweb.com, Newspapers www.newspapers.com Emedia http://emedia1.mediainfo.com/emedia…)**
- à partir d'un portail ou d'un site de référence spécialisé sur le sujet.

Pour la presse spécialisée, en France, on peut également utiliser des répertoires comme **Viapressepro** www.viapressepro.com ou le site de la **FNPS (**Fédération Nationale Presse Spécialisée) www.fnps.fr. Pour la presse spécialisée mondiale, utiliser **Publist** www.publist.com

Comment rechercher sur les sites de presse ?

La recherche sur les sites de presse pourra selon les cas se faire à partir des sommaires et tables des matières, ou bien par mots clés à partir du texte intégral des archives (pas toujours disponible gratuitement).

On peut enfin effectuer des recherches simultanées sur plusieurs sites de media avec des méta-moteurs spécialisés comme **Net2one** www.net2one.fr, **Dernieres.com** www.dernieres.com**, Newstrawler** www.newstrawler.com ou encore **Individual** www.individual.com.

Ces méta-moteurs proposent généralement gratuitement en plus une diffusion ciblée (push).

Enfin, de nouvelles revues sont apparues, sous forme uniquement électronique, sans équivalent papier. On les appelle **newsletters ou webzines** (ne pas les confondre avec les listes de diffusion). On peut y trouver, souvent gratuitement, de l'information exclusive et précieuse.

Pour identifier des revues électroniques et des newsletters, on peut utilise r le site de l'Université de Pennsylvanie www.library.upenn.edu/resources/ej/ej.html

PEUT-ON FAIRE UNE RECHERCHE PAR DATES ?

Un certain nombre d'outils permettent d'affiner sa recherche avec un critère temporel généralement dans leur recherche guidée ou avancée. Mais attention : c'est la date de dernière mise à jour des documents au moment de l'aspiration des pages par le crawler du moteur qui sert de référence, et non une date intégrée au document, ce qui ne garantit pas forcément la fraîcheur des informations. Par ailleurs, les sites à fort renouvellement de contenu (tels CNN) seront sur-représentés pour certains moteurs qui privilégient ces sites dans le rafraîchissement de leur index (exemple : AltaVista)

- ✓ Alta Vista.: Limitation possible par périodes (deux semaines, mois, deux mois, trois mois, année) ou par plages de dates (Du Au; Si rien n'est indiqué dans le champ "AU", c'est la date du jour qui est prise en compte).
- ✓ Google : Limitation possible par périodes (trois derniers mois, six derniers mois, année)
- ✓ Hot Bot : Limitation possible par périodes (une semaine, deux semaines, un mois, trois mois, six mois, un an, deux ans) ou par plages de dates (Avant Après)
- ✓ MSN : Limitation possible par plage de date (Modifié entre et)
- ✓ Voila (+ d'options)

La recherche d'événements par année est aussi possible sur certains outils : Voir notamment en anglais : dMarie Time Capsule http://dmarie.com/timecap/ (chansons, livres, événements de l'année de 1800 à 2001)

Plus général http://www.infoplease.com/millennium1.html (sport, science, etc.)

Sur le Xxème siècle uniquement : . http://www.multied.com/20th/index.html

Evaluation des sites web

L'évaluation de l'information sur Internet devient un enjeu important pour les professionnels. Il s'agit d'un acte d'expertise pour estimer la qualité des différentes ressources disponibles :le portail, le site web, la page web, l'article sur la page, la base de donnée accessible depuis la page, mais aussi le forum, la liste de discussion, le message posté sur une liste ou un forum, etc.

Les critères d'évaluation

Différentes catégories de critères sont à prendre en compte :

- ✓ Crédibilité : Organisation émettrice, type d'émetteur, auteurs des documents, source de financement ou sponsoring, webmaster, cibles et objectifs du site, type d'accès, etc.
- ✓ Fraîcheur : Date de création et de mise à jour
- ✓ Exhaustivité et l'exactitude : Type de document, citations des sources, bibliographie, contextualisation de l'information, qualité de la langue, etc.
- ✓ Adéquation : pertinence et utilité par rapport à la recherche ou à la veille menées.
- ✓ Ergonomie : arborescence, navigation, orientation, frames, etc.
- ✓ Design : présentation visuelle, conception graphique.

