Les moteurs de recherche sur Internet

EMAC - Paul Gaborit

juillet 1999

1 Introduction

Internet met à disposition de l'internaute près de 200 millions de pages WEB. Ce nombre continue à augmenter de jour en jour. Retrouver une information précise, exacte et pertinente dans ce « *fouillis* » devient évidemment de plus en plus difficile.

Dans ce document, nous proposons :

- quelques conseils permettant de faciliter cette recherche ;
- une liste d'outils pratiques (permettant de retrouver les moteurs de recherche) ;

Les moteurs de recherche peuvent être classés en quatre familles distinctes :

les moteurs généraux : ils indexent le contenu de toutes les pages WEB des sites qu'ils explorent. Ils se servent pour cela des mots-clés spécifiés dans l'en-tête des pages HTML et du contenu lui-même. Régulièrement, leurs robots d'exploration ré-indexent les sites anciens et explorent les nouveaux sites pour maintenir la base de données à jour. Deux des moteurs le plus connus de cette catégorie sont AltaVista [1] et Lycos [2].

les moteurs thématiques : ils classent les sites par thèmes selon les indications fournis par les créateurs du site ou éventuellement par des personnes chargées d'explorer le WEB. Toutes les pages ne sont évidemment pas classées. Le moteur le plus connu dans cette catégorie est Yahoo en français [3] ou internationnal [4] ;

les méta-moteurs : ce sont des moteurs qui ne savent qu'interroger automatiquement les moteurs des deux autres familles. Cela permet d'être sûr de ne rien manquer mais par contre la fusion des réponses et surtout leur classement n'est pas toujours efficace. Le plus connu de tous ces méta-moteurs est MetaCrawler [5] ;

les moteurs spécifiques : ils sont spécialisés par exemple dans un domaine particulier ou limités à un ensemble de sites ou encore liés à une base de données.

Il existe aussi des moteurs (encore expérimentaux mais déjà bien avancés pour certains) qui recherchent des images, des sons ou des films. Nous ne présenterons pas ici ce genre d'outils.

2 Quelques conseils

Conseil n° 1 - N'oubliez pas qu'Internet n'est pas la seule source d'information.

Bien sûr Internet est une source inépuisable d'information mais ce n'est pas la seule. Les livres, les journaux, les revues, la télévision, la radio, etc. restent aussi de très bonnes sources, souvent plus fiables et surtout mieux ciblées. Il est courant de découvrir une série de sites intéressants grâce à un ou plusieurs URL trouvés par ailleurs.

Chaque chaîne de télévision possède maintenant sa propre émission concernant Internet et même lors d'émissions plus classiques, il n'est pas rare d'entendre l'animateur citer un URL.

Les journaux offrent pour la plupart une rubrique Internet dans leurs colonnes.

L'avantage de passer par un canal d'information plus traditionnel réside dans le travail de recherche, de recoupement et de sélection qu'a fait le journaliste.

Ces informations sont d'ailleurs souvent reprises sur le site WEB du journal ou de l'émission.

1. http://www.altavista.com/
2. http://www.lycos.fr/
3. http://www.yahoo.fr/
4. http://www.lycos.com/
5. http://www.metacrawler.com/

Conseil n° 2 - Commencez toujours une recherche par les moteurs spécifiques puis passer aux moteurs thématiques et enfin utilisez les moteurs généraux.

De nombreux sites proposent un moteur spécifique qui n'indexe que les pages du site. Les informations retrouvées se limitent donc à celles que le site a décidé de mettre à disposition. En revanche, ces moteurs indexent souvent les différentes bases de données du site (Exemple à l'école des mines : le moteur de recherche documentaire [6]). Ils sont donc capables de retrouver une information non-indexée par les moteurs globaux (thématiques ou généraux).

De même, si vous voulez contacter une entreprise ou une personne (et donc retrouver son adresse e-mail), il est toujours plus efficace de passer par les moteurs spécifiques comme celui des pages jaunes [7] ou le bottin [8].

Les moteurs thématiques, quant à eux, classent les sites (ou les pages) par thèmes et sous-thèmes. Cela facilite la recherche lorsque on aborde un sujet général. Il est en effet inutile d'explorer 100 millions de pages WEB pour chercher des photos de Jupiter : la consultation des quelques sites d'astronomie suffit.

