Rechercher l'information sur Internet : approfondissement des méthodes

Support de cours commun

ADBS – Aout 2005

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

dd/06/yy

SOMMAIRE

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.

10. **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 leur caractère mouvant, décloisonné, international.

L'information disponible sur Internet

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
- ✓ 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).

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 en 2002, certainement 10 fois plus en 2005), 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 20 milliards de pages, sans compter les informations contenues dans les bases de données. Google a dépassé début 2005 les 8 milliards de pages indexées (chiffre doublé depuis l'an dernier).

Les études sérieuses sont malheureusement rares et commencent à sérieusement dater. Nous les citons ici pour référence.

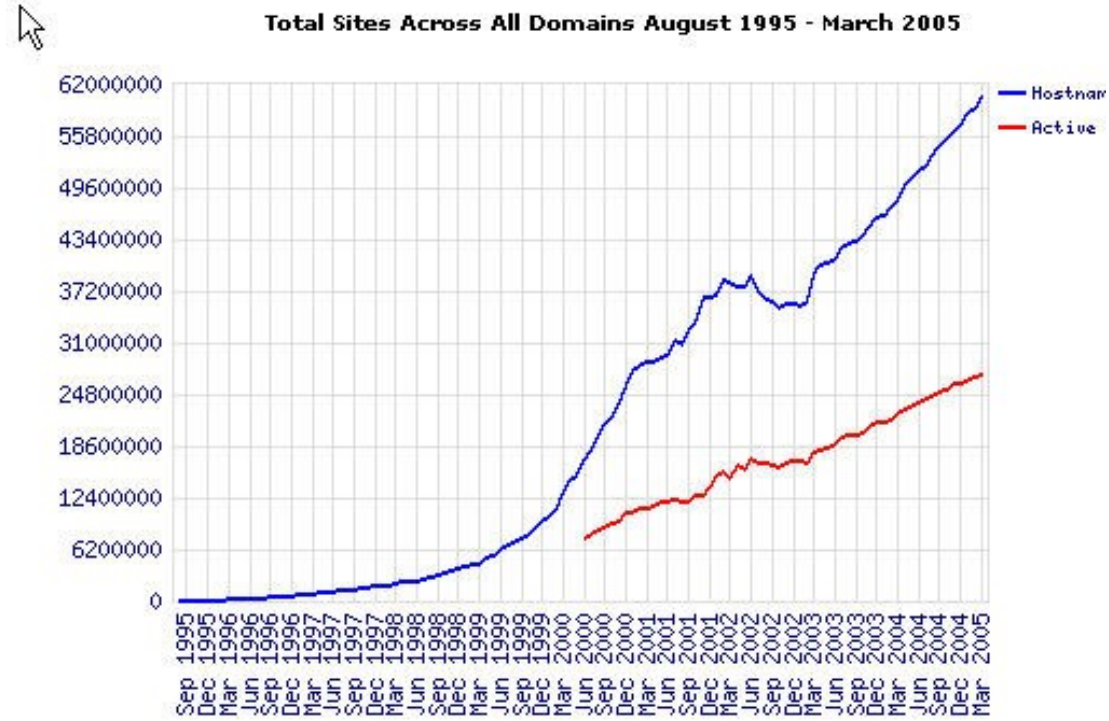
Voir sur le blog Motrech l'article de mars 2005 "Le web, un autre Univers en expansion" http://motrech.blogspot.com/2005/03/le-web-un-autre-univers-en-expansion.html. Cet article récapitule notamment les liens utiles concernant les études effectuées.

Voir aussi sur http://c.asselin.free.fr/french/webenchiffre.htm

- ✓ Une nouvelle étude, de deux chercheurs des universités de Pise (Italie) et de l'Iowa (Etats-Unis) réalisée en janvier 2005, donnerait une taille approximative de 11,5 milliards de pages indexables sur le web http://www.cs.uiowa.edu/ãsignori/web-size/ Elle semble toutefois en deça de la réalité, surtout si l'on en croit Yahoo qui annonce en août 2005 un index de 19,2 milliards de pages ! (donné à 6,6 milliards par l'étude citée).
- ✓ 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

Il y avait plus de 67 millions de sites web au niveau mondial en septembre 2005 selon l'organisme de référence depuis 1995 Netcraft (www.netcraft.com), pour 1 million en avril 97, 10 millions en 2000, 20 millions sept mois plus tard (50 millions en août 2004, 60 millions en mars 2005, 63 millions en mai).

Evolution du nombre de sites Internet depuis dix ans



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. Malgré tout, l'exploration est loin d'être exhaustive, car ne tient pas compte de ce qui se passe après le premier.

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/)

Les sites fédérateurs (portail vertical ou vortal)

Les sites fédérateurs ou portails sont des outils de recherche incontournables dans de nombreux domaines. Ils sont conçus au départ autour d'un ou plusieurs répertoires spécialisés (sites web, entreprises, adresses utiles, événements, etc.). Ils intègrent le plus souvent actualité et autres services. Ils peuvent évoluer vers le commerce électronique ou la place de marché.

Ainsi, les outils 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.

Le phenomene Weblogs et fils RSS

"De façon très synthétique, un "blog" (ou "weblog") est un site Web personnel composé essentiellement d'actualités (ou "billets"), publiées au fil de l'eau et apparaissant selon un ordre ante-chronologique (les plus récentes en haut de page), susceptibles d'être commentées par les lecteurs et le plus souvent enrichies de liens externes." (définition du weblog Pointblog, consacré au phénomène du blog, http://www.pointblog.com dans la rubrique l'ABC du blog). L'auteur considère que tous les blogs, dans leur énorme diversité actuelle, ont en commun leur caractère individuel ou "unipersonnel", l'utilisation d'outils dynamiques, la liberté de ton, et l'interconnexion.

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

12000 nouveaux blogs seraient créés chaque jour (400000 nouveaux billets chaque jour) pour un total d'environ 4 millions.. Source http://www.sifry.com/alerts/archives/000387.html à partir de Technorati. Notons toutefois que tous les weblogs (et loin s'en faut) ne sont pas actifs, et que beaucoup, héritiers des "journaux intimes" ne présentent pas le moindre intérêt pour les professionnels.

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 (cf page 10).

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 aussi la partie "Les agents d'actualité" page 54

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):

En anglais :

Site du Law Librarian Resource Exchange : http://www.llrx.com/llrx.xml

Site Librarian and Information Science News (nombreux fils thématiques)
http://www.lisnews.com/feeds.shtml

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

http://google.blogspace.com/index.xml (Google Weblog)

Librarian Index to the Internet LII http://lii.org/ et fil rss : http://lii.org/ntw.rss

En français :

Influx http://joueb.com/influx/ et http://influx.joueb.com/index.rdf

Figoblog : http://www.figoblog.org/ et fil =
http://figoblog.ouvaton.org/backend.php?format=rss092documents&charset=iso-8859-1

Biblioacid : http://www.biblioacid.org/ et fil = http://feeds.feedburner.com/BA_rss1

Blogokat : http://blogokat.canalblog.com/ et fil = http://blogokat.canalblog.com/rss.xml

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, javascripts modifiant le contenu, technologies propriétaires.
- ✓ 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)

On ne connaît pas du tout la taille du web invisible : Selon une étude de la société BrightPlanet (Completeplanet) parue en juillet 2000, il y avait à cette époque déjà 350 000 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). D'après eux, aujourd'hui, les 60 bases de données les plus importantes contiennent déjà environ 84 milliards de pages.

FAQ de la société Briht Planet sur le "deep web" (en anglais)

http://brightplanet.com/deepcontent/deep_web_faq.asp#DeepWebSize

Une certitude : le web invisible croît plus rapidement que le web visible, du fait de la multiplication des bases de données à interface web, et de l'explosion du web dynamique.

A noter : les fichiers pdf ou flash, autrefois partie intégrante du web invisible, sont aujourd'hui indexés par plusieurs moteurs, Google en tête.

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.

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.

- ✓ 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. Exemple, voir les forums de Liberation.

✓ Vers le web sémantique

« The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in coopération ». Tim Berneers-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001.

"Web intelligent dans lesquel les informations, auxquelles on donne une signification bien définie, sont reliées entre elles de façon à ce qu'elles soient comprises par les ordinateurs, dans le but de transformer la masse des pages web en un index hierarchies et de permettre de trouver rapidement les informations recherchées" (Grand dictionnaire terminologique. http://www.granddictionnaire.com/

Voir aussi :

- le site du World Wide Web Consortium consacré au sujet http://www.w3.org/2001/sw/
- l'explication de Charles Népote à un atelier à Autrans : http://autrans.crao.net/index.php/AtelierWebS%E9mantique

  "Le Web Sémantique est une extension du web, il est fondé sur le web. Il en utilise toute l'infrastructure technique -- les langages et les protocoles -- en ajoutant certains protocoles. Pour les utilisateurs finals, le Web Sémantique ne propose visuellement pas de modification des interfaces. Un site Web sémantique est visuellement identique à un site web classique. Le Web Sémantique vient s'ajouter au web, sans le remettre en cause. C'est une valeur ajoutée, une extension que l'on est pas obligé d'employer.
- le wiki consacré au sujet (communauté francophone) : http://websemantique.org/PagePrincipale
- le portail animé par Stefan Decker de l'université de Stanford : http://semanticweb.org

XML

XML (Extensible Markup Language, ou Langage Extensible de Balisage) est le langage destiné à permettre l'avénement du web sémantique.. 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…

Le langage RDF (Resource Description Framework) est à la base du web sémantique, en permettant d'attribuer un sens à une ressource en ligne, à l'aide de triplets "sujet-verbe-complément". RDF est en cours de standardisation pour les méta-données : il permet donc de présenter un élément d'information –qu'il s'agisse d'un site, d'une page, etc. – dans une syntaxe compatible XML. La grande difficulté résidera dans l'impossibilité d'imposer cette norme pour la publication sur Internet (pas de contrôle).