Les grilles d'évaluation existantes

La plus aboutie sur le Web (mais très lourde) dans le domaine de l'information santé :www.chu-rouen.fr/netscoring

Voir aussi

Sapristi (INSA Lyon) *csidoc.insa-lyon.fr/sapristi/fristi36.html*

Montréal *www.rrsss06.gouv.qc.ca/commpub/publications/grille.html*

Université Laval *www.fse.ulaval.ca/fac/href/grille/grille.gif*

Astuces pour l'évaluation des pages en cours de navigation

Pour rechercher le propriétaire d'un nom de domaine, on peut utiliser l'outil Whois (base de donnée d'informations sur les noms de domaine, les propriétaires de ces noms de domaines et autres données techniques) disponible sur de nombreux outils, dont Network Solutions (www.networksolutions.com) ou Andco (www.andco.fr) pour une interface en français.

L'utilitaire Alexa utilise un système du même type (cf ci-dessous)

Un outil comme Neotrace (www.neotrace.com) permet d'aller plus loin à partir d'une adresse IP, d'une URL ou d'un E-mail. L'outil montre sur une carte mondiale le chemin parcouru sur le réseau, de nœuds en nœuds jusqu'au serveur correspondant à l'adresse, et donne les informations sur le régistrant. Version d'évaluation gratuite.

Pour trouver des informations générales sur la page, on peut utiliser l'utilitaire Alexa gratuit (téléchargement de la barre d'outils www.alexa.com sous PC et Explorer 5),

propriété de Amazon.com. On obtient les coordonnées du "régistrant", mais aussi des statistiques sur le trafic du site, des témoignages d'internautes, le temps de chargement de la page, le nombre de liens vers cette page, etc. De plus, des sites/pages "similaires" sont proposées et on peut enfin obtenir, en cas d'erreur 404, la copie du document s'il existe dans la base d'archives d'Alexa. La fonction "Infos connexes" du navigateur Netscape résume certaines de ses fonctions.

DEUXIEME PARTIE
Outils de veille

Les agents évolués sur Internet

QUE SONT-ILS ?

La presse informatique a tendance à encenser ces outils logiciels destinés à automatiser des tâches récurrentes, à être mobiles sur les réseaux, à interagir avec l'environnement ou d'autres agents, à prendre des décisions autonomes, voire à faire preuve de facultés d'auto-apprentissage. Actuellement peu d'agents méritent vraiment leur qualification d' "intelligents", mais les meilleurs outils intègrent des technologies variées : Technologies linguistiques, intelligence artificielle, réseaux de neurones, logique floue, technologies mathématiques et statistiques, technologies push, vie artificielle…
Les méta-moteurs sont souvent considérés comme la " première génération " d'agents.

Voici aujourd'hui les grandes fonctions de ces agents sur Internet :

- ✓ Faciliter et guider la navigation via des fonctionnalités variées : meilleure gestion de l'historique, du cache, des bookmarks, informations sur les pages visitées, etc.
- ✓ Assister la recherche d'information : Méta-moteurs évolués, analyse linguistique des requêtes, filtrage collaboratif ("bouche à oreille électronique"), etc.
- ✓ Assister l'exploitation des résultats : Analyse, tri, indexation, résumés automatiques, exports des résultats, cartographie, etc.
- ✓ Permettre un suivi, une surveillance dans le temps : de recherches, de sites, de pages, de dossiers de pages, de produits d'informations spécifiques (actualités, offres d'emploi, infos financières, etc). L'agent gère la connexion à Internet et envoie un rapport de recherche.
- ✓ Permettre la personnalisation de la diffusion automatique d'information.

Ces agents, loin d'être indispensables pour une recherche d'information classique, s'avèrent rapidement incontournables dans une démarche de veille. Les compétences et la synthèse humaines restent toutefois indispensables.

Les "aspirateurs" de sites web

Ils enregistrent le site sur le disque dur pour une consultation hors ligne. Pour cela, ils offrent bien entendu un paramétrage très affiné de l'aspiration et permettent l'export (pour pouvoir consulter le site avec un simple navigateur, sans disposer du logiciel ayant servi à capturer les pages). A l'heure de la généralisation des connexions permanentes, cette fonction présente aujourd'hui moins d'attractivité qu'auparavant.

De nombreux utilitaires existent actuellement, avec des fonctionnalités plus ou moins sophistiquées. En terme de veille, ces outils sont intéressants pour leur capacité à mettre à jour les sites (souvent automatiquement) et repérer les changements.

Citons, avec une interface en français :

- ✓ Memoweb (Goto Software) : www.goto.fr
- ✓ Aspiweb (AalWay Software) : http://www.aalway.com/20/soft/aspiweb/
- ✓ Wysigot (ex e-catch La Mine) : www.wysigot.com

En anglais :

- ✓ Teleport Pro (Tenmax) : www.tenmax.com
- ✓ Web Whacker (Bluesquirrel) www.bluesquirrel.com
- ✓ Flash Site (Incontext) : www.incontext.com

Aller plus loin sur Wysigot

Wysigot est un logiciel de navigation/aspiration de sites Web orienté hors connexion et veille (mises à jour, recherches et comparaisons) conçu pour gérer des quantités de données importantes avec des réglages fins et/ou automatiques.

