TECHNO 21

Publication technique de la SmalS-MvM / Technische uitgave van de SmalS-MvM

La recherche d'information sur Internet

Introduction

Isabelle Boydens
est consultante à la section Recherches. Docteur en Philosophie et Lettres, orientation Sciences de l'information et de la documentation, elle enseigne à l'Université Libre de Bruxelles. Elle s'est spécialisée dans l'analyse critique et la modernisation des systèmes d'information administratifs (bases de données, groupware, workflow, systèmes d'indexation et de recherche documentaire) et est notamment responsable du sous-projet "instructions" de la DmfA (Déclaration multifonctionnelle – Multifunctionele Aangifte).
Contact: 02/509.59.92

Le recours à Internet en vue de rechercher de l'information est de nos jours incontournable. Toutefois, l'information y est aussi pléthorique qu'hétérogène. Internet, ce "fonds sans fond", se déploie en effet sans organisation ni structure. Quant au volume, il est impossible d'évaluer la taille exacte du Net puisque chaque moteur d'indexation et de recherche n'indexe qu'une partie du Web, l'ensemble demeurant inconnu dans l'absolu. De surcroît, chaque moteur peut indexer un sous-ensemble distinct du Web. Selon certaines sources, Internet contiendrait (au premier trimestre 2001) plus de deux milliards de pages, avec un taux de croissance de 7 millions de pages supplémentaires par jour[1]. Selon d'autres, il y aurait 550 milliards de pages sur le Web (étude de *BrightPlanet*, 1er août 2000) : on parle de "Web invisible" ou de "*Deep Web*"[2].

Des recherches menées par IBM, Altavista et Compaq à propos de la structure des liens sur Internet ont démystifié l'image de la "toile" et révélé une structure en "nœud papillon" ("*bow-tie theory*")[3]. L'étude, qui date de 1999, repose sur l'analyse de plus de 200 millions de pages[4] (Figure 1) :

- un tiers des sites correspondraient à un cœur fortement interconnecté (16 "*clics*" maximum permettraient de passer d'une page à une autre);
- un quart du Web serait constitué de "pages d'entrée" qui pointeraient de façon unidirectionnelle vers le cœur (*peu de pages pointent vers ces sites…)*, ce dernier pointant de façon unidirectionnelle vers un autre quart (*ces pages constituent une sorte de "cul de sac"… il est impossible ou presque d'en sortir par des liens*);
- enfin, environ un cinquième des pages serait constitué de sites qui sont reliés non pas au cœur du Web mais à ses extrémités.

A cela s'ajoute la présence d'îlots, groupes de sites isolés et déconnectés des autres sites.

---

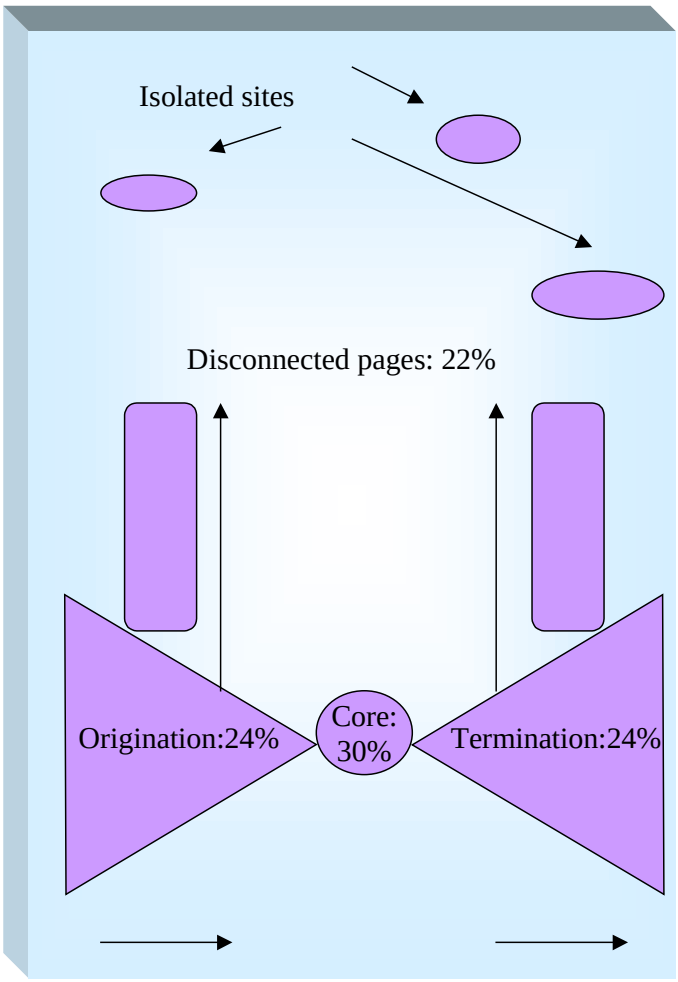
[1] *Quelle est la taille du Web actuellement ?* http://www.abondance.com/docs/ Voir aussi : http://www.cyveillance.com/

[2] *550 milliards de pages sur le Web ? (actualités du 1er août 2000).* http://www.abondance.com/actu/semaine.html.

[3] BRODER A., KUMAR R., MAGHOUL F., RAGHAVAN P., RAJAPOLAPAN P., STATA R., TOMKINS A. et WIENER J., *Graph Structure in the Web*, 1999 (http://www.almaden.ibm.com/cs/k53/www9.final/). MONTCULIER C., L'organisation des pages sur Internet. *O1 Informatique*, 12 janvier 2001, n° 1616, p. 16. ROUMIEUX O., Noeud pap' de rigueur sur le Web. *Archimag*, juin 2000, n° 135, p. 25.

[4] Outre le caractère restreint de l'échantillon, remarquons le fait que l'étude ne prend pas en considération les pages "dynamiques".

Figure 1. Structure du Web : "*the bow-tie theory*"



Source: IBM, Altavista et Compaq, 1999

Quant au fond, les documents publicitaires, les informations ineptes, les informations officielles et les informations scientifiques très "pointues" se cotoient sans qu'il soit possible de les distinguer a priori sur base d'une "simple" recherche par mots-clés. De telle sorte que certains parlent, pour qualifier le WWW (*World Wide Web*), de MMM (*Multi-Media Mediocrity*)[5] …

Dans les lignes qui suivent[6], nous présentons en premier lieu les difficultés majeures que soulève la recherche sur Internet (1). Ensuite, nous évaluons les principales classes d'outils de recherche disponibles sur Internet (2) et nous formulons enfin plusieurs recommandations syntaxiques en vue de les exploiter au mieux (3). Dans les conclusions de l'étude, nous présentons un tableau synthétique des usages recommandés par type d'outil de recherche et nous évoquons quelques perspectives d'évolution des outils de recherche actuels.

[5] SMITH A. G., Testing the Surf : Criteria for Evaluating Internet Information Resources. *The Public-Access Computer Systems Review*, 1997, vol. 8, n°3 (http://info.lib.uh.edu/pr/v8/n3/smit8n3.html). KOBOYASHI M. et TAKEDA K., Information Retrieval on the Web. *ACM Computing Surveys*, juin 2000, vol. 32, n° 2, p. 144-173.

[6] La validité des liens (URL) cités dans cette étude a été testée la dernière fois le 03 septembre 2001.

1. Les difficultés de la recherche sur Internet

Parmi les difficultés que soulève la recherche sur Internet, nous évoquons successivement la question de la polysémie[7] et de la synonymie[8] dont les effets pervers sont accrus en raison de la taille du Web et de la pratique croissante du "*spamming*", phénomène que nous exposons ci-dessous. Enfin, nous abordons plusieurs incidences du caractère évolutif du Web et des index qui y donnent partiellement accès.

1.1. Polysémie et synonymie à "grande échelle"

Deux exemples permettront d'illustrer les difficultés soulevées par la polysémie. Une recherche sur Altavista[9] sur le mot-clé *ESB* peut donner 45.465 résultats (pages Web[10]) référant à des thèmes aussi hétérogènes que l'*European Society of Biomechanics* (ESB), l'*Encéphalopathie Spongiforme Bovine* (ESB), le journal néerlandais "*Economisch Statistische Berichten"* (ESB) ou encore l'*European Soil Bureau* (ESB), … De même, une recherche sur le mot "*jaguar*" peut donner 452.230 résultats, incluant des sites relatifs au félin, à la marque de voiture ou encore, à la célèbre équipe de football américain ! Dans ces conditions, il semble difficile de cibler le contenu d'une recherche… La synonymie pose un autre type de question : par exemple, une recherche sur le mot-clé "*banque de données*" ne permet pas nécessairement de retrouver les sites relatifs à ce thème et indexés exclusivement avec des formes synonymes telles que "*base de données*".