Informations de base sur les méta-données

Il s'agit au départ de balises du langage html qui 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">
```

Pour pallier 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.

A noter : Voila est l'un des seuls moteurs à proposer la recherche avancée sur les balises descriptions et mots-clés.

Le nouveau paysage des outils

Les évolutions majeures

Depuis deux ans, le paysage des outils de recherche a beaucoup changé. Quelles sont les évolutions majeures :

- ✓ Baisse des recherches "annuaires" : les internautes utilisent beaucoup plus les moteurs, d'autant qu'il est souvent difficile de savoir d'où proviennent les résultats (cf évolution de l'interface de Yahoo).

  On remarque toutefois que les grands portails permettent une recherche plus aisée sur moteurs ou répertoires avec souvent un système d'onglets qui évite de retaper sa question (fonctionne aussi pour chercher sur les images, les news, les forums, etc.)

- ✓ Diminution du nombre d'outils généralistes : on a assisté
  - à la disparition de nombreux moteurs et répertoires : Looksmart France, Ctrouvé ; Excite, Northern Light, Ecila, Lokace, etc.
  - à une forte concentration : Infospace (nouvel acquéreur de Excite) détient Dogpile et Metacrawler, Ask Jeeves a racheté Teoma, Infonie repris par Tiscali, Ixquick a acheté aussi Debriefing, Profusion racheté par Intelliseek, Savvy Search racheté par Cnet… Mais surtout (c'était le scoop de l'année 2003), Overture a racheté le moteur AltaVista et la division WebSearch de Fast, l'éditeur du moteur AlltheWeb, avant de se faire racheter en juillet 2003 par Yahoo qui a également acquis Inktomi.
- ✓ Raz de marée Google, la part de marché mondiale est de l'ordre de 56 % (70 % en France).
- ✓ Pérennisation d'un modèle économique basé sur la publicité
- ✓ De plus en plus d'outils spécialisés : portails, répertoires, moteurs, voire méta-moteurs, avec parfois une insertion payante des sites.
- ✓ Un gros travail des outils autour de l'aide à l'utilisateur (reformulation des questions, correcteurs orthographiques, pages similaires, etc.) avec des interfaces évolutives.
- ✓ Une évolution vers la personnalisation (cf page 15).
- ✓ L'apparition de technologies innovantes sur le web (auparavant réservées aux applications verticales ou aux intranets) : clustering, cartographie (voir page 15).
- ✓ Les voies de la régionalisation, voire de la géolocalisation sont explorées : Voir par exemple, le moteur Mirago ou Indexa.
- ✓ Les outils explorent également la "navigation sociale", qui utilise des techniques de filtrage collaboratif (lien avec popularité important) : les gens qui ont aimé ce site (ce livre, ce service) ont aussi aimé, sont aussi abonnés, ont aussi acheté…
- ✓ De gigantesques bases de données se créent sur les internautes, leurs habitudes, leurs préférences (même si cela reste anonyme).
- ✓ De nombreux services en ligne (évolution parallèle à la personnalisation) dont les services d'alerte : nouveau site, nouvelle page, nouvelle actu répondant à certains critères définis par l'usager.

Qui "outille" qui ?

De nombreux sites moteurs ou répertoires travaillent avec des bases de pages crawlées ou des répertoires de sites et des technologies appartenant à d'autres (par exemple, le répertoire utilisé par Google est le Open Directory, le moteur MSN utilise actuellement Yahoo et Google travaille avec Ask Jeeves. Les accords se font et se défont, et il n'est pas toujours facile de suivre et de savoir qui travaille avec qui.

Pour vous aider :

- ✓ Le site Abondance qui propose un tableau bien pratique (attention aux mises à jour toutefois) : http://docs.abondance.com/portails.html
- ✓ Le "search engine decoder" www.search-this.com/search_engine_decoder.asp

Les moteurs de recherché personnels (desktop search)

Actuellement, la plupart des grands moteurs (voire méta-moteurs, comme Copernic) proposent gratuitement aux internautes de disposer de leur technologie pour indexer le contenu de leur disque dur, voire des réseaux internes de l'entreprise, et effectuer des recherches (de nombreux formats reconnus).

Voir par exemple :

- ✓ Copernic desktop search (très bon outil) : http://www.copernic.com/en/products/desktop-search/index.html
- ✓ Google Desktop Search : http://desktop.google.com (V2 disponible depuis peu avec de nombreuses fonctions nouvelles)

Les barres d'outils

Aujourd'hui, la plupart des grands moteurs proposent leur barre d'outils (toolbar) qui s'installent sur le navigateur (malheureusement souvent à Internet Explorer, mais pas toujours). Ces barres offrent alors un certain nombre de fonctionnalités très pratiques, dont la première reste bien sûr la recherche directe sur le web, sans avoir à aller sur le site de son outil préféré. Certaines permettent de rajouter les moteurs de son choix (exemple, celle de Copernic).

Fonctionnalités proposées :

- ✓ Recherche sur le web (pages, documents multimedia, etc.)
- ✓ Rajout de moteurs de son choix (exemple Copernic toolbar)
- ✓ Recherche au sein de la page visitée
- ✓ Informations sur la page visitée
- ✓ Traduction
- ✓ Mise en surbrillance des termes de la requête
- ✓ Blocage de fenêtres pop-up (exemple Google)
- ✓ Historique de recherche
- ✓ Personnalisation des affichages (exemple www.toolbarbrowser.com)

Voir la page dédiée sur le site de C. Asselin : http://c.asselin.free.fr/french/toolbar.htm

La personnalisation

C'est l'un des grands chantiers pour les outils de recherche (qui leur permet aussi de mieux fidéliser leurs "clients"), et un challenge pour l'avenir. Il s'agit de permettre aux internautes d'interagir avec leur moteur, en leur permettant, au-delà d'une simple personnalisation de l'interface et des préférences (devenue assez classique), de stocker des éléments d'information dans un espace dédié du serveur de l'outil, de conserver un historique de ses recherche, de surveiller des requêtes, de partager de l'information avec d'autres personnes, etc…

Toutefois, comme le dit très justement Jérome Charon dans son blog Motrech (motrech.blogspot.com) : "Mais la personnalisation est un sujet délicat. Il flirte dangereusement avec la confidentialité"…

Quelques exemples (tous de l'année 2004, c'est dire que le phénomène est récent) :

- ✓ **Ujiko** (www.ujiko.com), lancé par l'équipe du méta-moteur Kartoo à partir de la technologie Yahoo. L'outil joue à fond la personnalisation, en permettant la mémorisation et personnalisation des recherches. De plus, les URL des documents peuvent être cochés, anotés, filtrés, classés ou supprimés. Lorsqu'on clique sur un des résultats d'une recherche, la page est conservée en mémoire et ultérieurement placée en tête de résultats. Une nouvelle version a vu le jour tout récemment
- ✓ **Ask Jeeves** propose My AskJeeves (http://myjeeves.ask.com) qui permet de sauvegarder des liens obtenus comme résultats lors de requêtes sur le moteur et de les gérer par la suite (mention "save" à côté d'un résultat) dans des catégories, comme un bookmarks, et de les annoter. La version 1.2 permet de sauvegarder des images, d'utiliser des "dossiers virtuels" pour ranger les sites, de lancer de nouveaux filtres de recherche, etc.
- ✓ **Yahoo** (voir page 31) a lancé My Yahoo search, qui permet de sauvegarder les résultats (voire de les commenter et d'y effectuer des recherches). On peut aussi exclure un lien de futures requêtes. A noter une fonction de partage avec d'autres personnes.
- ✓ **Meceoo** (voir page 34) permet aux usagers de personnaliser leurs recherche grâce à une liste de sites exclus ou au contraire une liste de sites à explorer spécifiquement.
- ✓ **Amazon** a lancé le moteur A9 (www.a9.com) dont les résultats web sont fournis par Google. Cet outil, également axé sur le catalogue d'ouvrages de la librairie en ligne, permet notamment de conserver un historique de ses précédentes recherches
- ✓ **Looksmart a racheté Furl** (furl.net) gestionnaire de favoris qui propose des fonctionnalités du type : sauvegarde de résultats de recherche et de pages web, gestion et traitement de ces données dans des archives personnelles.

Le clustering ou catégorisation automatique

Les moteurs utilisant le clustering, après collecte et indexation automatique, répondent aux requêtes des usagers en structurant dynamiquement le corpus de résultats (une visualisation graphique peut y être associée, cf ci-dessous). Ces outils utilisent des technologies de textmining pour extraire directement des structurations de grands corpus de documents. Sur le web, où l'on ne peut pas faire référence à des dictionnaires ou ontologies préexistantes (type thésaurus ou autre), la catégorisation se fait dynamiquement en fonction de la requête.

Les techniques utilisées sont essentiellement statistiques (méthode des mots associés avec matrices de cooccurrences) pour constituer des clusters.

Voir notamment :

- ✓ **Les méta-moteurs Vivisimo** (voir page 35) et **Killerinfo** (page 35)
- ✓ **Exalead** (voir page 28)

Ces outils peuvent fonctionner comme une brique associée à un autre moteur : Exemple avec le logiciel récemment sorti TopGist qui permet de "thématiser" des recherches effectuées avec Google ou Yahoo.

La cartographie

La plupart des applicatifs disponibles sur le web traitent des relations typées entre données, et ne portent pas sur le contenu textuel des pages (sauf Kartoo).

La visualisation est aujourd'hui en progrès, avec des composantes dynamiques et contextuelles. En général, un ensemble d'information n'est pas représenté par une carte unique qui aurait souvent du mal à rendre compte de la complexité de l'ensemble, par une multitude de cartes ou vues reliées entre elles.

Les technologies de visualisation s'intéressent à des types de données de plus en plus diversifiés, et à des volumes de plus en plus grands. Elles s'interfacent naturellement avec des moteurs de recherche ou d'autres applications d'analyse de données, notamment les moteurs utilisant des technologies de classification automatique (voir ci-dessus)). Si la lisibilité et l'interprétation restent néanmoins souvent délicates, ces techniques sont néanmoins certainement promises à un bel avenir sur le web (et ailleurs) pour aider à la prise de connaissance d'un corpus, et générer de nouvelles hypothèses de travail.

Exemples :

- ✓ Technologie Kartoo appliquée à son méta-moteur : www.kartoo.com
- ✓ Technologie Groxis appliquée à Yahoo : Groker www.grokker.com (les bulles représentent les thèmes, les carrés les sites web, le passage de la souris ou un clic permettant d'avoir les infos utiles)