La version d'évaluation est illimitée dans le temps, mais ne permet pas de disposer de l'ensemble des fonctionnalités. Version payante : à voir, Wysigot est encore à sa version béta

Points forts

Mise en valeur des nouveautés dans les pages.

Recherche plein texte fonctionnelle dans les pages téléchargées

Saisie de formulaires hors connexion (page-réponse téléchargée lors de la prochaine connexion).

Prise en compte hors ligne des téléchargements futurs par un simple clic sur les liens désirés.

Fréquence des mises à jour des pages téléchargées automatique ou manuelle (les pages mises à jour sont signalées par le logiciel).

Téléchargement en parallèle jusqu'à 50 adresses simultanées

Sait se connecter, télécharger, et se déconnecter tout seul.

Export dans le format d'origine

Points faibles

Relative complexité d'utilisation par rapport à des outils comme Memoweb

LE PUSH (OU WEBCASTING)

Les principes du push

L'idée du Push : permettre aux internautes d'obtenir l'information désirée sans avoir à effectuer de fastidieuses recherches : l'information est ainsi " poussée " vers les destinataires et non plus " tirée ", comme elle l'est en "pull".

"Technologie fondée sur l'architecture client-serveur et dans laquelle un internaute s'abonne à une ou plusieurs chaînes thématiques auprès d'un serveur qui affichera automatiquement et à intervalles réguliers sur l'écran de l'utilisateur les renseignements sélectionnés" (Extrait de la terminologie internet de l'Office de la langue française www.olf.gouv.qc.ca).

Un logiciel client interroge donc régulièrement le serveur de push. S'il y a des données nouvelles à télécharger, le logiciel client les réclame et les enregistre sur le disque dur de l'utilisateur (réplication). Le logiciel "pousse" enfin les données vers l'utilisateur à sa demande.

Le modèle push est basé sur les mêmes principes de base que les techniques de diffusion hertzienne, via trois composants :

- ✓ L'émetteur d'information, qui permet la diffusion du contenu
- ✓ La chaîne ou canal, qui isole le contenu de ceux proposés par d'autres émetteurs
- ✓ Le récepteur, qui permet à l'utilisateur de recevoir le contenu (même principe qu'un tuner, mais avec consultation possible en mode asynchrone). Ce module est généralement gratuit et téléchargeable via l'Internet.

Le pionnier : Pointcast en 1995, avec un logiciel qui diffuse en direct sur un bandeau défilant une foule de titres de dépêches (Wall Street Journal, Fortune, CNN, The Times, Reuters etc). Il suffit de cliquer sur le titre pour lire l'article sur le site web de l'éditeur. Aujourd'hui propriété de la société Entrypoint, le phénomène Poincast a fait long feu – le mode de diffusion directe était notamment très gourmand en bande passante pour l'utilisateur.

Les navigateurs Explorer et Netscape ont intégré des agents push à partir de 1997, mais ces fonctionnalités sont abandonnées dans les versions actuelles.

Pourquoi le push n'a pas "décollé"...

- ✓ problèmes de contenu : pas assez ciblés.
- ✓ problèmes techniques : le push est plus adapté à une connexion permanente qu'à une connexion par modem, et est gourmand en ressources.
- ✓ problèmes de normalisation

Le renouveau du push

On assiste néanmoins depuis quelques temps à une « renaissance » du push

- ✓ au travers de sites qui diffusent de l'information (financière, actualité, emploi, appels d'offres…) en push. Mais la diffusion se fait là par e-mail (plus simple techniquement).

  Exemple : Net2one www.net2one.fr (actualité, revue de presse)
- ✓ au service des portails d'entreprises : OpenPortal4U (Arisem) avec Backweb, Portal One (Verity) avec Agentserver, Reuters avec Tibco…

Le phénomène Weblogs et les fils RSS

Les **weblogs**, ou 'blogues' sont nés de la rencontre du phénomène de simplification des techniques de publication sur Internet, et de celui de la volonté toujours présente de partager ses informations avec le plus grand nombre. Ce dernier phénomène, principe de base de l'Internet des premiers temps, connaît un jour nouveau avec ces nouveaux moyens de publication rapide, simple, souple...

(voir http://www.servicedoc.info/article.php3?id_article=28)

Intimement liée aux weblogs, mais sans en être une caractéristique, la **syndication** est une technique permettant d'afficher des données provenant (et offertes) d'autres sites, dans son propre site. C'est l'archétype de la gestion de contenu : c'est de l'information venant d'ailleurs, mise à disposition (éventuellement filtrée, reconfigurée...) du plus grand nombre.