Il est possible d'atténuer les difficultés précitées en faisant un usage judicieux des outils de recherche disponibles sur Internet (voir point 2 ci-dessous) et surtout en formulant correctement les équations de recherche (voir point 3 ci-dessous).

Au niveau de l'équation booléenne, les effets de la polysémie (donnant lieu à du "bruit" : obtention de résultats jugés non pertinents) peuvent être partiellement neutralisés en ajoutant au mot-clé d'autres termes qui en précisent le contexte, chaque terme étant relié par des "AND" booléens : dans notre exemple, une recherche (en français) sur la "maladie de la vache folle" pourrait se traduire par : (*ESB* AND "*Encéphalopathie Spongiforme Bovine"*) OR (*ESB* AND *"maladie de la vache folle"*)[11].

Les effets de la synonymie (donnant lieu à du "silence" : non obtention de résultats jugés pertinents) peuvent être réduits en incluant dans l'équation de recherche un maximum de termes synonymes, reliés par des "OR" booléens : dans notre exemple, une recherche sur les banques de données pourrait se traduire par : "*banque de données*" OR "*bases de données*" OR "*database*" OR "*databank*", etc.

Toutefois, une autre question non moins cruciale se pose : celle de la classification des sites ou pages obtenus en réponse par ordre de "pertinence" relative, étant donné le nombre important de résultats affichés. Souvent, on obtient plusieurs dizaines, voire plusieurs centaines de milliers de résultats et l'on considère que l'utilisateur "moyen" consulte tout au plus les 20 premières réponses. Le tri des résultats est donc fondamental. Or, ce tri est rendu particulièrement ardu en raison de la pratique croissante du *spamming*.

---

[7] On parle de polysémie quand un même mot ou une même expression revêtent des sens différents. Pour plus d'information concernant le traitement de la polysémie dans le domaine documentaire : LEFEVRE P., *La recherche d'informations. Du texte intégral au thesaurus*. Paris : Hermès, 2000, p. 26-36.

[8] Deux mots ou expressions sont synonymes s'ils ont le même sens. Pour plus d'information concernant le traitement de la synonymie dans le domaine documentaire : LEFEVRE P., *op. cit.*, Paris : Hermès, 2000, p. 23-26.

[9] http://www.altavista.com

[10] Le nombre de résultats obtenus en réponse peut être réduit sur de nombreux moteurs d'indexation et de recherche en sélectionnant l'option "*one result per Web site*".

[11] Ou en anglais : (BSE AND « *Bovine Spongiform Encephalopathy* ») OR (BSE AND « *Creutzfeldt-Jacob Disease* »).

1.2. La pratique croissante du spamming

Les moteurs de recherche utilisent des règles empiriques pour classer les réponses par ordre de pertinence relative. Ces règles de classement, dont l'algorithme est souvent opaque[12], reposent sur le comptage des occurrences de mots-clés répertoriés dans les sites obtenus en réponse. Diverses améliorations pondèrent l'importance des mots selon :

- le caractère discriminant d'un mot-clé évalué en fonction de sa rareté dans l'ensemble des termes indexés,
- l'apparition du mot dans les «méta-tags», les titres, au début des chapitres ou en caractères plus gros que le corps du texte (mais d'un genre «littéraire» à l'autre (roman ou thèse de doctorat), les mots contenus dans le titre peuvent être plus ou moins révélateurs du contenu..).

Ces règles de classement sont problématiques en raison des phénomènes de synonymie et de polysémie évoqués plus haut mais aussi en raison de la pratique du "spamming". Le mot «*spamming*»[13] est issu du mot anglais *spam*, marque de jambonneau couvrant tous les dialogues de la série télévisée britannique, *Monty Python Flying Circus*[14]. Le "*spamming*" désigne une pratique de "matraquage" consistant, dans le cadre de l'indexation des sites Web, à abuser des "mots-clés" les plus populaires afin d'améliorer artificiellement le classement de certains sites : ceux-ci incluent, par exemple, plusieurs fois dans leurs "méta-tags" des termes habilement choisis, n'ayant parfois aucun rapport avec le contenu du site indexé, ces mots-clés pouvant figurer dans des polices et des couleurs invisibles à l'œil.

Afin de lutter contre ce phénomène, certains moteurs d'indexation et de recherche refusent l'indexation des sites indiquant des mots-clés récurrents et calculent le taux de pertinence d'un site sur base de son taux de citation par d'autres sites faisant autorité[15]. Les moteurs privilégient ainsi des fonctionnalités de recherche "non spammables" ou difficilement "spammables" telles que la popularité et l'indice de clic[16]. La méthode "*PageRank*" du moteur *Google* (http://www.google.com), donnant un score d'autant plus élevé à un site qu'il est pointé par beaucoup d'autres sites, constitue un exemple remarquable de ce type de technique. Celle-ci restera à être affinée en vue de prendre en considération les nouveaux sites qui, par la force des choses, n'ont pas eu le temps d'être référencés de façon massive mais peuvent présenter de l'information "*up-to-date*" dans des domaines extrêmement pointus.

---

[12] "*Furthermore, systems often use surprising query transformations, unpredictable stemming algorithms, and mysterious weightings for fields. And in many systems, the results are displayed in a relevance renaking whose meaning is a mystery to many users (and sometimes a proprietary secret).*" SHNEIDERMAN B., BYRD B. et CROFT W. B., Sorting out Searching. A User-Interface framework for Text Searches. *Communication for the ACM*, avril 1998, vol. 41, n° 4, p. 95-98.

[13] Association IPEA (Internet Positioning European Association), *Charte de qualité et de déontologie : définition du spam* (2000). http://www.abondance.com/docs/spam.html

[14] Membres du projet Clever (IBM), Recherche intelligente sur l'Internet. *Pour la Science*, août 1999, n° 262 (http://www.pourlascience.com/numeros/pls-262/art-4.htm).

[15] ROUMIEUX O. Annuaires et moteurs. Les tendances du prêt-à-chercher. *Archimag*, juin 2000, n° 135, p. 25.

[16] ANDRIEU O., L'évolution se trouve du côté de la mixité annuaire et du moteur. *Archimag*, juin 2000, n° 135, p. 28.

La lutte affichée contre le *spamming* n'empêche pas une certaine hypocrisie puisque de nombreux services, payants ou gratuits, publient régulièrement sur Internet le "hit parade" des mots-clés les plus populaires, facilitant ainsi la pratique du "*spamming*"… Ainsi, pour l'année 2000, les mots "*Britney Spears*", "*Pokemon*" et "*Napster*" semblent être les termes les plus demandés sur les outils de recherche *Altavista*, *Lycos* et *Yahoo!*[17].

Enfin, le *référencement payant* (assurant le référencement d'un site sous un ensemble de mots-clés donnés mais sans garantie quant au positionnement) et le *positionnement payant* (assurant en outre un classement favorable du site pour un mot-clé donné) se généralisent sur le Web[18] et modifient à nouveau les règles du jeu. En février 2001, l'annuaire *Yahoo!* a officiellement proposé une offre de positionnement payant en ce qui concerne le classement des sites par catégorie. Le montant à payer varie avec l'indice de popularité de la catégorie et avec les "enchères" en cours[19]. Avec le développement de ces pratiques, le Web s'apparente de plus en plus aux pages jaunes du Bottin téléphonique : les plus offrants disposent de l'encart publicitaire (ou du positionnement) le plus avantageux.

1.3. Evolution continue du Web et des index

Une autre difficulté réside dans l'évolution continue du Web et, en corollaire, des adresses des sites et pages correspondantes. S'il est impossible de résoudre le problème des modifications d'adresses, certains logiciels permettent la réactualisation dans le temps du contenu des sites auxquels renvoient les bookmarks[20] (selon une périodicité déterminée par l'utilisateur). Il existe à l'heure actuelle de nombreux outils spécialisés dans la gestion et le "rafraîchissement" des bookmarks (voir par exemple : http://internet-gopher.com/toolkit/bookmark.htm)[21]. Toutefois, ceci n'empêche pas l'émergence des "liens cassés" (ou "*dead links*", menant à la fameuse "erreur 404"), correspondant à des adresses obsolètes. Il existe des services destinés à détecter les liens "cassés" au sein d'un site donné ou dans un ensemble de sites (par exemple : http://websitegarage.netscape.com/). Notons par ailleurs que certains outils tels qu'*Alexa*[22] *ou Google*[23] permettent de pallier partiellement cette fameuse "erreur 404" en conservant dans une mémoire "cache" l'image du site correspondant à l'ancien URL, devenu obsolète après l'indexation du site correspondant.