- ✓ Technologie The Brain appliquée aux résultats du Open Directory : http://www.webbrain.com/html/default_win.html
- ✓ Technologie Anacubis appliquée à Google : http://www.anacubis.com/googledemo/google/index.asp
- ✓ Technologie Mapstan appliquée à Societe.com www.societe.com
- ✓ Exemple de navigation graphique sur Renardus (www.renardus.org) : choisir un grand thème, un sujet, et cliquer sur "graphical navigation".

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 pour le Open Directory.
- ✓ Au niveau francophone, quelque 170000 sites sont répertoriés par Nomade et Yahoo, 75.000 sur les guides de Voila, de Lycos France ou de MSN, et pour environ 100000 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 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, mais elle reste néanmoins fondamentale. Les interfaces à bases

d'onglets mises en place par Google, Yahoo ou Voila par exemple sont nettement plus claires toutefois.

Les répertoires sont aussi utiles :

- ✓ pour se faire une idée du vocabulaire utilisé dans un domaine
- ✓ 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 (New York Times)	www.about.com	
Looksmart	www.looksmart.com	
Nomade (Tiscali)		nomade.tiscali.fr
Open Directory	http://dmoz.org	http://dmoz.fr
Voila (Guide)		recherche.wanadoo.fr ou guide.voila.fr
Yahoo (Guide web)	www.yahoo.com	www.yahoo.fr

Important : De nombreux autres portails intègrent bien entendu ces répertoires, et en particulier le Open Directory

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.

- ✓ 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 Google. 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 éventuelle des éditeurs : exemple 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 : http://www.vlib.org

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://signets.bnf.fr
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	http://www.webdopresse.ch
Presse scientifique	http://www.libs.uga.edu/ejournals/
Bibliothèques	http://sunsite.berkeley.edu/Libweb/index.html (Monde)

	http://sibel.enssib.fr/ (France)
Cartes géographiques Cartographie	www.mapquest.com http://www.sciences-po.fr/cartographie/cartotheque/cartotheques/cartes_diagrammes/milieu.html
Administration française	www.service-public.fr

Exemples de méta-pages thématiques :

Médecine	www.cismef.org	CISMEF – CHU Rouen
Juridique	www.legifrance.gouv.fr www.servicedoc.info/syndication.php3	Legifrance Cons. Constitutionnel
Collectivités	http://www.ait.asso.fr/Liens.htm	AIT
Economie	www.ccip.fr/rime	RIME (grandes écoles commerce)
Informatique	http://www.inria.fr/publications/infoweb/	Inria
Environnement	www.ulb.ac.be/ceese/meta/cdsfr.html	Université Libre de Bruxelles
Sciences sociales	www.sosig.ac.uk	Institute for learning and reserch technology
Culture	www.culture.fr (portails thématiques)	Ministère de la culture

Les répertoires d'outils de recherche

Ce sont des 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.

La page du site de l'ADBS "Panorama des outils de recherche sur le web" (Véronique Mesguich) http://www.adbs.fr/site/repertoires/outils/index.php est intéressante car simple et relativement complète. Voir aussi la page proposée par SearchengineWatch : http://searchenginewatch.com/links/

Répertoires "classiques"

- ✓ 7alpha (www.7alpha.com) ;
- ✓ Beaucoup (www.beaucoup.com)
- ✓ Finderseeker (www.finderseeker.com) ;
- ✓ Metamonster (www.metamonster.com) ;
- ✓ Searchability (www.searchability.com) ;
- ✓ Etc.

A noter que ces répertoires proposent aussi parfois un signalement géographique comme Beaucoup, Ariane6 (www.ariane6.com/moteurs.htm) ou Indicateur (www.indicateur.com), ou sont carrément spécialisés comme Search Engine Collosus (www.searchenginecolossus.com)

Spécialisés sur les outils francophones, tels

- ✓ Enfin (www.enfin.com)
- ✓ Le répertoire (www.lerepertoire.net)

Répertoires de méta-pages "académiques" les plus connus:

- ✓ The Argus Clearinghouse : www.clearinghouse.net
- ✓ Infomine http://infomine.ucr.edu/ (University of California)
- ✓ Virtual Library : www.vlib.org
- ✓ Strathclyde University, Ecosse : Bubl Link : http://bubl.ac.uk/link
- ✓ Internet Public Library (University of Michigan) : www.ipl.org
- ✓ Université de Göttingen : http://www.sub.uni-goettingen.de/0_infint-e.htm
- ✓ Library of California : Librarian's index to the Internet : www.lii.org

A noter : Renardus (www.renardus.org) : tentative pour fédérer plusieurs répertoires européens académiques. Accès en une seule requête à plus de 60.000 notices de ressources utiles provenant de 12 services.

En français, voir notamment les sélections de :

- ✓ Sciences-Po Paris : http://www.sciences-po.fr/docum/sites/index.htm
- ✓ La BNF (les "signets") : http://signets.bnf.fr
- ✓ Agence Science presse (la "bibliothèque") : www.sciencepresse.qc.ca/repertoires.html

Répertoires de "méta-pages" plus typés business :

- ✓ Les 1000 meilleurs portails sectoriels (Objectifs grandes ecoles) : http://www.objectifgrandesecoles.com/pro/secteurs/index.htm
- ✓ L'annuaire pro des sites francophones (Nomade Tiscali) : http://www.nomadepro.tiscali.fr/
- ✓ Guide des portails professionnels (Médiaveille) : http://www.mediaveille.com/outil/portail.htm
- ✓ Guide des sites pour managers (IAE de Paris) www.iae-paris.org/internautie/searchLinks.php4

- ✓ Métaportail Arist Bourgogne : http://bourgogne.arist.tm.fr/metaportail_arist_bourgogne.htm
- ✓ Competita express : http://www.competia.com/express/
- ✓ Search enginge Guide : (http://www.searchengineguide.com/searchengines.html

Les moteurs de recherche

LES MOTEURS DE RECHERCHE : PRINCIPES, IDÉES REÇUES, CHIFFRES CLES

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. ***Certains outils indexent également les méta-données***, mais ce n'est par exemple pas le cas de Google

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.

L'avantage d'un moteur de recherche par rapport à un répertoire, c'est la taille de son index, et sa capacité de réponse à des recherches précises (avec gestion de recherches complexes) en travaillant sur le contenu des pages et non pas seulement des descriptifs. Toutefois, les algorithmes de pertinence développés ne pallient pas les limites d'une indexation souvent "basique en texte intégral", et rappelons que les moteurs ne donnent pas accès au "Web invisible" (voir page 8) A noter : les moteurs évoluent rapidement actuellement dans leurs fonctionnalités, interfaces et modes de traitement de l'information (cf "Le nouveau paysage des outils" page 12)

Attention : les moteurs indexent rarement toutes les pages des sites visités et 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.

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.
- ✓ "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 (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 Les pages disparaissent, évoluent, etc.
- ✓ Les moteurs ont tous l'opérateur ET par défaut et acceptent les " " pour une expression : VRAI

Quelques chiffres sur les moteurs

Nombre de pages indexées par les principaux moteurs
Voir la page de SearchEngineWatch de janvier 2005 :
http://blog.searchenginewatch.com/blog/041111-084221

Voici le tableau extrait de cette page, mais les commentaires sont importants !

Search Engine	Reported Size	Page Depth
Google	8.1 billion	101K
MSN	5.0 billion	150K
Yahoo	4.2 billion (estimate)	500K
Ask Jeeves	2.5 billion	101K+

Notons que l'annonce récente de Yahoo, disant indexer 19,2 milliards de pages web, a fait l'effet d'une bombe, et que certains contestent la réalité de ce chiffre.

- ✓ **Statistiques d'utilisation**

On dispose aujourd'hui de statistiques de fréquentation des différents moteurs assez fiables (la part de marché Google en France approche ainsi les 70 % !), des organismes spécialisés fournissant des statistiques :

En France :

- Médiamétrie-estat : http://www.mediametrie.fr/web/
- Baromètre référencement : http://www.barometre-referencement.com/

A l'international :

- Onestat www.onestat.com

LES PRINCIPAUX MOTEURS FRANÇAIS ET INTERNATIONAUX

Yahoo a racheté en 2003 La société Overture, leader des liens sponsorisés et promotionnels qui venait elle-même de racheter le moteur Altavista et la division web search de Fast, l'éditeur du moteur Alltheweb. De son côté, Yahoo avait finalisé en mars l'acquisition de Inktomi, afin de posséder ses propres technologies de recherche. (En effet, jusqu'ici, le moteur utilisé par Yahoo était Google, un partenaire encombrant qui finalement lui captait de nombreux clients et internautes). **Microsoft a lancé sa propre technologie moteur début 2005.**

Attention : De nombreux sites moteurs ou répertoires travaillent avec des bases de pages crawlées ou des répertoires de sites et des technologies appartenant à d'autres (par exemple, le répertoire utilisé par Google est le Open Directory, Lycos utilise le moteur ASK pour l'international, AOL France a choisi Exalead, etc. Les accords se font et se défont, et il n'est pas toujours facile de suivre et de savoir qui travaille avec qui. Pour mémoire, sont cités dans ce tableau AOL et Lycos en italique, qui n'ont pas en dépit de leur notoriété de technologie moteur propre.

Pour vous aider :

- ✓ Le site Abondance qui propose un tableau bien pratique (attention aux mises à jour toutefois) : http://docs.abondance.com/portails.html
- ✓ Le "search engine decoder" www.search-this.com/search_engine_decoder.asp

Moteurs	Internationaux	Français
AOL	*www.aol.com (résultats Google)*	*www.aol.fr(technologie Exalead)*
Ask Jeeves	www.ask.com	
Exalead		www.exalead.com
Gigablast	www.gigablast.com	
Google	www.google.com	www.google.fr
Lycos	*www.lycos.com (resultats Ask)*	*www.lycos.fr (resultats Yahoo)*
Mirago	www.mirago.com (UK)	www.mirago.fr
MSN	search.msn.com	search.msn.fr
Teoma	www.teoma.com	
Voila		www.voila.fr
Wisenut	www.wisenut.com	
Yahoo	www.yahoo.fr	www.yahoo.com

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 (tri de pertinence)
- ✓ 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 quelques années, 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, racheté par Ask Jeeves en 2001, puis devenu Teoma, était la référence dans ce domaine et fut utilisés 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. Mais Ask Jeeves a ensuite décidé de centraliser ses efforts de développement sur le moteur Teoma.

- ✓ **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.

Les sociétés spécialisées dans le référencement cherchent bien entendu à connaître le plus précisément possibles les critères clés de chaque moteur. 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).