En dernier recours, rabattez-vous sur les moteurs généraux.

Conseil n° 3 - Les pages WEB ne sont pas la seule ressource disponible sur Internet. Pensez aux listes de diffusion (mailinglists), aux forum de discussion (newsgroups) et aux personnes elles-mêmes (via la messagerie).

L'information ne s'arrête pas à la lecture d'une ou plusieurs pages WEB. Pour creuser un sujet, il est parfois nécessaire de dialoguer avec d'autres.

Les forum ou groupes de discussion sont un lieu d'échange fantastique. Chaque groupe a sa propre thématique. De nombreux groupes disposent d'une FAQ (Foire Aux Questions) diffusée régulièrement sur le forum (une fois par mois en général). Ce document contient les réponses aux questions les plus fréquemment posées et donne très souvent une liste de sites WEB liés au sujet du forum. Il vous faut donc lire (ou plutôt feuilleter) un groupe de discussion pendant environ un mois pour y lire la FAQ ou vous apercevoir qu'elle n'existe pas (ne pas oublier que les FAQs sont créées et mises à jour par des bénévoles ce qui explique que tous les groupes n'en ont pas).

Si les réponses à vos questions ne sont pas dans la FAQ, la lecture du forum vous sera tout de même bénéfique. Elle vous aura permis de comprendre le sujet réel du groupe, de saisir l'état d'esprit des participants et surtout d'identifier les quelques internautes qui amènent le plus de réponses pertinentes.

Les serveurs de newsgroups ne gardent en mémoire que les messages des derniers jours (cela varie selon le groupe et le serveur entre deux et quinze jours en général). En revanche, il est possible de faire une recherche dans les serveurs d'archivage comme DejaNews [9] qui stockent parfois plusieurs années de discussion !

Pour d'autres sujets, il existe des listes de diffusion (ou mailinglists). Ces listes se présentent sous la forme d'une adresse e-mail qui regroupe en fait tout un groupe d'internautes à travers le monde. Tout message envoyé à cette adresse est rediffusé à tout le groupe soit automatiquement soit après approbation par les *modérateurs* (le ou les modérateurs sont en général les personnes ayant créé la liste de diffusion). L'abonnement est parfois libre, parfois sous contrôle des modérateurs et/ou soumis à certaines conditions. Le programme de gestion de liste de diffusion vous enverra un e-mail de confirmation d'abonnement. Conservez ce message car il contient en général la procédure permettant de se désabonner ! Une fois abonné, vous allez recevoir tous les messages envoyés à la liste et éventuellement y répondre ou poser vos propres questions. Certaines listes de diffusion proposent un archive parfois sur plusieurs années. Tout cela vous est en général expliqué lors de votre abonnement.

Tous ces moyens vous permettent d'entrer en relation avec de nombreuses personnes. Poursuivre une longue discussion en privé (par messagerie électronique) évite de polluer les forum ou les listes de diffusion. Après un temps d'adaptation, vous verrez qu'il n'est pas plus difficile de communiquer et d'instaurer de véritables dialogues avec un internaute qu'avec votre propre voisin (c'est même

6. http://www.enstimac.fr/cgi-bin/doc/find-doc.pl
7. http://www.pagesjaunes.fr/
8. http://www.bottin.fr/
9. http://www.dejanews.com/

parfois plus facile puisque vous avez pu définir au moins un sujet d'intérêt commun).

Conseil n° 4 - Sachez reconnaître un site officiel d'un site officieux ou personnel. Pour cela, demandez à votre navigateur d'afficher les URL consultés et apprenez à décoder ces URL.

Tout le monde peut créer son propre site. L'achat d'un nom de domaine (comme `larousse.fr`) ne coûte pratiquement rien (quelques milliers de francs tout au plus). Or, cela ne présage en rien de la qualité du contenu, puisque ce site (le jour où il existera) n'aura peut-être rien à voir avec l'éditeur bien connus.