Enfin, si le Web évolue, les index des outils de recherche évoluent également, ce qui a pour conséquence que deux recherches identiques menées le même jour via un même moteur de recherche ne donnent pas nécessairement les mêmes résultats. La périodicité de rafraîchissement des index est variable d'un outil à l'autre : les délais moyens (situation en mai 2000) de rafraîchissement (renouvellement complet) d'un index varient entre 4 à 6 semaines (*Altavista*) et 1 semaine (*WebCrawler*)[24].

---

[17] *Britney Spears, grande gagnante des mots-clés sur le Web* (le 05/01/2001), http://www.abondance.com/actu/actu0101.html.

[18] *Lettre "Actu Moteurs". Actualité des outils de recherche.* Semaine du 5 au 9 février 2001.

[19] L'outil de positionnement GoTo (http://www.goto.com) est l'un des plus actif dans ce genre de services auxquels de nombreux moteurs, tel qu'Altavista, ont recours. ANDRIEU O., *GoTo, nouvel acteur majeur dans le monde des outils de recherche.* http://www.abondance.com/trucs-et-astuces/outils20.html

[20] Les bookmarks ou "signets", renvoient à un URL (*Uniform Resource Locator*), adresse "absolue" identifiant une page dans l'environnement hypertextuel du Web.

[21] On trouvera des liens vers la plupart de ces outils à l'adresse suivante : http://www.abondance.com/ressources/bookmarks.html

[22] ANDRIEU O., *Un coup de barre, Alexa, et ça repart.* http://www.abondance.com//trucs-et-astuces/outils15.html

[23] ANDRIEU O., *Google, la mémoire du Web "caché"* (http://www.abondance.com/trucs-et-astuces/recherche22.html).

[24] http://www.abondance.com/outils/comparatif.html

2. Typologie des outils de recherche

Après la présentation de quelques liens directs vers certaines sources de base, nous évoquons successivement les outils de recherche suivants : annuaires et moteurs d'indexation et de recherche, méta-moteurs de recherche, logiciels de veille documentaire, logiciels de cartographie documentaire, outils de visualisation des copies d'écran, outils de recherche en langage naturel et enfin, outils de recherche reposant sur l'expertise humaine (ou "éditeurs humains").

Dans la pratique, ces outils de recherche sont complémentaires: nous les présentons ici successivement en fonction de leur ordre chronologique d'apparition dans l'environnement d'Internet. Nous exposons ci-dessous les principales fonctionnalités de chaque outil ainsi que des recommandations méthodologiques les concernant. Dans les conclusions de cette étude, nous présenterons un tableau synthétique des usages recommandés par outil en fonction du type de recherche que l'utilisateur souhaite réaliser.

2.1. Accès direct à quelques sources de base

Avant d'aborder les outils de recherche proprement dits, nous présentons plusieurs sources auxquelles l'utilisateur peut accéder directement en vue de trouver éventuellement une réponse rapide à ses questions :

- **les encyclopédies en ligne,** dont certaines sont gratuites (Webencyclo, http://www.webencyclo.com/) et d'autres payantes (*Encyclopaedia Universalis*, http://www.universalis-edu.com - *Encyclopaedia Britannica*, http://www.britannica.com/)
- **les outils d'accès à l'actualité** (dépêches de presse, etc.) : *E-revue* (http://www.e-revue.com/), sites Web des périodiques, comme le journal *Le Monde* (http://www.lemonde.fr/), ...
- **les FAQs (*Frequently Asked Questions* ou *Foire Aux Questions*)** rassemblent les réponses - généralement rédigées par les experts d'un domaine - aux questions les plus fréquemment posées dans un secteur considéré. Il existe des sites de FAQ (http://www.faqs.org/faqs/).
- **les archives de forums de discussion** : espaces communautaires permettant aux internautes d'échanger questions et informations sur un domaine considéré, les "forums de discussion" constituent une autre source intéressante. L'outil de recherche *Google* permet notamment d'accéder aux archives de ces sources (http://groups.google.com/).

Notons que certains outils de recherche, tel *Northern Light* (http://www.northernlight.com) offrent, outre les fonctionnalités de recherche envisagées ci-dessous, un accès à plusieurs milliers de sources non disponibles sur le Web, moyennant une somme modique (1 à 4 dollars par article) : journaux, livres, magazines, dépêches de presse, ... *Northern Light* inclut également un système de classification des résultats par thème de recherche, un outil d'alerte (envoi par mail de la liste des nouveautés répondant à une question donnée) ainsi qu'un utilitaire ("*Geo Search*") permettant d'affiner ses recherches sur une base à la fois thématique et géographique.

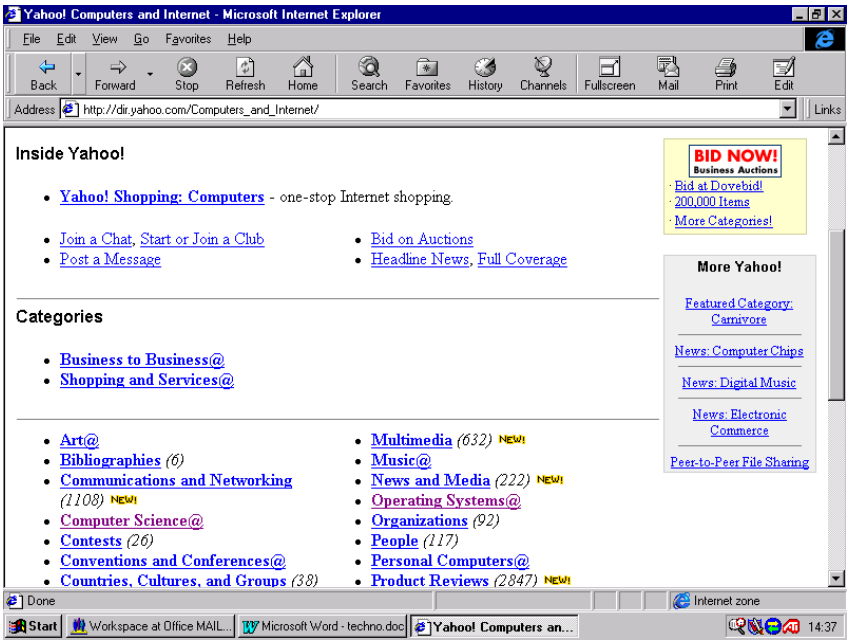
2.2. Annuaires et moteurs de recherche

La distinction entre annuaires et moteurs d'indexation et de recherche est fondamentale.

Les annuaires (Figure 2) :

- assurent une classification manuelle des *sites Web*, classification réalisée par des "netsurfers"[25] : le nombre de sites retenus est inférieur au nombre de sites retenus dans le cas des moteurs d'indexation et de recherche.
- correspondent à des catalogues référençant les sites Web par thèmes : de thème en sous-thème et d'étape en étape, l'utilisateur peut accéder à des listes d'adresses de sites référencés par catégorie. Cette structuration correspond à un langage documentaire classificatoire: le caractère monohiérarchique et figé de la classification ne simplifie pas toujours la recherche car il est parfois très difficile de savoir dans quelle catégorie rechercher un thème donné. Par exemple, sur *Yahoo!*, pour trouver les concessionnaires d'automobile, les deux chemins suivant sont possibles :
  - *commerce et économie ---> produits et services pour les particuliers ---> véhicules ---> quatre roues motrices ----> vente et distribution*
  - *sports et loisirs ---> véhicules*
- représentent une forme de garantie de "qualité" en ce sens que les site Web retenus dans la classification sont sélectionnés "manuellement" par des "netsurfeurs", lesquels traitent de 50 à 100 sites par jour et en rejettent en moyenne un tiers jugés inintéressants pour des raisons d'absence de contenu, d'obsolescence, de lenteurs, etc.[26]