Voir les sites spécialisés dans le référencement comme Webrankinfo (www.webrankinfo.com) qui donne de précieux conseils, notamment pour Google.

A noter : Le modèle du positionnement payant s'est aujourd'hui imposé, et tous les moteurs proposent des "résultats sponsorisés" (c'est aujourd'hui souvent l'indice de clic qui est favorisé) à ne pas confondre avec les résultats "normaux"

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 tels LawCrawler du site Findlaw dans le domaine juridique http://lawcrawler.findlaw.com ou Netsearcher (www.netsearcher.com) sur une sélection de sites internet.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").

Scirus : www.scirus.com : moteur spécialisé et méta-moteur spécialisé : 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.

Vous pouvez consulter le répertoire de moteurs spécialisés disponible à :

www.mylinea.com/moteurs-specialises

Revue de moteurs

(par ordre alphabétique : cette revue ne se veut pas exhaustive ni sur les moteurs bien sûr, ni sur les fonctions,)

ASK JEEVES www.ask.com voir page 47

EXALEAD www.exalead.com

- ✓ La technologie Exalead née en 2001 (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 (elle est "cousine" de l'ancienne fonction "Refine" présente autrefois sur AltaVista. : 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). Exalead propose une solution 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.

- ✓ Lancement récent d'un moteur de recherche qui référencerait déjà un milliard de pages. Il intègre des outils linguistiques tels correcteur d'orthographe, recherche phonétique, recherche approchée, génération de mots-clés
- ✓ Exalead outille AOL.fr, Netscape France, et Réacteur (réseau Abondance)
- ✓ 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 et la Possibilité d'équations complexes : téléphone mobile/portable équivaut à téléphone ET (mobile OU portable). Troncature implicite dès qu'il y a deux mots de la requête, mais si plus de mots, pas de troncature implicite * OK

GOOGLE www.google.fr

Google est la star actuelle des moteurs de recherche (70 % du trafic de la recherche francophone, 50 % au niveau mondial). Il annonce un index de plus de 8 milliards de pages web, sans compter les documents multimedia et les messages de forums dont les ¾ sont réellement disponibles en tant que documents web indexés en texte intégral. Voir à ce sujet le billet de Jean Véronis sur son blog "Technologies du langage" http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html

- ✓ Google utilise des algorithmes analogues à ceux des autres moteurs, mais donne davantage d'importance à 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 pages précédentes sur la pertinence)
- ✓ **Syntaxe**: L'opérateur par défaut est ET. Google accepte les guillemets. On peut utiliser l'opérateur OU, sachant qu'une équation booléenne est acceptée (pas besoin de parenthèses, le moteur traite d'abord le 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).

  Recherche dans le titre des pages (fonction intitle: ainsi que allintitle:) et dans l'url (inurl: ainsi que allinurl:) Utiliser allinurl pour signifier au moteur que l'on veut plus d'un mot dans l'URL ou dans le titre avec 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 management retrouve les fichiers de type pdf contenant le mot-clé management (il est obligatoire d'utiliser un mot-clé)

  Fonctionne avec les fichier texte (txt), word (doc), excel (xls), powerpoint (ppt), Adobe (pdf et ps), Works (wks, wps et wdb), RTF (rtf), ASP (asp), flash (swf), Lotus. Indexation récente du flash (texte présent dans les animations flash). Aujourd'hui, 13 principaux formats de fichiers indexés et recherchés par Google en plus du html
- ✓ 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.
- ✓ Le moteur utilise le répertoire Open Directory, mais reclasse dans chaque rubrique par popularité via son système.
- ✓ 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 a lancé en 2002 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

A noter : un répertoire dédié à Google : http://google.indicateur.com/ (ou http://google.indicateur.fr en français)

MIRAGO www.mirago.fr

- ✓ Ce moteur, à la technologie propriétaire, s'intéresse aux pages françaises (également Royaume Uni, Allemagne et Espagne)
- ✓ 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

Mozbot www.mozbot.fr

- ✓ Ce moteur de recherche est créé par Abondance, Raynette.com et Brioude Internet en utilisant les résultats de Google. C'est une amélioration du moteur Reacteur qui intègre de nouveaux services et fonctionnalités telles
    - Gestion de liste d'exclusion pour supprimer définitivement des résultats certaines pages et personnalisation de l'interface
    - Définitions du dictionnaire sur les mots demandés et suggestions de recherches connexes
    - Historique des requêtes effectuées sur le moteur
    - Mise en favori d'un lien
    - Informations sur le propriétaire d'un site renvoyé comme résultat
- Barre d'outils pour les navigateurs Explorer et Firefox qui permet de faire des recherches sur Mozbot, mais aussi Google, Yahoo et MSN (dont images et actus), surligner dans la page les mots demandés, et de bloquer les pop-ups.

MSN

- ✓ 5 milliards de pages rafraîchies tous les deux jours (avec 'intégralité des articles de l'encyclopédie Encarta appartenant à Microsoft).
- ✓ Ne tient pas compte de la casse des lettres, ni des caractères accentués.
- ✓ Fonctions de recherche classiques : Et par défaut, OU traduit par OR, site:, link:…
- ✓ Nouveaux opérateurs récents : filetype: ; linkdomain: (toutes les pages qui ont mis en place un lien vers n'importe quelle page d'un site) ; contains: (recherche sur des documents qui contiennent le mot demandé dans leur code html) ; inurl: (dans l'adresse de la page) ; inanchor (dans le contenu textuel des liens) ; intitle: ; inbody: (dans le texte à l'exception des titres) ;

TEOMA

- ✓ Ce moteur de la société Hawk Holdings , issu d'un projet né en 1998 à Rutgers University aux Etats-Unis, a été racheté par Ask Jeeves très rapidement après sa sortie en 2001. Il fournit une alternative aux résultats fournis par le système de questions-réponses Ask Jeeves. Il annonce actuellement un index de 2 milliards de documents proposés (200 millions en avril 2002), avec certains sites visités tous les jours, les formats pdf et flash indexés, une version "cache", des "related searches".
- ✓ 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. Liens établis à partir des communautés identifiées automatiquement.
- ✓ 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.

- ✓ Recherche avancée disponible, indexation des fichiers pdf récente.

YAHOO www.yahoo.fr

- ✓ Moteur lancé en février 2004, rompant ainsi son partenariat avec Google. Les moteurs Alltheweb et AltaVista ont aujourd'hui abandonné leur index pour adopter celui de Yahoo et sa technologie, même si les résultats ne sont pas toujours identiques. Le moteur annonce un index de plus de 4 milliards de pages, donc comparable à Google, et son interface générale est également proche. Yahoo a conservé le système de présenter en haut de la page de résultats les rubriques du guide web qui contiennent les mots de la requête. Ouverture directe possible des résultats dans une nouvelle fenêtre, fonction cache sont au RDV
- ✓ Nombre de résultats par page par défaut : 20
- ✓ Voir les fonctionnalités de recherche très complètes (le OR réclame des capitales) et recherche avancée. La fonction 'site:' est la même que sur Google (recherche sur un domaine, ou un sous-domaine) ; 'url:' permet de chercher un document en particulier, quand 'inurl:' permet de chercher une page ayant un mot-clé en particulier dans l'URL ; 'title:' cherche des pages ayant un mot-clé en particulier dans le titre.
- ✓ 'Fonction link:' comme sur Google, pour chercher les "backlinks" d'une page, mais l'URL complète doit être indiquée (avec le http) et contrairement à Google, on peut utiliser un mot-clé en plus (trouver les pages contenant un mot-clé et "pointant" sur). A noter également la fonction 'linkdomain:' qui permet de trouver les pages qui pointent sur le domaine donné (ex cnrs.fr) ou sur l'ensemble des pages d'un site donné (ex www.cnrs.fr). Il ne faut alors pas mettre de http. On peut ainsi trouver les pages émanant de sites publics français "pointant" sur l'un des sous-sites CNRS avec la requête : *linkdomain:cnrs.fr site:gouv.fr.* Ces fonctions peuvent être utilisées avec d'autres opérateurs, comme inurl ou intitle.
- ✓ Pas de limitation (comme le fait Google) sur le nombre de mots dans la requête.
- ✓ Si la page de résultats contient un "fil RSS", celui-ci est signalé (avec possibilité de le rajouter à My Yahoo).
- ✓ Toute une série de raccourcis de recherché sont disponibles sur Yahoo US à l'adresse http://help.yahoo.com/help/us/ysearch/tips/tips-01.html (des numéros d'immatriculation aux brevets en passant par la définition d'un mot).
- ✓ Nouvelles fonctions de personnalisation (voir page 14)
- ✓ CNN a remplacé Google par Yahoo en mai dernier.
- ✓ A noter cette fonction utile disponible sur Research Buzz pour chercher un mot séparé d'un (ou deux, trois, quatre ou cinq) mot du premier. http://www.researchbuzz.org/archives/002051.shtml
- ✓ Yahoo a sorti en février 2005 Y!Q (voir page 52).
- ✓ Yahoo!Mindset (beta) (http://mindset.research.yahoo.com) permet de moduler les résultats en fonction de la finalité de recherche de l'utilisateur (plus orienté "shopping" ou "recherche d'information") grâce à un simple curseur à faire varier.

WISENUT www.wisenut.com

- ✓ 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 en 2003 par Looksmart
- ✓ 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. On peut ouvrir la catégorie ou relancer une nouvelle recherche en utilisant la catégorie comme requête.
- ✓ Le moteur groupe les résultats par site et liste le nombre exact de pages d'un site définies comme pertinentes.

Les méta-moteurs sur le web

Présentation

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

Kartoo 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. Les sites sont représentés par des pages plus ou moins grosses, selon leur pertinence. C'est la Version 3 qui est actuellement en cours.