La technique utilisée est issue du XML, mais très simplifiée : elle est d'ailleurs nommée RSS pour Really Simple Syndication, en fait une version dépouillée de la norme RDF. On peut en profiter directement si on utilise un CMS (Content management system, comme SPIP par exemple) qui en tient compte, mais il est aussi possible d'insérer un simple code java script (voir http://www.servicedoc.info/article.php3?id_article=57)

Les fils RSS servent alors soit à afficher dans un intranet ou dans un autre site internet, les infos publiées sur le weblog, un peu comme une fenêtre d'actualité, soit à être collectés via des RSS-aggregators, des lecteurs de fils. De la même façon que l'on ouvre un utilitaire de messagerie ou un lecteur de news, on peut "s'abonner" à tel ou tel fil et lire en direct les infos provenant de ressources diverses

(voir http://www.servicedoc.info/article.php3?id_article=100)

Très anecdotique, notamment en Europe, jusqu'en 2002, cette méthode, pourtant ancienne (les premiers weblogs et fils RSS datent de 1997) a récemment explosé, tant et si bien qu'elle a sinon révolutionné le circuit de l'information dans certains secteurs (par exemple celui de l'information et de l'auto-formation à la recherche documentaire), au moins influer très sensiblement sur la politique d'indexation des gros moteurs de recherche.

Exemples de fils RSS sur la recherche documentaire (en général, le weblog associé est la racine du site hébergeant le fil):

http://www.llrx.com/llrx.rss

http://www.librarystuff.net/libraryblogs/index.rdf

http://talk.lii.org/tipoftheweek/index.rdf

http://morinn.free.fr/b2/b2rss.php

http://joueb.com/influx/rss.shtml

http://google.blogspace.com/index.xml

Les méta-moteurs clients "off-line"

Ils remplissent les mêmes missions de base que leurs confrères du "on-line", mais disposent de fonctions plus évoluées, variées selon les produits :

- ✓ Enregistrement des recherches dans des dossiers
- ✓ Traduction "sophistiquée" des équations de recherche (au-delà du ET, du OU, et de l'expression exacte)
- ✓ Traitement linguistique des requêtes (langage naturel)
- ✓ Interrogation de différents moteurs et bases de données spécialisées permettant d'accéder à du contenu non référencé par les moteurs classiques (web invisible). Certains outils laissent l'usager libre d'ajouter manuellement de nouveaux moteurs, bases de données, voire sites et pages web à interroger dans le cadre de nouveaux groupes de sources.
- ✓ Téléchargement des pages de résultats, édition de rapports personnalisés en html
- ✓ Mise à jour des recherches, voire automatisation de la surveillance : paramétrage de la périodicité des requêtes, alertes par mail
- ✓ Raffinement des recherches (ou filtrage) : La fonction "raffiner" ou "filtrer" permet d'effectuer une recherche spécifique sur des documents préalablement téléchargés. On utilise alors le moteur de recherche intégré au métamoteur, qui offre des fonctions avancées de recherche avec les opérateurs classiques mais aussi le PRES (permet de rechercher une page où les mots-clés sont distants d'un nombre défini de mots) et les parenthèses. On peut ainsi télécharger un corpus important de pages web sur une thématique assez large, et effectuer ensuite rapidement des recherches beaucoup plus précises pour l'étude des sous-thèmes.
- ✓ Suivi des liens hypertextes des liens considérés comme pertinents
- ✓ Surveillance de pages de résultats, éventuellement groupées dans des dossiers : Les changements sont indiqués par la présence d'une icône modifiée, ou envoyés par mail.
- ✓ Relevance feed-back : l'avis de l'usager est demandé sur les documents ramenés
- ✓ Traitement des documents résultats : traductions, résumés automatiques, mise en exergue des extraits pertinents
- ✓ Traitement automatique de l'ensemble du corpus de résultats, cartographies

Parmi les plus utilisés à l'heure actuelle, citons :

Copernic : www.copernic.com (voir ci-après)

Bullseye (Intelliseek) : www.intelliseek.com

Strategic Finder (Digimind) : www.strategicfinder.com (voir ci-après)

La plupart de ces outils sont aujourd'hui proposés en version "serveur" pour être installés au sein des entreprises clientes, accessibles par exemple via l'intranet. Ainsi, en mars 2002, Copernic a lancé une application logicielle serveur pour les entreprises Copernic Empower : la solution compte 4 modules complémentaires (indexation, module de recherche en parallèle sur internet, intranet), module de veille (monitoring de documents), module de résumé (identifie les concepts clés et extrait les phrases les plus "importantes" du document).