Figure 2. Exemple d'annuaire : langage de recherche classificatoire



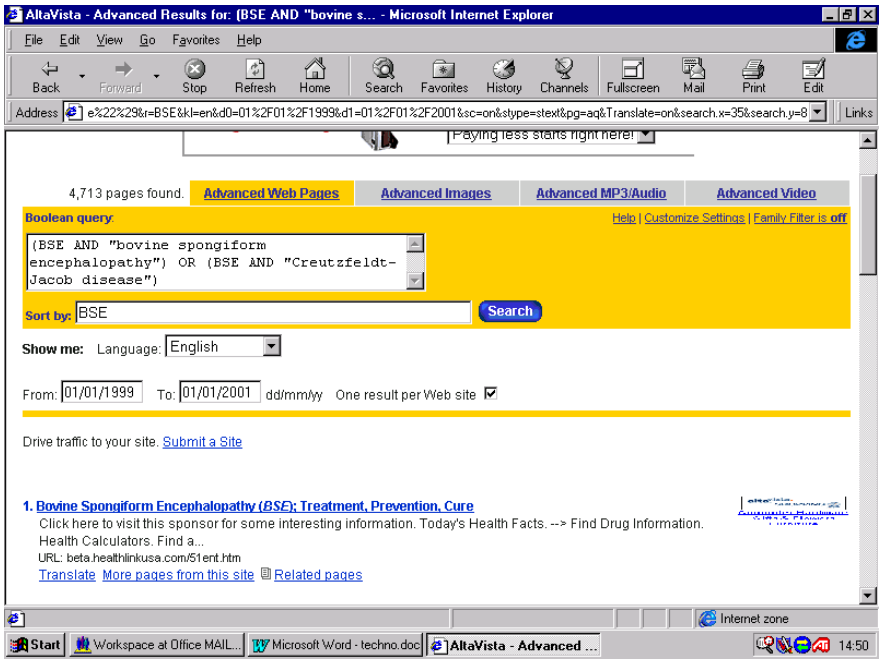
[25] Signalons l'initiative de l'*Open Directory* qui fait appel à des volontaires en vue d'assurer une classification plus appofondie et plus "réactive" de l'annuaire. Par rapport aux annuaires traditionnels, l'Open Directory présente le risque d'une certaine disparité. http://www.dmoz.org/

[26] ANDRIEU O., *Facteurs bloquants : 18 façons de rater son référencement*. Juin 2000. (http://www.abondance.com/docs/contraintes.html).

Les moteurs d'indexation et de recherche (Figures 3 et 4)

- assurent une indexation automatique des *pages Web*[27].
- reposent sur une recherche documentaire de type combinatoire (Figure 3 : combinaison booléenne de mots-clés).

Figure 3. Exemple de moteur : langage de recherche combinatoire



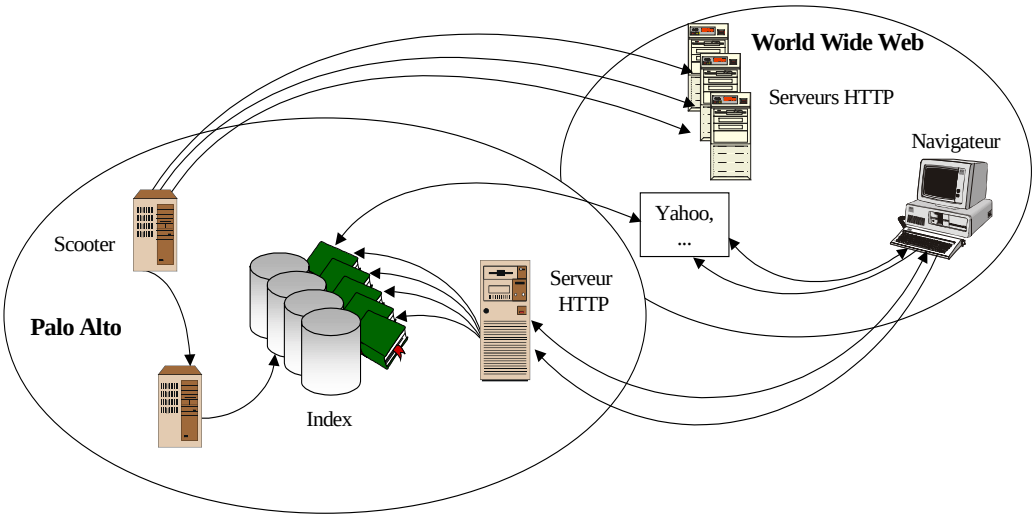
- fonctionnent comme suit : un robot (*scooter*) parcourt le réseau, identifie les nouvelles pages ainsi que les nouveaux liens entre pages; ensuite, les nouveautés et modifications par rapport à l'état antérieur sont transmises à un "indexeur" qui ajuste les fichiers d'index. Enfin, un gestionnaire de requêtes situé sur un serveur HTTP interroge les serveurs d'index et retourne les réponses aux browsers, clients légers via lesquels nous accédons au Web (voir Figure 4, présentant à titre d'exemple le moteur Altavista (http://www.altavista.com)[28].
- ne représentent pas l'ensemble du contenu de l'internet (en 2000, on estime que 60% environ des pages Web ne sont pas vues; Altavista indexerait de 12 à 14% du Web). Ce taux assez faible est dû en partie au fait que les moteurs d'indexation et de recherche excluent de plus en plus radicalement les "doublons" ainsi que les sites suspectés de "spam"[29].

[27] Selon les cas, le moteur prend en considération les termes placés entre les balises "méta", les mots du titre ou du résumé ou encore, tous les termes figurant dans l'ensemble du code HTML des pages.

[28] LELOUP C., *Moteurs d'indexation et de recherche. Environnements client-serveur, Internet et Intranet.* Paris : Eyrolles, 1998, p. 168.

[29] en 1998, on estime que 200 millions sur les 500 millions de pages de l'internet seraient alors indexées par l'ensemble des moteurs d'indexation et de recherche. SAMIER H. et SANDOVAL V., *op. cit.*, p. 15. Voir aussi : LAUR B., *Intranet : la synthèse. Séminaire Cap Gemini 3-5 mai 2000*. Paris : Cap Gemini, 2000, p. 1/15. Notons que ce problème ne se pose pas dans le cadre d'un intranet.

Figure 4. Schéma de fonctionnement du moteur Altavista



Source: Leloup C., Moteurs d'indexation et de recherche, Environnement client-serveur, Internet et Intranet. Paris: Eyrolles, 1998, p.168

Dans la pratique, annuaires et moteurs sont complémentaires. Il est recommandé de recourir aux annuaires en vue de réaliser des recherches sur des catégories générales et de recourir aux moteurs d'indexation et de recherche lorsqu'on effectue une recherche sur des mots-clés très pointus. De plus en plus, les outils de recherche, tel que Google, associent les deux types de langage, classificatoire et combinatoire.

2.3. Méta-moteurs de recherche

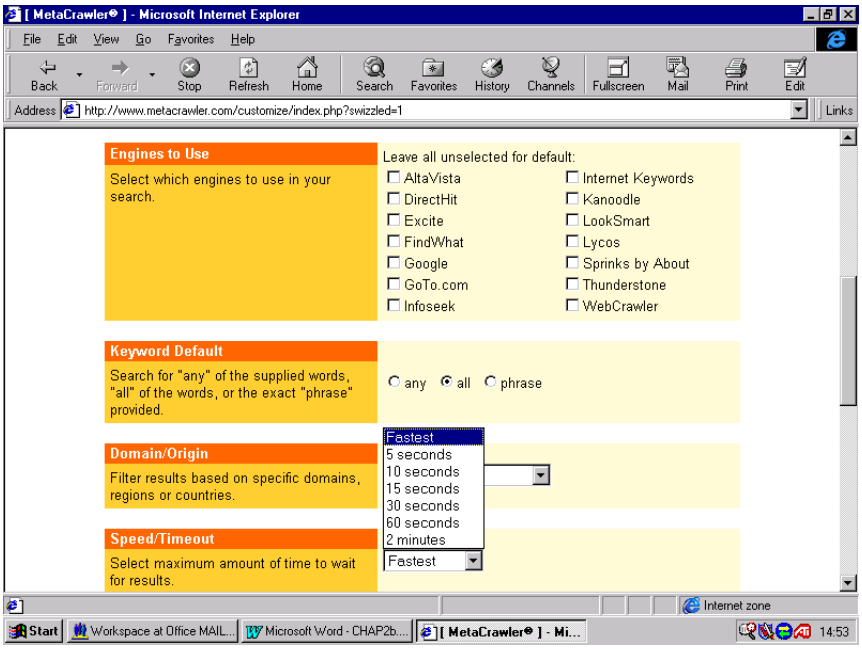
Les «méta-moteurs» (Figure 5) permettent une recherche en parallèle dans plusieurs moteurs de recherche et n'ont pas d'index propre sur leur site. Les méta-index[30]:

- fonctionnent sur base de programmes qui offrent une interface rapide et évoluée entre l'utilisateur et les moteurs, traduisent la requête et la soumettent simultanément à plusieurs moteurs de recherche;
- reçoivent les 10 à 50 premières réponses de chaque moteur, les compilent en éliminant les doublons et fournissent les résultats ainsi que l'identification des moteurs qui en ont permis la sélection;
- se différencient sur base:
  - du nombre de moteurs consultés: de 6 (Highway61) à 13 (Metacrawler) et 17 (Dogpile);
  - du nombre de réponses prises en compte par moteur;
  - du respect de la formulation des requêtes pour chaque moteur (vu leurs caractéristiques respectives);
  - du format des résultats.