Kartoo utilise la syntaxe suivante : - OR : cherche les sites contenant au moins l'un des mots ou expressions indiqués- URL : pour chercher des pages dont l'adresse contient un mot donné- LIKE : pour chercher des sites similaires- HOST : pour chercher sur un site donné- TITLE : pour chercher des pages dont le titre contient un mot donné- DOMAIN : pour chercher des pages dont le domaine est donné- TEXT : pour chercher dans le texte de la page en priorité- LINK : pour chercher un mot qui se trouve dans un lien hypertexte- IMAGE : pour chercher des images- NEAR est utilisé pour chercher 2 mots proches l'un de l'autre sur une page

Meceoo **www.meceoo.fr ou www.meceoo.com**

Le métamoteur Meceoo est proposé par le Réseau Abondance, effectue ses recherches sur plusieurs moteurs majeurs des Web francophone et anglophone (choix automatique des moteurs en fonction de la requête, de la langue, etc.).

Meceoo offre un vrai + en permettant aux utilisateurs de créer leur propre "liste d'exclusion" (afin d'éliminer de ses pages de résultats certains sites estimés peu pertinents) et aussi de lancer une recherche uniquement sur le contenu de sites web sélectionnés (liste sauvegardée).

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 développa le méta-moteur Bulls-Eye aujourd'hui arrêté, Profusion permet aujourd'hui 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 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").

MyWebSearch **http://www.mywebsearch.com**

Dixit C. Asselin (mars 2005) "Dans le classement Hitwise des moteurs de recherche US (nombre de visites), l'outil MyWebSearh apparaît en 7e position; Ce multimoteur de Ask Jeeves, peu connu en France permet de visualiser par onglets les résultats de Google, Yahoo, Ask et Looksmart."

Vivisimo www.vivisimo.com

Ce méta-moteur, issu de Carnegie Mellon University a pour particularité de classer les résultats des recherches dans des dossiers (catégorisation automatique, ou clusterisation), comme le faisait 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. L'outil Traduit la requête AND + OR NOT NEAR dans le langage des moteurs ; La pertinence n'est pas recalculée mais est fonction des algorithmes des moteurs utilisés.

Voir aussi la technologie appliquée à PubMed (Medline) : http://vivisimo.com/clustermed

Vivisimo a récemment lancé Clusty (http://clusty.com/), qui propose de chercher dans les images, les actualités, les encyclopédies.

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 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 Copernic (voir plus loin)

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
Médecine	Omnimedicalse	http://www.omnimedicalsearch.com/
Sciences	Scirus	www.scirus.com

Comment trouver… ?

COMMENT TROUVER DES LISTES DE DISCUSSION ET DES FORUMS ?

Cf page 9

Listes de discussion

Au niveau francophone, Francopholistes (www.francopholistes.com) reste le répertoire incontournable avec plus de 6200 listes indexées. La société propose aussi une recherche centralisée sur les archives récentes de l'ensemble des listes francophones (recherche parmi 142000 messages)

Citons aussi Pidinfo (www.pidinfo.com) lancé par Histen Riller en 2003 qui signale 300 lettres d'information professionnelles.

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

- ✓ La société Topica 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 : http://lists.topica.com/
- ✓ 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.
- ✓ Google s'est mis sur le créneau (voir ci-dessous).
- ✓ http://www.lsoft.com/lists/listref.html

Forums de discussion

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 845 millions de messages depuis 1981. Les fonctionnalités de recherche sont assez complètes : par newsgroup, par sujet, par auteur, par langue et par date. http://groups.google.com

Une nouvelle version du site s'est mis en place qui laisse plus de place à la personnalisation et gére également les listes de discussion.

Citons aussi, pour la France, le serveur mis en place par Voila http://news.voila.fr

Pour les webforums, peu d'outils disponibles. Voir en anglais www.boardreader.com fondé en mai 2000 par des ingénieurs et étudiants de l'Université du Michigan.

Voir aussi le récent Lycos Discussion Search (http://discussion.lycos.com/)

Comment trouver des sites fédérateurs ou portails ?

Cf page 7. 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 des répertoires professionnels vus auparavant. Attention, ils ne sont jamais exhaustifs, et peu critiques pour la plupart.
- ✓ 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 en utilisant des mots-clés comme portail, liens etc. associés à la thématique souhaitée.
- ✓ Le "bouche à oreille" (y compris sous forme électronique avec les listes de discussion et forums) : l'information sur les bons sites circule…

Comment identifier des ressources du web invisible ?

Cf page 8

- ✓ L'identification 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://dadi.enssib.fr (près de 900 bdd gratuites avec classification Dewey)

  - en anglais, The invisible web directory de Chris Sherman et Gary Price http://www.invisible-web.net (près de 1000 bdd majoritairement gratuites)

  - en anglais, le répertoire de la société Bright Planet www.completeplanet.com (plus de 103000 outils organisés selon un système de 4000 catégories !) : des bases de données, mais aussi des moteurs, dont des moteurs internes à des sites. Notons que tant l'identification des outils que la classification sont entièrement automatiques via la recherche de formulaires et l'application d'un algorithme propriétaire. Options de recherches sophistiquées néanmoins…
- ✓ Pour les news et les sites à très fort renouvellement (dont weblogs), utiliser des agrégateurs ou outils ad hoc.
- ✓ Utiliser des méta-moteurs spécialisés, si toutefois il en existe dans son domaine.

COMMENT TROUVER DES WEBLOGS ET "FILS RSS" ?

Cf page 7 – mise à jour janvier 2005

Outils internationaux (mais majoritairement anglophones)

1/ Répertoires et listes internationales

Blogarama www.blogarama.com
20100 blogs listés (7850 en février 2004) avec un "vrai" système de catégories (avec arborescence). Recherche sur les descriptifs (avec le AND, le OR, et les guillemets disponibles) et recherche avancée disponible.

Eaton web http://portal.eatonweb.com/
20800 blogs (16400 en février) environ répertoriés sur ce site . Les catégories sont en fait des mots-clés prépropposés. On peut aussi avoir une liste par ordre alphabétique, par pays et par langue + moteur mots-clés sur les descriptifs.

Newsisfree http://www.newsisfree.com
14000 sources (8200 en février). Répertoire de sources par catégories : http://www.newsisfree.com/sources/bycat/ : on peut aussi chercher sur les noms et descriptions des blogs.
On peut aussi choisir d'avoir les derniers articles par catégories. Egalement moteur sur les derniers "posts" récupérés (voir ci-dessous dans "moteurs internationaux").

Syndic8.com http://www.syndic8.com
Liste par ordre alphabétique (environ 80000 sources, 28000 en février 2004), par catégories (pas très utilisable), et recherche par mots-clés.
Liste pour la France (non complète, néanmoins) :
http://www.syndic8.com/feedlist.php?ShowLanguage=fr&ShowStatus=all
Des fils RSS sont ajoutés, voire créés sur proposition des internautes. Dans la liste, un fil orange émane directement du producteur. Un fil bleu a été "forgé" ("scraped" par Newsisfree) (exemple ci-dessous).

Weblogs.com http://www.weblogs.com/
Blogs récemment (dans les trois heures) mis à jour : on peut s'abonner au fil RSS correspondant.

2RSS.com http://www.2rss.com/
Un nouveau repertoire qui traite actuellement quelque 7800 fils.

2/ Moteurs internationaux

Blogpulse www.blogpulse.com

Mis en ligne en mai 2004 par la société Intelliseek. Moteur de blog (avec fonctionnalités de recherche avancées) et permettant également de voir les termes les plus discutés, qui sont classés (graphiques). On peut générer des graphiques en utilisant le "Trend search" Blogpulse propose également un classement des Top links, Key phrases et Key people, avec pour ce dernier un baromètre des positions.

Bloogz – Worl Wide blogs http://www.bloogz.com/

Originellement italien, le moteur Bloogz permet aujourd'hui une recherche également en français, anglais, espagnol, italien et allemand. Le nombre de blogs pris en compte n'est pas indiqué, mais les éditeurs peuvent "référencer" leur blog. On peut choisir un tri par pertinence ou par date.
Nouveauté : "Blogs popularity index" (en fonction du nombre de visiteurs et de citations) et Agrégateur Rss disponible gratuitement en ligne

Daypop http://www.daypop.com

Capacité de recherche dans quelque 60.000 sources (même chiffre annoncé en février 2004) utilisant le format RSS (35000 en octobre 2003) : sites de news, weblogs et "fils" RSS". Recherche avancée permet de chercher par pays et par langue. On a le "post" en cache et les citations sur le blog dans la liste des résultats.

A note la fonction link: pour savoir qui pointe sur un blog (par rapport à l'index de Daypop)

Blogstats pour avoir les statistiques de popularité d'un blog

Feedster http://www.feedster.com

Moteur spécialisé dans la recherche sur les weblogs et fils rss (800.000 sources et 5000 nouvelles par jour ! pour 446.000 en février), nés de la fusion de Feedster avec RSS Search : fonctionnalités évoluées de recherche. Basiquement, on choisit des résultats triés par pertinence ou par date.

On peut créer un fil RSS à partir d'une recherche par mots-clés

Possibilité d'avoir une visualisation graphique des provenances des résultats (carte du monde).

"My Feedster" permet de syndiquer son contenu en ligne gratuitement, en sauvegardant des posts et des recherches.

Nouveau : possibilité de savoir qui pointe sur un blog (par rapport à l'index de Feedster, bien sûr).

Newsisfree http://www.newsisfree.com
14000 sources. Recherche par mots-clés des articles récupérés dans les deux derniers jours. Formulaire permettant de choisir entre le Et et le Ou entre les mots, le tri de pertinence et la langue. Une recherche par catégories est également disponible (cf ci-dessus).

Technorati www.technorati.com

Plus "moteur de popularité" que moteur de recherche, Technorati est l'un des premiers outils à avoir scruté la "blogosphère" (actuellement environ 4300000 weblogs surveillés et 650000000 posts) selon cet axe. Le moteur d'avoir une liste de blogs qui ont fait un lien vers une source donnée (site web, blog ou article). Inutile de mettre http:// ou même www, l'outil les rajoute automatiquement.

Waypath www.waypath.com

Ce nouveau moteur (oct 2004) permet de faire une recherche sur le contenu de trois millions de blogs, avec des fonctions avancées (opérateurs, proximité, etc.)

L'outil propose par ailleurs l'agrégation de différents blogs traitant du même sujet.