Copernic

Le logiciel Copernic a été lancé fin 1997 par la société Agents Technologies Corporation, et compte aujourd'hui vingt millions d'utilisateurs avec une couverture de 46 % aux Etats-Unis, 47 % en Europe et 7 % en Asie. Il effectue des recherches sur plusieurs outils francophones ou internationaux (paramétrage des outils et du nombre de résultats par moteur).

En octobre 2002, la gamme "Copernic Agent" remplace le logiciel Copernic 2001, avec une architecture et une interface renouvelées. La version "Basic" reste gratuitement téléchargeable, et donne déjà une bonne idée du produit. Les versions Personal et Professional offrent bien sûr plus de fonctionnalités, notamment la mise à jour automatique des recherches selon la périodicité souhaitée et avec alertes par mail. Elle permet aussi d'automatiser le téléchargement ou la validation de documents ainsi que le raffinement des recherches.

A noter : Copernic a passé un accord avec Espotting en février 2002 (en Grande Bretagne, puis sur le web français et le web allemand au fur et à mesure de l'installation d'Espotting, et aussi d'Overture, car cela se fera aussi) pour l'affichage de 5 liens maximum en fonction des mots-clés demandés

Nouvelles fonctionnalités :

- ✓ Nouvelle interface plus complète, mais aussi beaucoup plus complexe !
- ✓ Intégration d'un agent d'alerte, comme Bullseye et Strategic Finder (surveillance automatique de changements dans les pages web)
- ✓ Résumés des pages (extraits pertinents : technologie "Copernic Summarizer"
- ✓ Intégration avec IE et Microsoft Office
- ✓ Catégories de recherche personnalisables (mais impossibilité de "rentrer" de nouveaux moteurs
- ✓ Filtrage des résultats selon la langue, le domaine, etc. et groupement des résultats selon ces mêmes filtres
- ✓ Améliorations diverses : fonctions automatisées de veille et de recherche, recherche de mots-clés dans les pages web, suppression de résultats non pertinents, personnalisation

Les versions Personal et Professional permettent d'accéder à plus de 1000 sources d'information spécialisées, groupées dans quelque 125 catégories de recherche.

Strategic Finder

Strategic Finder, lancé en 2000 par la société Digimind permet la recherche, non seulement sur un certain nombre d'annuaires et moteurs, mais aussi sur des sites spécialisés (actualité, entreprise, juridique…). On obtient la liste en cliquant sur «Sources». La version payante de SF permet d'ajouter de nouvelles sources ou «plug-ins» (certains de ces plug-ins étant proposés sur abonnement en sus). http://www.strategicfinder.com/.

SF accepte les parenthèses, les guillemets, les opérateurs AND, OR, NOT, les équations étant traduites en fonction du langage d'interrogation de chaque outil. La requête peut être soumise aux catégories de sources souhaitées, avec la possibilité de paramétrer pour chaque source le nombre de résultats souhaités.

La version 2 du logiciel est beaucoup plus rapide et permet d'interroger jusqu'à 4000 sources réparties dans plus de 50 secteurs d'activité.

- ✓ Téléchargement des pages souhaitées
- ✓ Un "résumé" est proposé permettant d'avoir le contexte entourant le mot-clé de la requête
- ✓ Mise à jour des recherches
- ✓ Création de dossier de pages que l'on peut mettre sous surveillance (une icône spécifique indique les modifications) ; Les pages d'un dossier peuvent provenir de différentes recherches et on peut même y intégrer une autre page (en utilisant le navigateur intégré)
- ✓ Filtrage : Par défaut, Strategic Finder utilise la requête utilisée pour lancer la recherche. Mais on peut personnaliser le filtrage en modifiant la requête qui est affichée. Pour cela, vous pouvez créer une requête booléenne (en utilisant les opérateurs AND, OR, NOT, (), "").
- ✓ « Approfondir » est une fonctionnalité qui permet de prolonger sa recherche en trouvant de nouvelles informations en ramenant plus de résultats pour telle ou telle source.

  te des catégories de sources. Il vous suffit alors de cocher ou décocher les catégories de votre choix et de cliquer sur "OK" pour lancer la recherche à nouveau. Vous pouvez aller plus loin en rentrant dans le détail de chaque catégorie (bouton Détail) et en modifiant le nombre d'informations à ramener par chaque moteur dans une catégorie.
- ✓ Possibilité de se créer ses propres catégories de sources (au prix de quelques efforts !)

Les agents d'alerte

Ils signalent par mail les modification d'une page ou d'un site web, selon des critères plus ou moins fins.