[30] SAMIER H. et SANDOVAL V., *La recherche intelligente sur l'internet*. Paris : Hermes, 1998, p. 26-28.

- offrent un gain de temps et une vision synthétique des résultats d'une recherche (mais en contre partie, les résultats ne sont, par définition, pas exhaustifs);
- impliquent la prise en compte de la ***syntaxe hétérogène de chaque moteur de recherche*** utilisé simultanément (type d'opérateurs, format des entrées d'index, etc.).

Figure 5. Méta-moteur de recherche : exemple



Il est conseillé de recourir aux méta-moteurs au seuil d'une recherche, afin d'avoir rapidement un aperçu général et synthétique de l'information disponible sur un thème donné.

2.4. Logiciels de veille documentaire

Appelés par certains "agents intelligents" ou "logiciels de veille", les logiciels de recherche semi-automatique offrent des fonctions voisines de type "push" (système d'alerte par mail en vue d'avertir l'utilisateur des nouveautés). En fonction des services offerts, ces outils (par exemple : http://www.spyonit.com/) sont tantôt gratuits, tantôt payants[31].

Les outils d'aide à la consultation de moteurs de recherche permettent de remplir tout ou partie des fonctions suivantes[32] :

- la consultation en parallèle de plusieurs moteurs de recherche (éventuellement sélectionnés par l'utilisateur) en vue de poser des questions récurrentes à intervalles réguliers;
- la compilation des résultats de la recherche;
- la consultation de chaque site donné en réponse et le chargement des pages de résultats (avec élimination de pages identiques);

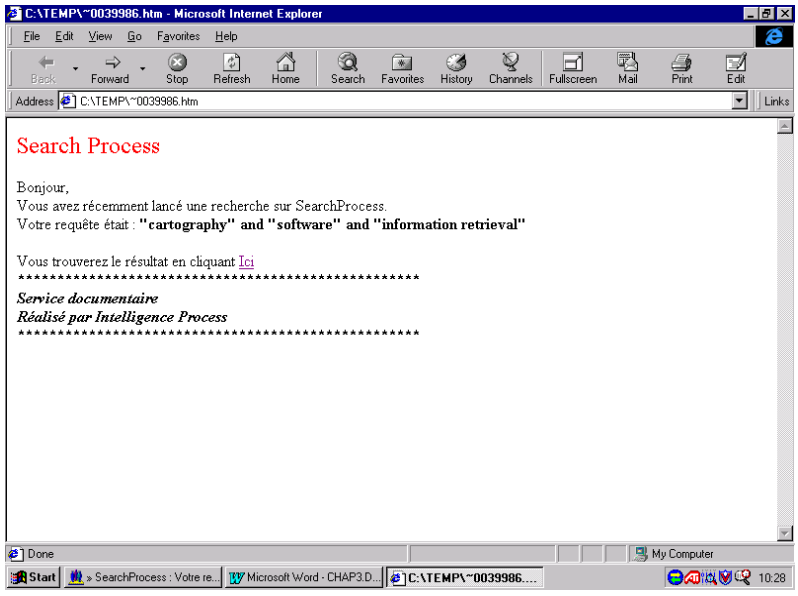
[31] ANDRIEU O., *Les espions du Web veillent pour vous...* (http://www.abondance.com/trucs-et-astuces/outils12.html)

[32] http://www.netmind.com - http://www.getupdated.com/ - http://spyonit.com - http://informant.dartmouth.edu

- la présentation synthétique de l'ensemble des résultats via une page HTML reprenant le titre et le résumé de chaque page;
- l'indexation en texte intégral de l'ensemble des pages retrouvées (une recherche locale est ainsi possible);
- la création d'un carnet d'adresse (*bookmarks*) et la mise en œuvre de techniques de rafraîchissement de ces *bookmark*s (selon une périodicité déterminée par l'utilisateur) [33]. Il existe maintenant des services permettant le stockage de ces carnets sur des serveurs externes, de telle sorte que l'utilisateur puisse y accéder depuis n'importe quelle plate-forme et depuis n'importe quel endroit[34] ;
- la mise à disposition de fonctions de cartographie, s'appuyant sur des techniques d'analyse linguistique et statistique. Les logiciels de cartographie documentaire[35] ont pour objet :
  - d'indiquer les liens physiques entre sites;
  - d'indiquer les liens sémantiques - sur base de dictionnaires sémantiques - entre "termes associés" (tels que "pétrole" et "guerre du Golfe", par exemple) issus d'un ou plusieurs documents ;
  - d'indiquer, sur base de dictionnaires sémantiques, les termes absents des dictionnaires et susceptibles d'être révélateurs de nouvelles tendances;
  - de permettre à l'utilisateur de visualiser les thèmes principaux des résultats d'une recherche et de distinguer leur importance relative.

A titre d'exemple, *Search Process* a proposé gratuitement ce type de services (figure 6), services qui maintenant sont payants, cette évolution étant symptomatique des tendances du Web [36] :

Figure 6. Processus de recherche sur *Search Process*



[33] On trouvera des liens vers la plupart de ces outils à l'adresse suivante : http://www.abondance.com/ressources/bookmarks.html

[34] Online bookmark-diensten. *Personal Computer Magazine*, juillet-août 2000, p. 82-83.

[35] CHANTAL E., Cartographier l'information. *Le Monde interactif*, 03/05/2000, p. 5.

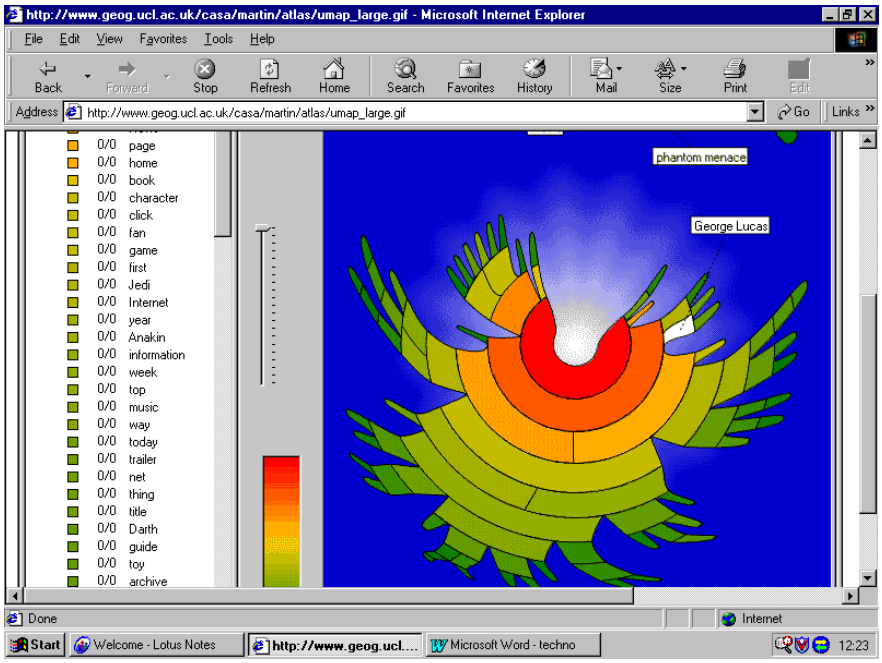
[36] Search Process, l'agent double. *Archimag*, avril 2000, n°133, p. 10.

Le problème essentiel que soulèvent ces outils (problème que l'on retrouve également dans le cas des "méta-index") réside dans la traduction d'une requête compréhensible par les moteurs de recherche, étant donné la syntaxe hétérogène des méthodes de recherche préconisée par chaque moteur.

2.5. Logiciels de cartographie documentaire

Les logiciels cartographiques[37] déjà évoqués au point précédent ont pour objet d'améliorer la navigation dans une collection de documents textuels et/ou d'interpréter les résultats d'une recherche documentaire[38] sur base des similarités thématiques entre documents en fonction de la présence/absence conjointe des mots-clés demandés. Au cœur du graphique (Figure 7) se trouvent les documents incluant tous les mots-clés. Plus on s'éloigne du cœur, moins il y a de mots-clés conjointement présents. Des familles de documents sont par ailleurs représentées sur cette base. Si l'utilisateur sélectionne une zone du schéma, il peut relancer la recherche sur les mots-clés correspondants.