Une fonctionnalité à retenir : le bouton "Waypath it!" en bas à droite de chaque résultat vous permet trouver les posts ayant un sujet proche (qui donne d'ailleurs lieu à un bookmaklet pour faire de même en cours de navigation.

Outils francophones

1/ Répertoires francophones

Blogonautes www.blogonautes.com

Répertoire de blogs francophones (Annublog, cité dans l'article, renvoie désormais sur Blogonautes) : il annonce actuellement 3922 weblogs (2470 en février) et offre une recherche multicritères : nom de l'auteur du blog, pays et ville d'origine, sexe et tranche d'âge (attention, selon les déclarations du blogueur lui-même), mots-clés (recherche sur la description du blog). On peut trier les résultats par différents critères.

Pas de classement hiérarchique des weblogs.

La page d'accueil permet de voir les nouveautés sur les blogs répertoriés (fil RSS d'ailleurs).

BlogArea http://www.blogarea.net/
Encore un annuaire avec 379 blogs référencés, et 20 catégories pour 43 sous-catégories.

Blogolist http://blogolist.com

Ce répertoire francophone indexe environ 740 blogs (600 en février) en permettant une recherche par origine géographique. Recherche dans les url, les titres, les descriptions, les mots-clés (une liste de mots-clés est proposée). Pas de classement hiérarchique des blogs. Blogs récemment mis à jour, articles et blogs les plus cités

Les pages joueb http://pages.joueb.com
636 blogs (id en février, c'est inquiétant !) classés par mots-clés.

Retronimo www.retronimo.com

Annuaire de fils RSS (650 actuellement + 83 flux "exclusifs" générés pour des sites ne proposant pas de fils RSS + quelques flux thématiques associés à un moteur de recherche sur le "fond documentaire" ainsi constitué), de blogs (URL directe www.retronimo.com/blog), et métamoteur de blogs (URL directe www.retronimo.com/bse)

RssReporter http://www.rssreporter.net/html/
Un petit nouvel annuaire de fils RSS ou Atom (actuellement 246 sites recensés) qui comporte une vraie arborescence (certes encore un peu pauvre !)

Weblogues.com http://www.weblogues.com/

Ex-Moteur spécialisé sur les blogs francophones avec recherche intégrale sur le texte des "billets", Weblogues a restreint son ambition en devenant un outil de repérage des blogs (quelque 35280 weblogs, 4500 en février, 2500 en octobre 2003)..

On peut rechercher sur la description des blogs, ou par une liste de mots-clés assez anarchique. La page s'ouvre sur les blogs mis à jour, ce qui constitue d'ailleurs un fil RSS auquel on peut bien sûr s'abonner.

2/ Moteurs francophones

Easy RSS : moteur de recherche de fils RSS (actuellement 130400 fils indexés environ) lancé par deux bruxellois avec l'objectif ambitieux de devenir "le premier portail RSS européen".

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.

- ✓ On peut télécharger la barre d'outils Alexa ou utiliser le moteur www.alexa.com qui rajoute les infos sur la page aux résultats de Google avec une interface très agréable et donne des sites "similaires" à la page visitée

- ✓ Il convient bien sur de ne pas oublier pour autant la technique "classique" en recherche d'information : extraire la terminologie du site pertinent, voire les noms d'expert pour relancer une recherche sur des mots-clés a priori performants.

Où trouver des archives du web ?

Rien n'est exhaustif dans le monde du web, mais le service proposé par l'association The Internet Archive (qui reçoit des donations et soutiens de différents acteurs, dont Alexa) est très impressionnant : on peut ainsi visualiser un site tel qu'il était à différentes dates depuis 1996, et même suivre des liens sur ces archives.

The way back machine : www.archive.org

Depuis la fin 2003, un service en beta permettait d'aller beaucoup plus loin, en permettant une recherche plein texte, par date, sur plus de 11 milliards de pages archivées. Différentes fonctionnalités étaient accessibles à partir des résultats des sites répondant le mieux à la recherche : graphique permettant de voir la fréquence d'apparition du mot-clé sur la période, thèmes traités par le site, concepts proches, etc. Malheureusement, le site ne répond plus depuis plusieurs mois, mais devrait prochainement revenir sur le devant de la scène : http://recall.archive.org

A noter : La Bibliothèque Naitonale de France va archiver le web français, dans le cadre d'un dépôt légal des sites. Lire l'article sur 01net.com de Yannick Arrieux de juin 2004 http://www.01net.com/article/246302.html , et sur le site de la BNF bien sûr http://www.bnf.fr/pages/infopro/depotleg/dli_intro.htm

Comment trouver des bookmarklets ?

Les bookmarklets sont des programmes contenus dans des liens, c'est à dire des éléments de code java qui se mettent dans les favoris comme des URL classiques, mais qui déclenchent quand on les appelle une action particulières. Ils déclenchent souvent ouverture de fenêtre pop-up (ce qui pose d'ailleurs un problème quand on utilise un "anti pop-up" : obtenir le premier résultat du moteur Google directement, faire un lien direct vers un paragraphe de page html, traduire, éditer les urls présents sur une page à la fin de celle-ci, intégrer un nouveau bookmark si l'on est sur un service en ligne de gestion de favoris, etc.

Pour en trouver, et pour démarrer votre recherche :

- http://outilsfroids.joueb.com/texts/OutilsBookmarklets.shtml
- www.bookmarklets.com

Comment trouver des fichiers audio, des videos ?

Les grands moteurs commencent à proposer ce type de services, avec plus ou moins de bonheur, car la pertinence de l'indexation est encore discutable (ne prend pas en compte les images, notamment). : Voir par exemple Yahoo (http://video.search.yahoo.com/) ou Google (http://video.google.com) . Ces services sont basés sur l'indexation de retranscriptions textuelles des vidéos, ou des pages web qui les accompagnent. Yahoo a lancé récemment un moteur de recherche audio (http://audio.search.yahoo.com) avec recherche possible par artiste, chanson, album, par format, durée, source… Quelque 50 millions de fichiers audio sont concernés, dont pas mal d'artistes français. Mais le pb des droits d'auteurs ne semble pas résolu, de nombreux morceaux étant téléchargeables gratuitement.

On peut aussi utiliser le moteur Blinkx (http://www.blinkx.tv/) qui permet de chercher dans les fichiers video issus de la télévision : le moteur cherche directement sur le web les fichiers videos qui sont lus et, via une technologie de reconnaissance de la parole, indexés sous forme de texte.

Speechbot est un projet de recherche de Hewlet Packard qui permet de chercher dans le contenu de plus de 15000 heures de programmes radios, indexés comme Blinkx après reconnaissance vocale.

Notons que le moteur de Google vient de s'enrichir d'un lecteur video permettant de lire directement les fichiers dans les pages de résultats (sur PC / windows et navigateurs Explorer et Firefox). Le moteur propose maintenant non seulement des programmes télé, mais aussi des documents fournis par des internautes.

Comment..? Est-il possible de… ?

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 ou les archives de Alexa.

- ✓ **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.

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 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. Sinon, travailler du plus général au plus spécifique (mais choisir les synonymes appropriés pour le terme générique : par exemple si je m'intéresse à un film qui s'appellerait "Demain, dès l'aube", on pourrait écrire 'film OR cinéma "demain, dès l'aube"').

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. C'est pourquoi il est parfois plus judicieux de repérer une notion discriminante de son sujet de recherche plutôt que de d'utiliser sans réflexion le SAUF.

Majuscules, minuscules, accents ? De façon générale, les moteurs sont insensibles à la casse des caractères et retourneront le même nombre de résultats pour python, PYTHON, ou Python. La situation est plus contrastée en ce qui concerne les accents : si MSN, Exalead, Voila et en théorie Google traitent de manière identique les mots accentués ou non (l'expérience montre que c'est loin d'être toujours évident sur Google), Yahoo par exemple procède différemment : ils ne retournent pour un mot-clé accentué que les mots contenant l'accent, mais pour une requête non accentuée, ils retournent les mots avec ou sans accent. Bref, il convient de faire attention à l'utilisation des accents avant d'utiliser un moteur.

Troncatures ? La troncature permet de remplacer plusieurs caractères sur la fin des mots, mais cette possibilité devient fort rare sur le web : les trois grands moteurs Google, Yahoo et MSN ne proposent pas cette option, et notamment, un mot-clé indiqué au singulier sera traité comme tel. Il est donc important de prévoir au moins l'alternative du mot au pluriel, sous peine d'occulter de nombreux résultats. pertinents. En revanche, le challenger Exalead supporte la troncature avec le caractère * Sur le Guide du web de Voila, une recherche au singulier ou au pluriel donne en revanche les mêmes résultats (faux sur le moteur Voila).

Ordre des mots ? Il peut avoir de l'importance selon les moteurs, non pas bien entendu pour le nombre de réponses, mais pour le classement des résultats : c'est par exemple vrai pour Google ou pour Voilà.

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 pour le français et l'anglais http://dico.isc.cnrs.fr/
- ✓ Utiliser un thésaurus de son domaine (en ligne gratuit, ou acheté comme par exemple celui de la base INSPEC (www.iee.org.uk). Glossarist (www.glossarist.com) est un répertoire de glossaires en anglais
- ✓ Utiliser un moteur de recherche travaillant à partir de dictionnaires, encyclopédies, thesaurus, tel pour les termes en anglais FreeDictionary http://www.thefreedictionary.com/ . Voir aussi www.thesaurus.com.
- ✓ Faire une recherche sur une base de données bibliographique du domaine dans lequel se situe le sujet, utilisant une indexation manuelle (dewey, autre plus spécialisée avec un thésaurus par exemple). Repérer alors comment sont indexés quelques documents pertinents, quelle est la terminologie retenue.
- ✓ Utiliser l'option define: sur Google (aujourd'hui disponible en français). On a aussi depuis peu l'option Google Suggest http://www.google.com/webhp?complete=1&hl=en