On distingue :

- ✓ les agents d'alerte web "serveurs" (Digimind Monitor, Infominder, Get-Updated) : l'utilisateur se connecte sur le serveur de la société éditrice du produit, donne ses directives et reçoit ses alertes généralement par mail ou les consulte sur un espace privé. L'agent peut aussi être directement installé sur le serveur de l'entreprise cliente. Il fonctionne alors selon le même principe général, mais avec une installation "privée" en intranet ou extranet.

Digimind Monitor : www.digimind.fr

Infominder : www.infominder.com

Get updated : www.getupdated.com

- ✓ Les agents d'alerte "clients" qui nécessitent le téléchargement d'un logiciel particulier : Webspector, WebSite Watcher ou Watznew

Webspector www.illumix.com/webspector

Website Watcher aignes.com

Notons que certains agents d'alerte se spécialisent par grandes thématiques, tel TracerLock pour l'actualité. D'autres existent aussi pour les webmasters qui, en plaçant un bouton sur leur site, permettent à leurs visiteurs d'être avertis des nouveautés par mail.

Le paysage des agents d'alerte est loin d'être stabilisé actuellement. Notons que la plupart des grands méta-moteurs intègrent aujourd'hui la surveillance de pages web.

Les outils de "text-mining"

Ils traitent automatiquement de grandes quantités d'informations textuelles issues de bases hétérogènes et peuvent faire :

- ✓ de l'indexation automatique
- ✓ de l'extraction terminologique
- ✓ de la détection de liens entre les mots (algorithmes statistiques)
- ✓ de la visualisation graphique (notamment cartographique)
- ✓ de l'interaction dynamique avec l'utilisateur

Quelques exemples

Leximine	www.lexiquest.fr
Wordmapper / Question Data	www.grimmersoft.com
Pericles	www.datops.fr
Semio	www.semio.com
Tetralogie	http://atlas.irit.fr
Tropes	www.acetic.fr
U-map (module de See-K)	- > http://www.trivium.fr/fr/index.htm

Principes d'une veille efficace sur Internet

Dire que l'on "fait de la veille sur Internet" est un abus de langage. En fait, on utilise Internet comme un outil de surveillance des entreprises, des marchés, des technologies, des évolutions de la société…

L'apport d'Internet par rapport dans une démarche de veille :

➔ Une information ouverte, disponible à tout moment, souvent à faible coût

➔ Une information régulièrement actualisée

➔ Des informations multi-sources, multidisciplinaires (le fonctionnement réseau étant idéal pour la veille).
➔ Une information numérisée, pouvant être triée et exploitée rapidement.

Mais il ne faut pas oublier les aspects négatifs :

➔Risque de désinformation : une information "orientée" et donc pas toujours fiable.

➔ Risque de se "noyer" dans l'information.

➔Une information parfois difficilement accessible (barrières des langues, services payants,...).

➔ Une information en perpétuelle évolution et donc instable

➔ Une relation temps-coût / valeur intrinsèque de l'information obtenue pas toujours facile à maîtriser.

MÉTHODOLOGIE À METTRE EN ŒUVRE

- ✓ **Définition des cibles de veille**

La mise en place d'un processus de veille sur Internet s'appuie sur un ciblage de la veille défini à partir des objectifs et du positionnement stratégique de l'entreprise ou organisation sur ses différents marchés.

Concrètement, c'est la réponse aux questions : Qui surveiller sur Internet ? Sur quel thème ?

- ✓ **Inventaire des sources connues sur Internet**

Lesquelles sont pertinentes par rapport à l'étape précédent, pour quel thème ?

- ✓ **Recherche d'autres sources pertinentes**

Pour cette étape, on procédera d'abord à la constitution évolutive d'une liste arborescente des mots-clés des différents thèmes stratégiques, traduits en anglais, et si nécessaire, dans d'autres langues.

Cette liste peut évoluer en fonction des ressources trouvées, et de l'évolution du vocabulaire du domaine.

Il s'agit ensuite de constituer les équations de recherche les plus pertinentes pour chaque thème de veille pour les proposer à différents moteurs.

On peut aussi travailler à partir de répertoires hyper-spécialisés et suivre les liens proposés (les répertoires généralistes sont de peu de secours, les thèmes de veille étant généralement assez pointus).

- ✓ **Mise sous surveillance des couples "ressource Internet"/ thème de veille**

On obtient donc une liste de ressources clés sur Internet qui pourra évoluer dans le temps (ne pas oublier les forums et listes de diffusion).

Après un choix d'agents à utiliser (agent d'alerte on-line ou off-line), les pages clés (par exemple pour un concurrent les pages Produits, News et Offres d'emploi) sont mises sous surveillance automatique.