Figure 7. Exemple de logiciel cartographique



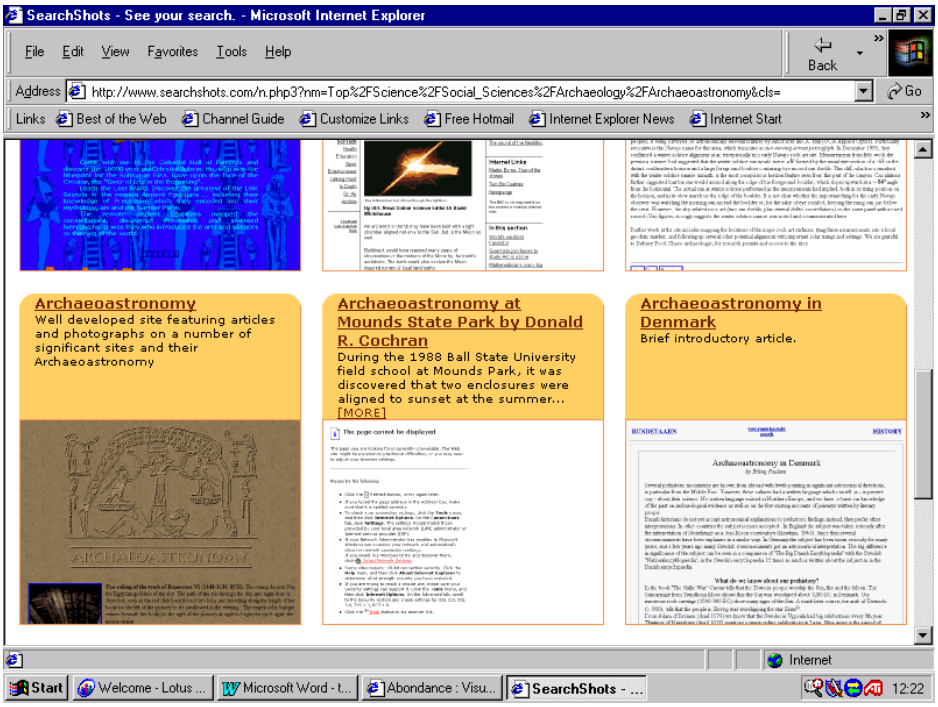
[37] Les copies d'écran qui suivent sont extraites du site "An Atlas of Cyberspaces" : http://www.geog.ucl.ac.uk/casa/martin/atlas/info_maps.html

[38] ROUSSINOV D., Information Forage through Adaptative Visualization. *1998 Symposium on Digital Libraries*. ACM : Pittsburgh, 1998, p. 303-304. HELZER B. et MILLER N., Four Critical Elements for designing Information Exploration Systems. *CHI98 Workshop on Information Exploration* (http://www.fxpal.com/CHI98IE). WANG BALDONADO M., Interfaces for Information Exploration : Seeing the Forest. *CHI98 Workshop on Information Exploration* (http://www.fxpal.com/CHI98IE).

2.6. Outils de visualisation des copies d'écran

Certains outils de recherche, tel que l'annuaire *SearchShots* (http://www.searchshots.com), permettent de visualiser les copies d'écran des résultats d'une recherche[39] (Figure 8). L'objectif consiste à donner à l'utilisateur un premier aperçu graphique des sites obtenus en réponse. Dans certains cas en effet, le *lay out* d'une *homepage* constitue un indicateur de la nature du contenu du site. A l'heure actuelle, seuls des annuaires, répertoriant un nombre restreint de sites, permettent d'offrir un tel service, qui ne serait pas tenable, en terme de performance, pour des outils manipulant des centaines de millions de pages.

Figure 8. Exemple d'outil de visualisation de copie d'écran

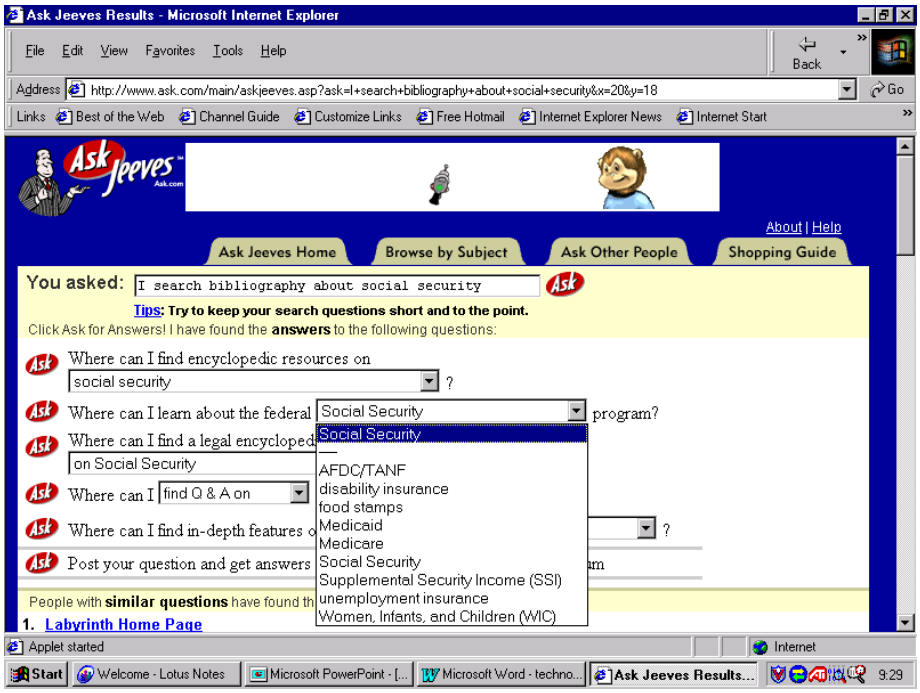


2.7. Outils de recherche en langage naturel

Certains outils de recherche, tel que *Ask Jeeves* (http://www.ask.com), offrent une technique de recherche en langage naturel (Figure 9). Le système repose sur une base de "questions-réponses" préexistantes, s'enrichissant progressivement en fonction des requêtes des utilisateurs : à la question de l'utilisateur, l'outil répond par plusieurs autres questions censément plus fines, dont les réponses sont accessibles *on line*. Ce type de technologie pose plusieurs difficultés : suivi dans le temps et dans l'espace de l'adéquation des questions aux réponses stockées et traitement linguistique du langage naturel. En effet, la question du sens des mots est extrêmement complexe, voire impossible à résoudre automatiquement, ainsi en cas de mots polysémiques appartenant à la même catégorie grammaticale (ex : un *café* désigne à la fois une boisson et un lieu) ou encore lorsque des phrases ayant une structure analogue ont des sens distincts ("*servir des boissons aux fruits*" ou "*servir des boissons aux invités*").

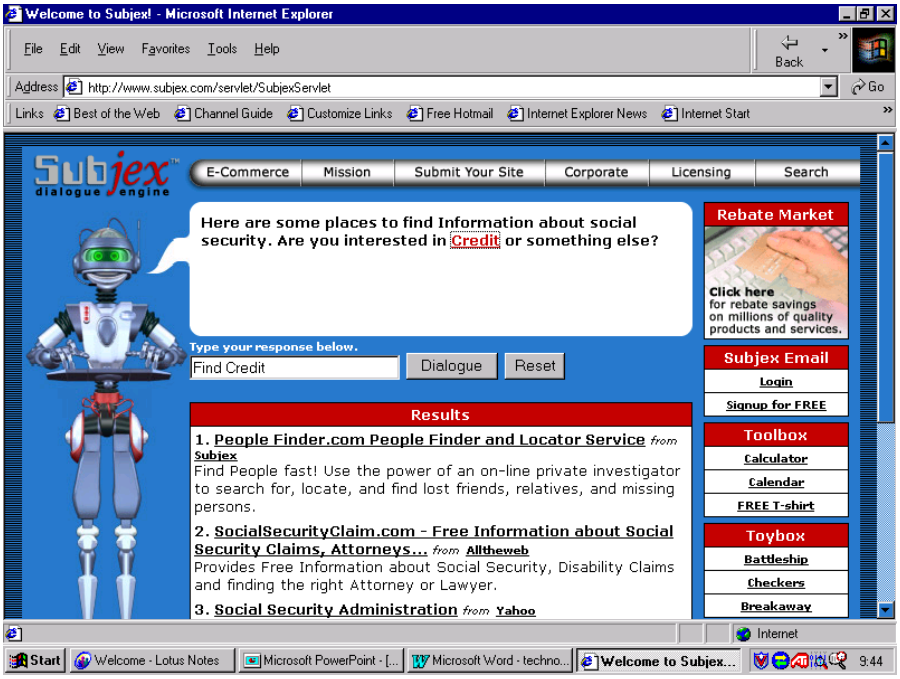
[39] ANDRIEU O., *Visualisez des copies d'écran de vos résultats de recherche* (http://www.abondance.com/trucs-et-astuces/outils18.html).