Les URL contenant un ~ (tel que ma page personnelle[10]) sont ceux liés à un utilisateur particulier. La qualité et la pertinence des informations sont donc directement liées à la personne qui crée ses propres pages. Certains fournisseurs d'accès (tel que Wanadoo[11] ou ALTERN[12]) proposent et hébergent les pages personnelles de personnes physiques ou morales. Là aussi, apprenez à reconnaître ce genre d'URL pour mieux en juger la pertinence.

Souvent en se connectant à l'URL de la racine du serveur (son URL est constitué de http:// suivi uniquement du nom de la machine machine.domaine.pays), on peut plus facilement identifier la nature de l'organisation qui héberge la page trouvée.

Conseil n° 5 - Apprenez à (re)connaître les sites qui sont de bonnes sources d'informations.

Une *bonne* recherche concernant un thème particulier ne s'effectue pas d'une seule traite. Il faut prendre le temps de lire les différents documents, d'en évaluer la pertinence et de classer les sites qui les hébergent en fonction de la qualité des informations trouvées. De plus, certains sites pourtant de très bonne facture sont absents des moteurs de recherche : la qualité du référencement ne va pas toujours de pair avec la qualité du contenu.

10. http://www.enstimac.fr/~gaborit/
11. http://perso.wanadoo.fr/
12. http://www.altern.com/

Conseil n° 6 - Utilisez les signets (ou bookmarks) afin de consulter régulièrement les sites pertinents sur un sujet donné.

Dans vos signets, regroupez les sites par sujet ou par thème. Lors d'une recherche ultérieure, visitez ces sites en premier (profitez-en pour nettoyer les URL qui ne fonctionnent plus). Un site vivant et bien tenu donne en général des informations à jour et constitue donc une bonne base de départ pour continuer une recherche. Méfiez-vous tout de même des sites qui évoluent trop vite : il arrive parfois que l'information intéressante disparaisse !

Conseil n° 7 - Pensez à spécifier la langue voulue... ou (au pire cas) utilisez les traducteurs automatiques.

Il est toujours dommage de trouver 243 pages WEB en réponse à une requête très précises pour s'apercevoir quelques minutes plus tard que 10 sont en japonais, 97 en allemand et le reste en anglais alors qu'on ne connaît que le français et quelques mots d'espagnol.

Certains moteurs se limitent aux pages francophones. D'autres permettent de spécifier la langue demandée. Pensez dans ce cas à choisir la langue. Certains mots sont évidemment spécifiques à une langue donnée et n'apparaissent donc que dans des pages rédigées dans cette langue. Mais méfiez-vous des faux amis...

Une dernière remarque : la plupart des moteurs indexent les mots sans tenir compte de l'accentuation mais ce n'est pas toujours le cas. Pensez-donc à saisir les accents si le système utilisé le permet.

Conseil n° 8 - Ne vous limitez pas aux premières pages trouvées.

Le moteur utilisé pour la recherche produit une liste classée par ordre de pertinence... mais sa définition de la pertinence n'est pas la vôtre ! Explorez donc les pages trouvées à la recherche d'autres liens que le moteur ne vous a pas signalés où qu'il a classé trop loin pour que vous les voyez.

C'est parfois au détour d'une page secondaire ou de troisième niveau que l'on découvre l'information recherchée...

Conseil n° 9 - Si elle existe, pensez à lire l'aide en ligne des moteurs que vous utilisez.

La plupart des moteurs propose un mode d'emploi. Lisez-le et comprenez-le. Puis choisissez un ou deux moteurs et apprenez à bien vous en servir plutôt que de changer tout le temps de moteur.

Le liste suivante illustre tout cela à partir de AltaVista (à votre avis, AltaVista fait-il partie de mes moteurs favoris ? `;-)`).

`html`
: 105 422 472 pages

`text:html`
: 6 471 533 pages

`+text:creation +text:html`
: 316 179 pages

`+text:creation +text:html -programmation`
: 286 701 pages

En effet, la plupart des pages WEB sont écrites en html et elles contiennent donc toutes le mot *html* sous une forme ou sous une autre. En précisant `text:html`, vous obligez le moteur AltaVista à ne rechercher le mot *html* que dans le texte lui-même et non pas dans le source HTML de la page. Le `-programmation` permet de spécifier qu'on ne veut pas des pages contenant le mot *programmation*.