Les équations de recherche peuvent être soumises régulièrement aux moteurs de recherche sélectionnés (voire méta-moteurs) pour être averti de la présence de nouveaux acteurs intéressants.

L'utilisation parallèle de logiciels de cartographie sur les résultats de ces requêtes, (téléchargés préalablement sur le disque dur) peut permettre de repérer des évolutions faibles ou tendances sur des marchés mouvants.

Avec ces outils, il peut être intéressant de travailler en plus sur des thèmes de veille élargis.

- ✓ **Collecte et Sélection des informations recueillies**

Rappelons que dans une optique de veille, on ne se base pas sur des données rétrospectives, ni même quantitatives et certaines, mais sur des signaux fragmentaires dits "faibles" : en ne conservant que les informations réellement stratégiques pour l'entreprise, la sélection consiste à affiner le travail de collecte et permet l'analyse.

L'évaluation de la fiabilité de la source et de l'information sont bien sûr très importantes, mais peuvent se faire a posteriori.

On quitte alors le "cycle Internet" pour intégration des données dans le système d'information de l'entreprise, diffusion et exploitation.

LA VEILLE AUTOMATISÉE

- ✓ Surveillance de pages web
- ✓ Surveillance de sites web
- ✓ Surveillance de dossiers de pages web
- ✓ Surveillance de recherches sur un outil web / plusieurs outils web / une base de données / plusieurs bases de données
- ✓ Surveillance de catégories d'un répertoire
- ✓ Surveillance de catégories de ressources (actualités, articles de presse, appels d'offres, offres d'emploi, communiqués de presse, informations financières, etc.)

La veille "manuelle" (sans l'utilisation des agents)

- ✓ **Repérer les nouveaux sites dans un domaine :**

La meilleur méthode : bouche à oreille, abonnement à des listes de diffusion, à des e-zines et newsletters.

Les services "Nouveautés" des moteurs sont trop généralistes pour être efficaces. Si votre veille s'exerce sur un secteur géographique donné, n'oubliez pas les annuaires et moteurs géographiques. Il existe aussi des sites qui informent de la création de nouveaux sites web (ex : Interneto http://www.interneto.fr/ ou Actusite www.actusite.com), mais les classifications ne sont souvent pas fines.

- ✓ **Suivre l'actualité :**

Cela est possible grâce aux services de diffusion personnalisée, en push, comme Newspage ou Net2one (voir plus haut).

- ✓ **S'abonner aux périodiques électroniques des sites portails importants**

Y sont indiqués le plus souvent non seulement les nouveautés du site, mais aussi du secteur concerné.

- ✓ **Quelques pistes en veille technologique :**

➔Utiliser les newsgroups et les listes de diffusion scientifiques (généralement de bonne qualité)

➔Utiliser les fonctions d'alerte des grands fournisseurs d'information : Uncover Reveal (diffusion de tables des matières sur profils via e-mail), ou le TOC Alert de Publist.com, Inist (veille documentaire)…

➔Accès plus facile et moins cher à des bases de données, par exemple de brevets (INPI www.inpi.fr)

- ✓ **Quelques pistes en veille concurrentielle ou marketing:**

➔Suivre les sites web de sociétés avec un agent d'alerte comme The Informant ou Webspector, ou un aspirateur de sites,… ou manuellement

➔Utiliser les services Push type PRLINE ou Companynews (www.prline.com)

➔Utiliser les newsgroups en faisant des recherches par noms de sociétés (attention à la fiabilité de l'information !) Cela peut être toutefois un bon moyen de détecter les rumeurs et les bruits qui circulent.

En guise de conclusion

On a vu les limites actuelles des outils de recherche "classiques" : annuaires, moteurs et méta-moteurs. On a vu également que les "agents intelligents" sont prometteurs mais ne sont pas encore totalement adaptés aux besoins des professionnels de l'information. Ces outils vont encore évoluer, en incluant des fonctions de plus en plus évoluées (résumé automatique, traduction, gestion du langage naturel..). Une perspective d'évolution intéressante concerne l'avenir des méta-données et le XML.

L'évolution des méta-données

D'après une étude du cabinet eMetrie sur 30000 pages (résultats de 300 requêtes sur 10 outils de recherche francophones), les balises méta seraient peu utilisées (50 % des pages ne comportaient aucune balise).

Par ailleurs, les outils de recherche tiennent de moins en moins compte de ces balises pour leur tri de pertinence (Google, Fast, Lycos ne les utilisent pas du tout).