Figure 9. Exemple d'interrogation en langage naturel



Certains systèmes ("*dialog engines*") proposent d'aller plus loin encore (Figure 10) et posent des questions complémentaires aux internautes dans le but d'affiner la réponse (par exemple : *Subjex* http://www.subjex.com/).

Figure 10. Exemple de "dialog engine"

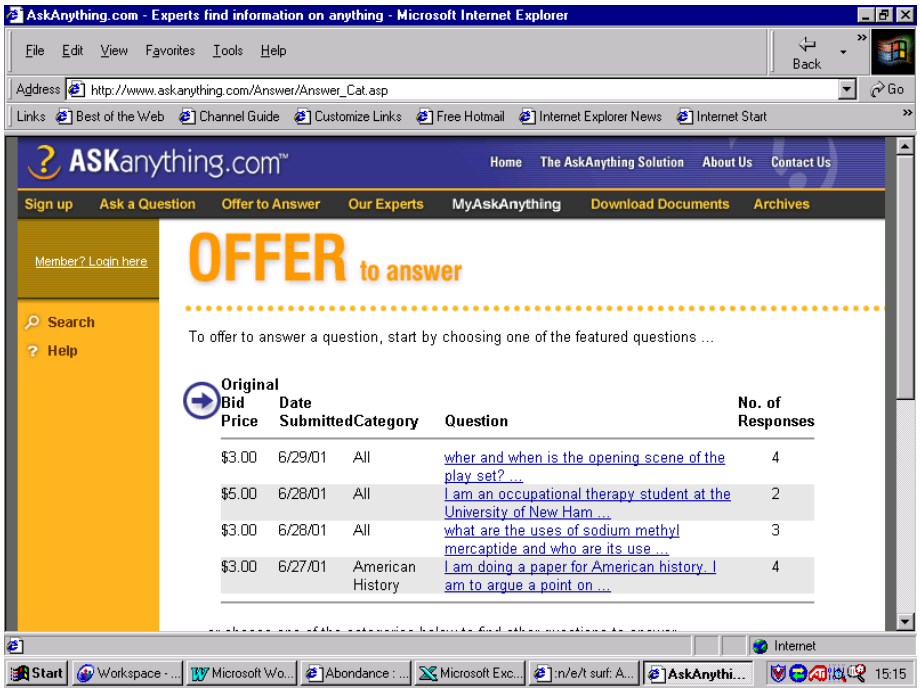


2.8. Retour à l'expertise humaine

En vue de compléter les outils "automatisés", des éditeurs de recherche "humains" (*WebWizards*) se développent de plus en plus[40]. Des experts humains répondent, dans un délai donné, aux questions posées. Certains services (de type "call centers"), tel que *Webhelp*, offrent gratuitement des réponses rapides et pertinentes à des questions simples et générales (leur mode de financement reposant uniquement sur les publicités).

D'autres services, payants et spécialisés dans un domaine particulier (Figure 11), offrent des réponses approfondies à une question pointue : les experts sont payés à la question[41] et effectuent une recherche plus ou moins poussée en fonction du prix que l'utilisateur déclare vouloir payer pour sa question. On trouvera une liste quasi exhaustive de ces sites de recherche "humains" à l'adresse suivante : http://www.netsurf.ch/askexperts.html

Figure 11. Exemple d'éditeur humain payant



[40] "*The search engines are now recognizing the limits of the massive quantity and lack of quality of information on the web. Hence, they are preparing a number of strategies for adding editorial context to the data. Infoseek and Excite have "channels" or "guides" where editors (yes, real people) identify quality sites in particular aress and guide the user to those, rather than just let them type in a word and search ... Now however, most searchers realize it is completely irrelevant if the search engine retrieves 100.000 or 1 million records. This is completely unwieldy....* » THOMAS A., Tendance : déferlante de moteurs humains. *Archimag*, juin 2000, n° 135, p. 30. Par exemple : http://www.woonoz.com ou http://www.equesto.com ANDRIEU O., *Posez vos questions à des être humains!* (http://www.abondance.com/trucs-et-astuces/recherche41.html).

[41] Certains systèmes mettent la question aux enchères avec une limite dans le temps et des experts y répondent. C'est celui qui pose la question qui fixe le prix et le délai. Par ailleurs, dans le cas de *AskAnything.com*, tout internaute peut également répondre aux questions posées pour gagner de l'argent.

3. Quelques recommandations méthodologiques

Nous avons rappelé plus haut (point 1) plusieurs recommandations de base en vue d'atténuer les conséquences de la synonymie et de la polysémie. Nous présentons ici un ensemble de recommandations méthodologiques spécifiques à l'interrogation sur Internet.

3.1. Interprétation des opérateurs de recherche sur les moteurs d'indexation

Il est important de noter que, d'un moteur de recherche à l'autre, la syntaxe des opérateurs de recherche est loin d'être homogène. Dans la mesure où cette syntaxe est susceptible d'évoluer, il est fondamental de consulter l'aide en ligne des outils de recherche. A titre illustratif, les remarques qui suivent, extraites du site Web *Abondance*[42], concernent la plupart des systèmes de recherche sur Internet :

- **l'ordre des mots** est fondamental dans certains cas[43], les mots se situant en premier lieu ayant plus de poids dans le classement des réponses mais n'est pas pris en compte dans d'autres systèmes[44];
- **le recours au signe "+"** est fondamental si l'on souhaite marquer un "AND", deux mots reliés par un espace étant souvent considérés comme étant reliés par un "OU".
- **le signe "-"** indique le "SAUF";
- **le recours à la troncature** est implicite sur certains outils, tel que *Yahoo!* (une recherche sur "capi" donnera "capital", "capitaine", etc.) mais doit être explicitement indiqué dans d'autres systèmes, tel que *Altavista* (capi*). La troncature gauche n'est, pour des raisons de performance, quasi jamais permise.
- **le recours aux guillemets** est fondamental pour toutes les recherches sur une expression, ce qui revient à recourir à un opérateur d'adjacence entre chaque terme de l'expression (dans ce cas, il n'est plus possible d'utiliser des opérateurs de troncature à l'intérieur de l'expression);
- **la casse des lettres** (recours aux majuscules ou aux minuscules) est tantôt prise en compte (sur *Altavista*, une recherche sur "ibm" donnera lieu à une recherche sur ce terme quelle que soit la casse de chacune des lettres dans les pages Web mais dès qu'une des lettres est en majuscule dans la requête (par exemple, recherche sur "Ibm"), le système sera sensible à la casse). La casse est tantôt ignorée (sur *Yahoo!* par exemple).
- **l'accentuation** est diversement prise en compte :
  - elle n'est pas prise en compte sur *Yahoo!*, par exemple;
  - toutes les occurrences du mot sont prises en considération si celui-ci n'est pas accentué dans la requête (sur *Altavista*, par exemple);
  - la graphie exacte est prise en compte (sur Lycos, par exemple).

[42] http://www.abondance.com/trucs-et-astuces/recherche.html

[43] Altavista, Infoseek, Northern Light, Google, Ecila.

[44] HotBot, Excite, Lycos, WebCrawler, Alltheweb, Lokace, Yahoo!, Looksmart, Open Directory, Nomade.

3.2. Eléments syntaxiques propres à Internet

Par ailleurs, il sera utile de prendre en considération les niveaux de recherche proposés et notamment :

- l'adresse (URL) : par exemple, la requête url:bookmark (ou « *link* », « *hostlist* », ...) associée aux termes de la recherche a pour objet de sélectionner les sites qui contiendraient une collection de liens consacrés à l'objet recherché.
- au sein de l'URL, la sélection d'un domaine spécifique permet de cibler le type de site recherché (exemples de noms US : .ac, académique, .com, commercial, .edu, universités, écoles, .org, organismes publics, .gov, organisations gouvernementales, .int, organisations internationales, .net, administrateurs de réseau, .mil, défense , etc.).

L'indice de popularité peut également être utile pour les recherches documentaires sur le Web[45], en vue de rechercher des sites "cousins" ou "complémentaires". L'indice de popularité doit être interprété en fonction du caractère plus ou moins récent du site examiné (plus il est récent moins d'autres sites auront eu le temps de créer un pointeur vers ce dernier).