Utiliser les générateurs de mots clés des grands moteurs publicitaires. Attention, on travaille alors à partir des requêtes des utilisateurs beaucoup plus qu'à partir de la terminologie des documents traitant du sujet (même si parfois on a aussi l'indication des mots-clés accompagnant souvent le mot demandé dans les pages web). Ces outils servent en général à mieux référencer un site web. Ils peuvent néanmoins donner des idées (pour compléter cette liste, voir la rubrique consacrée au sujet par le site Abondance : http://ressources.abondance.com/generateur-mot-cle.html

- Générateur de mots-clés du programme Google Adwords : https://adwords.google.com/select/main?cmd=KeywordSandbox
- Overture : http://inventory.overture.com (ou pour la France http://inventory.fr.overture.com ; pour les autres pays, on remplacera "fr" dans l'adresse par le code correspondant au pays souhaité).
- Espotting : http://fr.espotting.com/popups/keywordgenbox.asp
- Outiref (www.outiref.com). Voir aussi WebRankinfo http://www.webrankinfo.com/outils/semantique.php

✓ 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 : Notons que Surfwax a mis en ligne (mars 2005) WikiWax, qui combine l'outil de suggestion de terme de Surfwax et l'encyclopédie collaborative en ligne Wikipedia. http://www.wikiwax.com/. Enfin, Surfwax propose aussi sa technologie "LookAhead" en action sur Wikiwax sur son "News Accumulator" très intéressant : des articles récents en provenance de 4000 sources que l'on obtient un tapant un mot (50000 sujets possibles) : on choisit alors l'entrée souhaitée pour avoir les articles correspondants.

✓ Explorer les balises méta (keywords) de quelques documents pertinents

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". On peut aussi faire une recherche moteur avec un mot-clé large en français et en anglais et un mot-clé "profond" en français seulement : wine vin soutirage peut me donner des infos terminologiques sur le soutirage en anglais.

COMMENT ÉVALUER UN SITE 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, sachant qu'il convient de croiser une évaluation de la source avec une évaluation du contenu :

✓ 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 http://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*

Il est intéressant de consulter le cours en ligne "L'évaluation de l'information sur Internet" à l'adresse www.uhb.fr/urfist/Supports/StageEvalInfo/EvalInfo_cadre.htm , élaboré par Alexandre Serres, responsable URFIST Bretagne

Astuces pour l'évaluation des pages en cours de navigation

- ✓ *Chercher des informations sur l'éditeur sur le site*. En cas de difficulté, chercher le copyright en bas de page. On peut aussi repérer sur le plan du site la page Contact qui va fournir un email. Voir alors la seconde partie de l'adresse mail (après le @) qui peut renvoyer à un domaine particulier que l'on cherchera alors sur le web.
- ✓ *Chercher des informations sur la société indiquée*. On utilisera alors des bases de données d'informations sur les sociétés (R5CS, organismes de régulation boursiers).
- ✓ *Pour rechercher le propriétaire d'un nom de domaine* (noms des responsables techniques et administratifs). Attention, les informations sont loin d'être toujours mises à jour, donc il y a des risques d'erreur, et parfois besoin de recoupements.
    - Pour les noms de domaine se terminant par un ".fr" on utilisera le moteur proposé par l'AFNIC, centre d'information et de gestion des noms de domaine pour la France (et pour l'île de la Réunion .re) : www.afnic.fr
    - Pour les noms de domaine "gTLD" (generic Top Level domains), c'est à dire les .com, .net, .org, et plus récemment les .biz et les .info, c'est plus difficile car les bases de données ne sont plus unifiées (auparavant, la base Whois gérée par l'Internic). On utilisera donc un méta-moteur comme Betterwhois, qui permet d'interroger les bases des "régistrants" (prestataires assurant la gestion administrative et technique du nom de domaine) les plus importants : www.betterwhois.com. Voir aussi Allwhois http://www.allwhois.com/
    - Pour les autres noms de domaine par pays, on peut passer par un service générique http://www.generic-nic.net/dyn/whois, ou bien chercher préalablement l'organisme national pays par pays sur Yahoo : http://dir.yahoo.com/computers_and_internet/internet/domain_name_registration/top_level_domains__tlds_/registry_operators/International_Country_Codes/
- ✓ *Pour trouver des informations générales sur la page*, on peut utiliser le moteur Alexa www.alexa.com, 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 proposés.

- ✓ *Utiliser également le "URL info" de Fagan Finder :* http://www.faganfinder.com/urlinfo
- ✓ *Ne pas oublier non plus de faire des recherches sur le web* en prenant le nom du site comme mot-clé, et avec la fonction link : (recherche par popularité : qui a un lien sur cette page).
- ✓ *On peut aussi utiliser l'interface de recherche développé par un journaliste Jean-Marc Manack pour se simplifier la vie dans la validation des informations :* Plus de 200 outils classés par rubriques (moteurs de recherche, administratif – URL, dictionnaires, référence, actualités, blogs, etc.) sont disponibles à partir d'un seul formulaire, les résultats apparaissant dans la partie gauche de la page. La différence avec un méta-moteur classique, est que l'on peut mettre soit un mot-clé, soit une url. (utilisable aussi en mode "sidebar" dans le navigateur : http://manhack.net

PEUT-ON FAIRE UNE RECHERCHE PAR DATE ?

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 la date de mise à jour du document, ce qui ne garantit pas forcément la fraîcheur des informations. Si dans le cas d'articles issus de revues "reconnues", la date est clairement indiquée, il n'en est pas de même pour des pages web "classiques". On peut alors tenter, une fois sur la page souhaitée, de taper dans la barre d'adresse du navigateur, à la place de l'URL, la commande : 'javascript:alert(document.lastModified)', qui permet dans la moitié des cas environ d'obtenir la date de dernière modification de la page (ne fonctionne pas avec la plupart des pages dynamiques).

Voir cet article de First Monday

http://www.firstmonday.org/issues/issue9_10/wouters/index.html#w2

et cet article de Armelle Thomas dans Veille Magazine

http://www.inforizon.com/b/s/242/fiches/FILE1_3548414169106327424171O.pdf

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 :

- ✓ 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 2003)

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

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

PEUT-ON COMPARER LES RÉSULTATS DES MOTEURS DE RECHERCHE ?

Depuis l'arrivée sur le marché des moteurs des deux poids lourds Yahoo et MSN, l'internaute dispose d'une vraie alternative à Google, dans la mesure où les résultats semblent fort différents d'un moteur à l'autre. L'étude conduite en avril et juillet 2005 par le méta-moteur Dogpile, en collaboration avec des chercheurs de l'Université de Pittsburgh et de Pennsylvannie, montre que le taux de recouvrement est très faible sur les premiers résultats (http://CompareSearchEngines.dogpile.com/OverlapAnalysis) : seuls 1,1 % des liens proposés seraient communs aux 4 moteurs testés (Ask Jeeves faisait partie de l'étude), 89,4 % étant uniques à un seul moteur, et 11,4 % étant proposés par deux moteurs.

Oui, on peut donc comparer les résultats (souvent plus facilement sur des requêtes en anglais, les utilitaires étant anglo-saxons) et voici les outils à disposition :

- Pour avoir les résultats des deux moteurs Yahoo et Google côte à côte : **Google-Yahoo comparison** http://www.googleguy.de/google-yahoo/
- Pour savoir quels sont les résultats communs de Yahoo, Google et Ask Jeeves et uniques à une requête, (par exemple : quels sont les résultats de Yahoo qui ne sont pas dans Google ?) : **Jux2** www.jux2.com
- Pour avoir une visualisation graphique comparant les 100 premier résultats de deux moteurs au choix parmi Google, Yahoo, Alltheweb, **AltaVista, MSN,** Teoma et Wisenut : **Thumbhots** http://ranking.thumbshots.com
- Voir aussi l'outil développé en flash par le méta-moteur Dogpile (www.dogpile.com). Si le meta-moteur permet de voir grâce à des fenêtre flottantes les 10 premiers résultats de chaque moteur souhaité, une autre interface permet elle de voir les zones de chevauchement et de spécificité entre Google, Yahoo et AskJeeves (sur les tous premiers résultats) : http://missingpieces.dogpile.com/missingpiecestool.aspx Cet outil est toutefois moins pratique que Jux2.

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, les sociétés telles Albert (www.albert-inc.com) ou travaillant le plus souvent pour des projets intranet / internet de clients spécifiques.

Voir par exemple 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) quatrième moteur de recherche américain après Google, Yahoo et MSN, peut être utilisé avec le langage naturel., même s'il se rapproche aujourd'hui beaucoup plus d'un moteur classique qu'à sa création (où il fonctionnait à partir d'une base de données ayant atteint dix millions de questions-réponses, donnant la page web la plus appropriée à sa question), donnant les liens de son moteur TEOMA. Ask Jeeves a été rachetée par la société AC/InterActiveCorp dirigée par le milliardaire Barry Diller (Rappelons que Ask Jeeves est propriétaire de la marque Excite.com et de Teoma, ainsi que de Blogger.
- ✓ Notons qu'après la tentative Infoclic qui a cessé son activité fin 2001, pas d'autre clone français se montre pour le moment.
- ✓ A tester également Brainboost (www.brainboost.com), the "answer engine".
- ✓ Yahoo a sorti en février 2005 Y!Q (http://yq.search.yahoo.com) qui permet de rentrer dans la fenêtre de requête, non seulement des mots ou expressions, mais aussi des textes entiers. Ceci permet de trouver des sites associés à la sélection de texte donnée. Le moteur filtre en fait la question pour éliminer les termes "inutiles" (de son "point de vue"), puis découpe la requête en plusieurs termes ou expressions, que l'on peut par la suite sélectionner ou désélectionner.

PEUT-ON CIRCULER DE FAÇON ANONYME SUR LE WEB ?

On le sait, la navigation sur le web laisse des traces (voir notamment à ce sujet le site de la CNIL www.cnil.fr). Il existe néanmoins des services permettant de masquer les adresse IP d'origine et d'empêcher les cookies et autres techniques de marquage de fonctionner, c'est à dire de garantir une meilleure confidentialité de surf sur internet

Anonymiser http://www.anonymizer.com/ (payant)

Technologie Safeweb, acquise par Symantec

Voir aussi la catégorie adéquate du Open Directory : http://directory.google.com/Top/Computers/Security/Internet/Privacy/

Voir enfin le portail Stay Invisible qui propose définitions, actualités, tests, un forum de discussion sur le sujet ainsi qu'une liste d'outils **:** www.stayinvisible.com