Pour pallier à la "faiblesse" des balises méta classiques, certains groupements travaillent à mieux décrire les documents sur Internet. On pourra utilement se référer au "Dublin Core", métadonnée de 15 éléments destinée à la description générale des documents, qui est d'ores et déjà utilisée via les balises méta par certains organismes, y compris en intranet. Le Dublin Core, considéré comme un bon candidat pour une norme internationale, est le fruit du travail depuis 1995 d'une cinquantaine de chercheurs et professionnels issus du monde de la documentation et des bibliothèques, de l'informatique, de la codification des informations. L'ensemble fut initié par l'OCLC (Online Computer Library Center) en accord avec le NCSA (National Center for supercomputing applications). Le Dublin Core doit son nom à la première réunion de travail en juin 95 à Dublin Ohio dans les locaux de l'OCLC.

Notons le format RDF (Resource Description Framework) en cours de standardisation pour les méta-données : il permet de présenter un élément d'information –qu'il s'agisse d'un site, d'une page, etc. – dans une syntaxe compatible XML (voir ci-dessous). La grande difficulté résidera dans l'impossibilité d'imposer cette norme pour la publication sur Internet (pas de contrôle), mais il semble qu'avec son avatar, le RSS (Really Simple Syndication), le RDF tende à devenir une norme de fait.

Du html au xml

XML (Extensible Markup Language, ou Langage Extensible de Balisage) est le langage destiné à succéder à HTML sur le World Wide Web. Comme HTML c'est un langage de balisage (markup), c'est-à-dire un langage qui présente de l'information encadrée par des balises.

Mais contrairement à HTML, qui présente un jeu limité de balises orientées présentation (titre, paragraphe, image, lien hypertexte, etc.), XML est un métalangage, qui va permettre d'inventer à volonté de nouvelles balises pour isoler toutes les informations élémentaires (titre d'ouvrage, prix d'article, numéro de sécurité sociale, référence de pièce…), ou agrégats d'informations élémentaires, que peut contenir une page Web.

La tâche est aujourd'hui de définir des ensembles de balises et de règles pour les différents domaines, et de très nombreux groupes de travail se sont mis en place. Il vont pouvoir standardiser la structure d'un document chimique comme d'un type de contrat. Le langage permet également une utilisation plus flexible des liens hypertextes placés dans un fichier spécial… ***Toutefois, les répercussions à court terme se jouent principalement dans les intranets d'entreprise, et non sur le Web.***

POUR EN SAVOIR PLUS…

RECHERCHE D'INFORMATION

- . Trouver des informations sur le web / Olivier Andrieu - Eyrolles, 2001
- La recherche d'Information sur Internet : Outils et méthodes (Risi) / Jean-Pierre Lardy - ADBS, Coll. Sciences de l'information, série Recherches et documents, 2001 -
- Intelligence stratégique sur internet / Carlo Revelli. - Dunod, 2000.
- Recherche et veille sur le web visible et invisible / Béatrice Foenix Riou. - Technique et documentation Lavoisier, 2001
- The invisible web : Uncovering information sources search engines can't see / Gary Price, Chris Sherman, 2001

SITES D'AUTOFORMATION A L'INTERNET

- **Apprendre l'Internet** *www.learnthenet.com/french*
- **Netexpress** *www.wanadoo.fr/animation/internautes/netexpress*
- **UNGI** *www.imaginet.fr/ime*

SITES CONSACRES A LA RECHERCHE D'INFO SUR INTERNET

- **GIRI** *www.bibl.ulaval.ca/vitrine/giri*
- **RIsI** *www.adbs.fr/adbs/viepro/sinfoint/lardy/risi.htm*
- **Sapristi !** (**info scientifique)** *http://csidoc.insa-lyon.fr/sapristi/digest.html*
- **Netsesame (info économique)** *www.devinci.fr/infotheq (rubrique « Outils Internet »)*
- **Abondance (outils de recherche)** *www.abondance.com*
- **Agentland (agents)** *www.agentland.fr*
  - La lettre du bibliothécaire québecois www.sciencepresse.qc.ca/lbq/lbq.html 7000 sites commentés sur http://www.sciencepresse.qc.ca/repertoires.html
  -

SITES DES ORGANISMES DE L'INTERNET

- **The World Wide Web Consortium** www.w3.org
- **AFNIC** *www.nic.fr*
- **Internet Society (ISOC)** www.isoc.asso.fr
- **IAB** *www.iab.org/iab*
- **Internet.gouv.fr** *www.internet.gouv.fr/francais/index.html*

LISTES DE DISCUSSION

- ADBS-INFO adbs-info@cru.fr

- BIBLIO-FR biblio-fr@cru.fr

- MOTRECH motrech-abonnement@egroups.fr