3.3. Recours à des mots-clés spécifiques

En relation avec les mots-clés que l'utilisateur aura placés dans l'équation de recherche, les pages de résultats de certains outils peuvent afficher un lien avec :

- le "***Real Name***" correspondant[46] : ces "mots-clés Internet" sont achetés par un particulier ou par une entreprise et reliés à un site officiel. Ces mots-clés ne sont attribués par la société "*RealNames*" (http://www.realnames.com/) qu'après vérification de l'adéquation au site correspondant : une cohérence sémantique est donc garantie. Bien entendu, un nom n'est attribué qu'à un seul interlocuteur. Plusieurs outils (*Altavista*, *Google*, ...) ont recours à cette technologie (via la base de données centrale de ces "mots-clés Internet") qui permet à l'internaute d'accéder directement au site Web d'un organisme potentiellement lié à l'objet de sa requête.
- les "**AdWords**" correspondants (essentiellement dans le cas de *Google*) : la pratique des AdWords permet d'acheter pour certains mots-clés la possibilité de publier des encarts publicitaires. Contrairement aux "RealNames", les "AdWords" ne font pas l'objet d'une vérification a priori. Dans certains cas, cette pratique s'apparente à une "publicité self service", sans contrôle systématique et peut donner le jour à certains effets pervers, tels que la diffusion de textes en faveur de l'église de scientologie, lorsque l'utilisateur tape le mot "*drogue*"[47].

---

[45] "*Imaginons que vous ayez un site web, répondant à l'adresse www.votresite.com et que vos deux principaux concurrents, nommés Concur1 et Concur2, disposent des sites web www.concur1.com et www.concur2.com. Pour trouver des sites à qui proposer un partenariat d'échange de liens réciproque, tapez sur www.Altavista.com la requête :*

+link:concur1.com +link:concur2.com -link:votresite.com

Que dit cette ligne de recherche ? Elle va fournir comme résultat les pages qui ont mis en place un lien vers les sites de vos deux principaux concurrents, mais pas vers le vôtre. Vous pensez que ce n'est pas normal. Comme cette page "pointe" vers les deux sociétés Concur1 et Concur2, elle ne fait pas partie d'un site concurrent. Il vous reste alors à aller sur ces pages listées par Altavista, à contacter le webmestre qui en est responsable, et à lui signaler l'existence de votre site en lui proposant poliment de rajouter un lien vers celui-ci. " ANDRIEU O., *Comment se servir de l'indice de popularité dans vos recherches ?* (http://www.abondance.com/trucs-et-astuces/recherche16.html).

[46] ANDRIEU O., *Interview de Jean-Marie Hulot* (décembre 1999), http://abondace.com/docs/interview_realnames.html.

[47] GONZAGUE A. et STEIN S., Quand Google fait de la pub aux scientologues. 06/04/2001 (http://www.traansfert.net/fr/cyber_societe/

4. Conclusions

Nous avons évoqué les difficultés de la recherche sur Internet, difficultés liées à l'étendue et à la mouvance du réseau, au phénomène massif de polysémie et de synonymie, à la pratique du spam et à l'évolution rapide des index : *"face à ce foisonnement, on a vraiment l'impression de chercher une aiguille dans une meule de foin, avec la complication suivante : la meule de foin est à géométrie variable et elle ne cesse de grossir de façon exponentielle*".[48]

Afin de faire face à ces difficultés, nous avons présenté brièvement les principaux outils de recherche existant à l'heure actuelle et évoqué, le cas échéant, plusieurs recommandations méthodologiques en vue de les exploiter au mieux : accès direct à quelques sources de base, annuaires et moteurs, outils de recherche par méta-index, logiciels de veille documentaire, logiciels de cartographie documentaire, outils de visualisation de copies d'écran, outils de recherche en langage naturel et recours aux éditeurs humains. A cela, il convenait d'ajouter plusieurs recommandations méthodologiques quant à la formulation des équations de recherche sur Internet : en effet, dans le cadre du Web, l'application de la logique booléenne connaît plusieurs extensions syntaxiques et sémantiques.

Voici un tableau synthétique reprenant l'usage recommandé de chacun des outils évoqués.

Type d'outil :	**Type de recherche préconisée :**
Sources de type encyclopédique, actualités, archives de FAQ ou de forums de discussions	Recherche (éventuellement rétrospective) sur un sujet précis.
Annuaires	Recherche sur un thème général.
Moteurs d'indexation et de recherche	Recherche sur un sujet précis dont la formulation peut nécessiter une requête booléenne complexe combinant plusieurs concepts
Méta-moteurs de recherche	Recherche « rapide » sur plusieurs moteurs donnant un aperçu synthétique des réponses à une question. Le recours à ces outils est conseillé au début d'une recherche pour se faire une première idée de l'état d'un domaine.
Logiciels de veille documentaire	Recherche récurrente sur un thème donné avec mise en place éventuelle d'un système d'alerte.
Logiciels de cartographie documentaire	En complément des outils précédents, recherche permettant de visualiser la répartition des réponses par mots-clés sous forme graphique.
Outils de visualisation des copies d'écran	Recherche permettant de visualiser la page d'accueil des sites obtenus en réponse.
Outils de recherche en langage naturel	Recherche destinée aux utilisateurs qui ne souhaitent pas recourir à la logique

[48]LELOUP C., *Moteurs d'indexation et de recherche. Environnements client-serveur, Internet et Intranet.* Paris : Eyrolles, 1998, p. 130.

	booléenne et veulent formuler leurs questions en langage naturel. Dans certains cas, les résultats sont « approximatifs »
Outils de recherche reposant sur l'expertise humaine	Recherche destinée aux utilisateurs qui ne souhaitent pas recourir à la logique booléenne en recourant à l'aide d'experts de la recherche sur Internet (dans des domaines spécialisés).

Très souvent, la mise en forme physique de l'information a eu une influence déterminante sur l'appréhension de son contenu. Ainsi, avant l'imprimerie, l'innovation la plus importante a résidé dans la modification de la présentation physique de l'ouvrage manuscrit. Le passage du rouleau de l'antiquité au codex (assemblage de feuilles ou de pages cousues) se généralise en effet à la fin du IVème siècle et est facilité par l'abandon du papyrus (qui ne peut être aisément cousu) au profit du parchemin (qui ne peut être aisément roulé)[49]. C'est ainsi qu'ont pu naître les index et classements permettant la "recherche documentaire" et, en corollaire, une lecture plurielle des écrits. Au XIVème siècle, on observe ainsi l'apparition des professionnels faisant de l'index leur métier (comme en témoignent les archives de la cour pontificale d'Avignon, sous Jean XXII)[50].

Dans cet ordre d'idées, ce qu'on appelle "*l'information overloading*", la surcharge d'information, se complexifie dans le cadre d'Internet, avec la nature même de l'hypertexte qui assure la mise en forme de l'information sur le Web. Le livre, au sens traditionnel, est téléologique et fini, s'inscrivant autrefois entre l'*incipit* et le *finit* (ou *explicit*), comme en témoigne l'expression ancienne "*Here ends the romance*". Par contre, dans l'environnement hypertextuel, l'histoire ne s'achève jamais : les textes sont sans fin et récursifs. En soi, un lien hypertextuel est en quelque sorte un outil de spoliation. Le *spamming* pourrait être vu comme une dérivation "naturelle" de cette fonction de spoliation propre à l'hypertexte.

Au niveau des outils de recherche sur Internet, plusieurs tendances se dessinent à l'heure actuelle en vue de lutter contre le *spamming* et/ou de faciliter la recherche : qu'il s'agisse du recours accru à l'expertise humaine ou encore du positionnement et du référencement payants. Face à cela, il est probable que les moteurs et annuaires du futur seront de plus en plus spécialisés : les uns, plus spécifiques, pourraient continuer d'offrir un accès gratuit à des informations dont le caractère commercial n'est pas déterminant (sites universitaires, culturels, administratifs, ...) et les autres, commerciaux, assureraient la diffusion payante de l'information.

[49] CHARTIER R. et MARTIN H.-J., éds, *Histoire de l'édition française. Tome I. Le livre conquérant. du Moyen-âge au milieu du XVIIème siècle.* Paris : Fayard- Cercle de la Librairie, 1989.

[50] ROUSE M. A ET ROUSE R. H., *La naissance des index.* In CHARTIER R. et MARTIN H.-J., éds, *Histoire de l'édition française. Tome I. Le livre conquérant. Du Moyen-âge au milieu du VVIIème siècle.* Paris : Fayard- Cercle de la Librairie, 1989, p. 95-108.