PEUT-ON EFFECTUER DES TRADUCTIONS DE TEXTES SUR LE WEB ?

Des outils gratuits sont disponibles en ligne pour traduire des textes, voire des pages web. Les résultats sont certes souvent discutables, mais pour une première approche, ces technologies peuvent être d'une aide réelle à la recherche.

Sur Voila (technologie Systran)	http://trans.voila.fr
Sur Google (technologie Systran)	http://www.google.fr/language_tools?hl=fr
Sur Alta Vista (technologie Systran)	http://babelfish.altavista.com/
Sur Reverso (technologie Reverso)	http://www.reverso.net

Les agents évolués sur Internet

PRESENTATION

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.

Notons qu'il existe trois grandes familles d'outils :

- Les agents "monopostes" : Ce sont des logiciels à installer sur l'ordinateur. Chaque usager doit installer et paramétrer son agent. Ils sont généralement assez bon marché.
- Les services web : Le service est disponible en ligne, l'usager ne conserve généralement pas non plus les données sur son poste. On paie un abonnement au mois, à l'année, etc.
- Les applications client-serveur : Elles s'installent sur le serveur de l'entreprise et permettent de travailler en réseau, généralement avec une interface web pour les usagers. Ce sont bien entendu les solutions en général les plus coûteuses. Certains éditeurs ont abandonné le marché des logiciels monopostes (exemple Digimind pour son méta-moteur Stratégic Finder ou Intelliseek pour son méta-moteur BullsEye pour se concentrer sur le créneau plus rémunérateur des intranets). D'autres y reviennent, comme par exemple Arisem avec la solution Kaliwatch (qui est tout de même à 8000 euros pour un poste). Nous ne traiteront pas de ces outils dans le cadre de cette formation.

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. Ils se conçoivent aussi dans un objectif d'"archivage du web".

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. Ils rejoignent donc dans cette optique les agents d'alerte : Wysigot, Web Wahcker.

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

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 :

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

Encore des bugs et des imperfections

Les méta-moteurs clients

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

Le plus utilisé au monde, et aujourd'hui pratiquement le seul sur le marché dans la catégorie des logiciels monopostes est Copernic (Copernic technologies) www.copernic.com

La plupart de ces outils sont en effet 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).

Aller plus loin sur 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.

Quelques fonctionnalités

Nouvelle interface plus complète, mais aussi beaucoup plus complexe !

Intégration d'un agent d'alerte, pour surveiller automatiquement les 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

Export dans le format d'origine

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

LES AGENTS D'ALERTE

Ils signalent les modification d'une page ou d'un site web, selon des critères plus ou moins fins, et de manière plus ou moins variée

On distingue :

- ✓ les agents d'alerte web "serveurs" : 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. Exemples :

  Digimind Monitor : www.digimind.fr

  Infominder : www.infominder.com

- ✓ Les agents d'alerte "clients" qui nécessitent le téléchargement d'un logiciel particulier. Exemples :

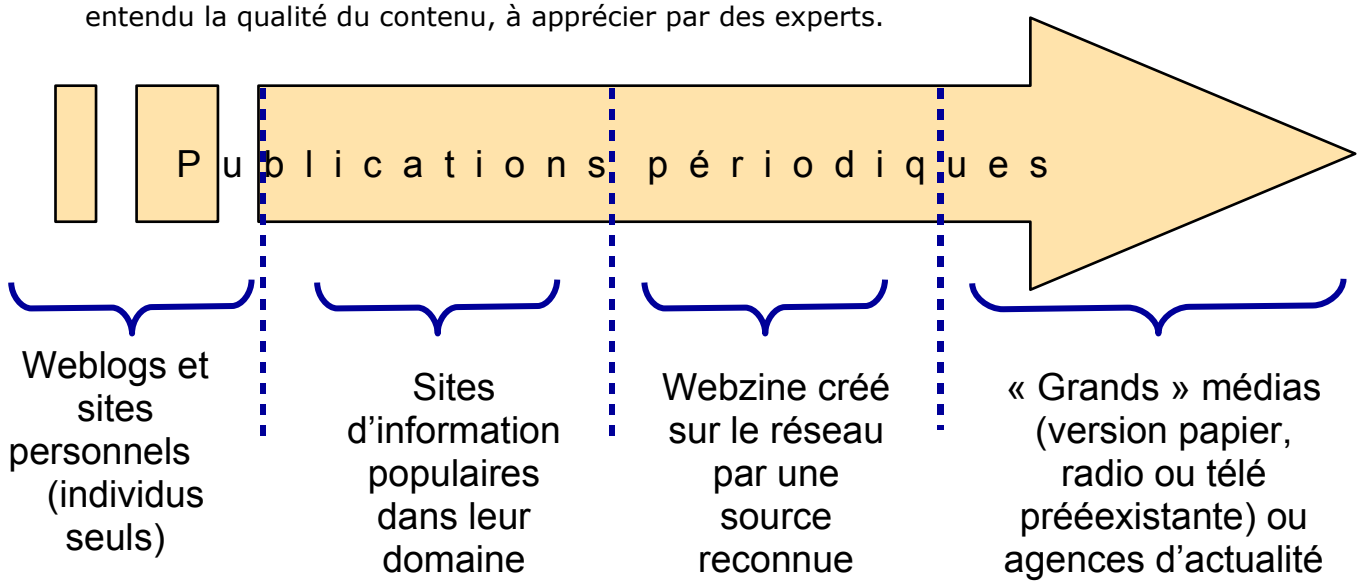
  Website Watcher http://aignes.com

  Webspector www.illumix.com/webspector

Les agents d'actualité

Sur quel type de publications périodiques "travaillent" ces agents ? C'est une question importante à se poser, car le paysage news se complexifie sur internet, du "weblog perso" au grand média à la version papier. Les sources vont du plus ou moins officiel.

Un élément important reste la mise à jour fréquente de la source, sa popularité, et bien entendu la qualité du contenu, à apprécier par des experts.



Trois types d'outils sont disponibles pour avoir des nouvelles fraîches :

- ✓ Les bases d'articles avec archives (serveurs type Factiva)
- ✓ Les agrégateurs d'actualités, qui ne travaillent que sur du flux, sans archives disponibles : Les grands moteurs jouent aujourd'hui presque tous ce rôle (Google, AltaVista, Yahoo, etc.) ; Les agrégateurs ou revendeurs de fils d'information ciblés (Moreover, Yellow Brix, Newshub, RocketNews, Newstrove, NewsIndex, NewsNow (ce dernier peut aussi proposer des archives).
- ✓ Les agégateurs personnels : l'usager choisit ses sources, soit dans un catalogue prédéterminé, soit de manière beaucoup plus large, en ajoutant toute nouvelle source considérée comme pertinente : ex pour le premier cas : Net2One, PIN, mais aussi un outil comme NewsEdge, dans le deuxième cas, on utilise un agrégateur personnel, outil spécifique dédié à la technologie des fils RSS (ou atom).

La syndication personnelle

- ✓ Il s'agit de se créer son propre "fil d'information", à partir de sources spécifiques sélectionnées, mais aussi éventuellement à partir de flux issus de plusieurs sources (flux thématiques, géographiques), voire à partir de flux issus de la surveillance de mots-clés sur un ensemble de sources.
- ✓ Les agrégateurs personnels utilisent un format de fichier dérivé de XML pour récupérer seulement le nouveau contenu d'un site (sans avoir, comme les agents d'alerte, besoin de procéder par comparaison en surlignant les nouveautés) : RSS ou Atom (cf page 7)
- ✓ Un fichier RSS (ou fil RSS = RSS feed en anglais) contient simplement une liste d'éléments (titre, résumé, lien vers URL, voire date, nom de l'auteur). Il a pour

extention .xml ; .rss ; .rdf ou autres qui représentent les plus récentes modifications d'un site

Agents de syndication personnelle

On trouve des agrégateurs en ligne, ou des logiciels à télécharger et installer sur son ordinateur.

- ✓ Agrégateurs en ligne : par exemple RSS4You, Bloglines, mais aussi My Yahoo
- ✓ Logiciels :
  - Autonomes : Sharpreader, Feeddemon
  - Associés à un navigateur ou à client de messagerie : Pluck (Internet Explorer), ou Lektora, nouveau francophone qui s'intègre à Internet Explorer, ou Newsgator

Repérer les bonnes sources

Pour vous aider, consulter la liste d'outils "weblogs et fils RSS" page 39

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

➔ Un très grand volume d'information à disposition

➔ 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

On a vue la richesse d'internet pour la mise en œuvre d'une veille. Cependant, l'exploitation manuelle est souvent délicate du fait de tâches très consommatrices en temps. Les outils (agents, cf "Les agents évolués sur Internet" ci-dessus) permettent une automatisation de tâches répétitives :

- ✓ Outils de collecte (moteurs d'indexation et de recherche, méta-moteurs, agents d'alerte) : ils permettent de surveiller des pages et des sites web (voire des dossiers de pages web), des catégories d'un répertoire, différentes catégories de ressources (actualités, articles de presse, appels d'offre, communiqués de presse, informations financières, etc.). Ils peuvent travailler sur un moteur, un répertoire, plusieurs outils simultanément, une base de données, voire plusieurs bases de données.
- ✓ Outils de tri et d'aide à l'analyse (résumés, traduction, text-mining, cartographies, etc.)

- ✓ Outils d'aide à la diffusion (logiciels push, outils de création de newsletters, outils pour dossiers documentaires, portails, etc.).

Aucune solution informatique ne permet l'automatisation complète de la veille, et certaines technologies sont plus ou moins bien adaptées à telle étape ou tel type de documents (exemple : l'analyse de documents structurés). On utilise assez fréquemment l'association de plusieurs "briques technologiques" pour mener à bien un process de veille automatisé.

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.

- ✓ **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.

POUR EN SAVOIR PLUS (via le web)

SITES D'AUTOFORMATION A L'INTERNET

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

SITES CONSACRES A LA RECHERCHE D'INFO SUR INTERNET

- **GIRI** **www.bibl.ulaval.ca/vitrine/giri**
- **Le "RISI" par Jean-Pierre Lardy http://urfist.univ-lyon1.fr/risi/risi.htm**
- **Sapristi!(Insa Lyon) http://csidoc.insa-lyon.fr/sapristi/digest.html**
- **Netsesame (info économique) www.devinci.fr/infotheq**
- **Infothèque Léonard de Vinci : http://www.devinci.fr/infotheq/guiderec/Mguide.htm**
- **Abondance : www.abondance.com**
- **Indicateur : www.indicateur.com**
- **Intelligence Center : http://c.asselin.free.fr/**
- **Outils froids : http://joueb.com/outilsfroids/**
- **http://www.secrets2moteurs.com/ (ex 1ère position)**
- **Agentland (société Cybion) : www.agentland.fr**
- **La lettre du bibliothécaire québecois : http://www.sciencepresse.qc.ca/lbq/lbq.html**
- **Bases Publications : http://www.bases-publications.com/** permet notamment d'accéder au texte intégral des articles des deux revues, depuis 2001 dans un premier temps, avec toutefois bien sûr un "embargo" d'un an
- **Revue de presse de l'Urfist de Toulouse (pas toujours très à jour, mais intéressante) www.urfist.ccit.fr**
- **http://www.dsi-info.ca/ Le site de Marc Duval au Canada, très bien documenté (notamment langage des moteurs de recherche)**

Citons aussi quelques blogs spécialisés fort utiles :

- Kesaco.canalblog.com
- Aixtal.blogspot.com
- Moteurs.blog.com
- www.zorgloob.com
- influx.joueb.com
- motrech.blogspot.com
-

Les incontournables en anglais

- **Search Engine watch : www.searchenginewatch.com**

  Le blog : http://blog.searchenginewatch.com/blog/

  La lettre journalière : http://searchenginewatch.com/searchday/
- **Websearch : http://websearch.about.com**
- **Research Buzz : http://www.researchbuzz.com/**
- **Fagan Finder blog : http://www.faganfinder.com/blog/**
- **Searchengine showdown : www.searchengineshowdown.com/blog**
- **Resource shelf : http://resourceshelf.freepint.com/**
- **Library Stuff : http://www.librarystuff.net/**
- **Incite / Rubrique de Bellinda Weaver (adapter l'url au mois choisi) : www.alia.org.au/publishing/incite/2003/09/weaver.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 http://sympa.adbs.fr/wws/info/adbs-info

- BIBLIO-FR http://listes.cru.fr/sympa/info/biblio-fr

- MOTRECH http://fr.groups.yahoo.com/group/motrech/

Dico du Net http://www.dicodunet.com/

Dictionnaire collaboratif en ligne sur les e-technos (référencement, mesure d'audience, recherche, hébergement, etc.