Conseil n° 10 - Recoupez les informations entre différentes sources.

Ce conseil n'est bien évidemment pas du tout spécifique à Internet. Cela devrait être la base de toute recherche documentaire.

Attention aux rumeurs : certaines informations sont reprises à travers Internet sans aucune vérification et deviennent très vite des légendes. L'annonce de certains faux virus en est un exemple très classique.

Conseil n° 11 - Utilisez les combinaisons de mots clés les plus spécifiques possibles. Utilisez les mots clés les plus génériques possibles.

Ces deux conseils semblent apparemment contradictoires. Ils sont pourtant vitaux.

En effet, une recherche sur un simple mot clé peut ramener des milliers voire des centaines de milliers de documents. Il est évidemment utopique de consulter tous ces documents pour y retrouver ce qu'on cherche. Il faut donc mieux focaliser la recherche en combinant différents termes. Préciser plusieurs mots dans la recherche permet au moteur de croiser les résultats. La liste suivante (créée avec AltaVista[13]) illustre ce conseil :

`verre`
: verre : 111 282 pages

`verres`
: verres : 20 310 pages

`contact`
: contact : 31 446 156 pages

`verre de contact`
: verre : 111 282 pages
: contact : 31 446 156 pages
: de : 294 245 739 pages

`"verres de contact"`
: verres de contact : 398 pages

Dans d'autres cas, la combinaison est trop spécifique ou utilise un mot que personne n'utilise.

Il faut donc être capable :

- de trouver de nombreux synonymes ;
- d'exprimer sa requête en combinant différents termes ;
- de remonter vers des informations plus génériques ;
- de traduire les mots recherchés dans différentes langues.

3 Quelques liens pratiques

Attention : ces liens sont donnés à titre indicatif. Les choses évoluent très vite sur Internet. Un moteur de qualité aujourd'hui peut rapidement devenir obsolète.

Tout d'abord quelques moteurs de recherches francophones (à titre d'exercice, le lecteur pourra essayer de retrouver le site d'origine de ces listes) :

Chut (France) : http://www.chut.com/

Ecila (France) : http://www.ecila.fr/
http://www.ecila.com/

Excite (États-Unis) : http://www.excite.fr/

Euroferret (Angleterre) :
http://www.euroferret.com/french/

13. http://www.altavista.com/

Guide des Services Francophones (France) :
http://ns1.jetmultimedia.fr/gsf/

Lokace (France) : http://www.lokace.com/

Lycos (Etats-Unis) : http://www.lycos.fr/

Netscan (France) : http://www.netscan.fr/

The Blue Window (Suisse) :
http://www.sear.ch/index_f.html

Voilà (France) : http://www.voila.fr/

Quelques méta-moteurs :

All-in-One Search :
http://www.media-prisme.ca/all/all1srch.html

Ariane : Méta-moteur de recherche
http://www.espace2001.com/moteur/index.html

Debriefing : Utilise les principaux annuaires et moteurs de recherche
http://www.debriefing.com/france/

EDirectory : Moteurs de recherche du monde, classés par pays http://www.edirectory.com/frindex.html

Général Moteur : L'annuaire des annuaires
http://www.acorus.fr/general/

Méga Francité : Recherche dans 1500 moteurs
http://mega.francite.com/

Metacherche : Recherche directe sur les principaux moteurs et méta-moteurs. Propose un classement par rubriques des annuaires spécialisés et des sites portails http://www.metacherche.8m.com

MetaW! :
http://www.multimania.com/www/metaw.html

MetaW3 : http://www.metaw3.com/indexfr.html

Rexweb : Interroge huit outils simultanément
http://www.rexweb.com/

Subway Search : Assistant de recherche sur Internet
http://altern.org/subway/

Swiss InterWeb : Recherche à partir d'une seule requête sur plusieurs annuaires et moteurs de recherche (dont certains sont suisses)
http://search.siw.ch/

Trouvez ! : Recherches dans 16 répertoires et moteurs de recherche http://www.trouvez.com/

Ces listes ne sont évidemment pas exhaustives ! Une bonne recherche sur Internet devrait vous permettre de trouver vous-mêmes le ou les moteurs qui vous conviennent.