enssib
école nationale supérieure des sciences
de l'information et des bibliothèques

MÉMOIRE DE DEA
Sciences de l'Information et de
la Communication

option : Systèmes d'information documentaire

LE FILTRAGE D'INFORMATION SUR INTERNET : CONVERGENCES ET DIVERGENCES ENTRE OUTILS DE RECHERCHE

Cynthia DELISLE

Sous la direction de : Omar LAROUK
(Université de Bourgogne, IUT de Dijon)

Septembre 1999

Université Lumière
Lyon 2

Ecole Nationale Supérieure
des Sciences de l'Information
et des Bibliothèques

Université Jean Moulin
Lyon 3

LE FILTRAGE D'INFORMATION SUR INTERNET : CONVERGENCES ET DIVERGENCES ENTRE OUTILS DE RECHERCHE

Résumé :

Le repérage de l'information sur Internet est actuellement une tâche ardue, dont le succès (ou l'insuccès) est tributaire en bonne partie de l'efficacité des outils de recherche. Nous présentons les caractéristiques des diverses catégories d'outils de même que les résultats de quelques tests effectués en ligne. Ces derniers permettent de mieux appréhender les nuances de traitement entre outils et leurs implications pour le processus de recherche d'information.

Descripteurs : système d'information ; repérage de l'information ; filtrage de l'information ; Internet ; Word Wide Web ; outil de recherche ; annuaire ; moteur ; métamoteur ; agent intelligent.

Abstract :

Information retrieval on the Net remains a difficult task. Whether it succeeds or not depends greatly on the search engines' effectiveness. We review the categories and main features of WWW search engines and present the results of some online tests. These results allow us to conceive of the subtle differences that exist between search engines regarding information processing. The resulting implications in the information retrieval process are also discussed.

Keywords : information system ; information retrieval ; information filtering ; Internet ; World Wide Web ; WWW search engine ; subject directory search engine ; keyword search engine ; meta search engine ; agent.

Remerciements

Ma gratitude va à monsieur Omar Larouk, mon directeur de recherche, ainsi qu'à tous mes proches et amis du Québec, pour leur affectueuse sollicitude et leur soutien moral «virtuels».

Un merci tout spécial à R., qui se reconnaîtra.

TABLE DES MATIÈRES

«It is surprising what some search engines find and what others do not find.»

(Alan Poulter, «The design of World Wide Web search engines: a critical review», 1997)

Introduction

Internet peut être défini comme un réseau informatique global où des milliers d'ordinateurs, identifiés de manière unique par un numéro IP et un nom de domaine, communiquent entre eux via le protocole TCP/IP[1]. Émanant à l'origine du milieu académique américain, ce «réseau des réseaux» n'a cessé de croître depuis sa naissance, en 1969, et plus particulièrement depuis l'arrivée du World Wide Web, en 1993. La simplicité d'utilisation et le caractère attractif de ce dernier ont favorisé, en effet, une augmentation radicale du nombre des internautes, de même qu'une diversification de ceux-ci : les scientifiques et universitaires des origines ont été rejoints puis massivement dépassés par les entreprises et le grand public, séduits par le potentiel commercial et les pages bariolées du nouveau médium.

Le Réseau déploie maintenant ses ramifications aux quatre coins du monde. On assiste à l'explosion de nouveaux marchés qui menacent de concurrencer sérieusement les Nord-Américains, utilisateurs de la première heure : l'Europe (notamment l'Europe de l'Est), l'Asie, l'Amérique du Sud. La quantité de documents et de services disponibles augmente de manière incontrôlée et, la mondialisation entraînant inévitablement une diversification des langues, des sujets et des types de ressources, on peut affirmer sans risque d'exagération que tout un chacun, ou presque, est désormais susceptible de trouver son bonheur sur Internet.

Le Réseau, en effet, constitue à la fois un mode de communication attrayant (par exemple, en ce qui a trait au courriel, au bavardage en direct ou aux transactions financières sécurisées) et une exceptionnelle mine de renseignements accessibles simultanément à plusieurs usagers, en tout temps et pour un coût relativement minime. Il a notamment hérité de ses origines universitaires une grande richesse en «littérature grise», c'est-à-dire en textes académiques à diffusion restreinte, généralement difficiles à obtenir ou à repérer à partir de sources documentaires

[1] TCP est le sigle de *Transmission Control Protocol*, IP celui de *Internet Protocol*.

traditionnelles. Internet s'affirme également, de plus en plus, comme le média de prédilection pour suivre l'évolution de l'actualité mondiale, nationale ou régionale, puisque l'information peut y être diffusée en permanence de façon quasi simultanée et ne connaît ni restrictions d'espace (comme dans la presse écrite) ni limites de temps (comme à la télévision ou à la radio).

Mais, pour exploiter pleinement toutes ces richesses, encore faut-il en connaître l'existence. Et c'est là que le bât blesse, puisque le repérage des ressources constitue précisément le talon d'Achille d'Internet, comme le souligne J.-P. Lardy :

> *Des ressources informationnelles impressionnantes sont disponibles sous des formes très variées et peu structurées, dispersées sur des milliers de serveurs : ce qui les rend très difficiles à repérer, identifier et évaluer. Il est quasiment impossible à un utilisateur d'avoir une idée précise de ce à quoi il peut accéder sur le réseau C'est une situation déroutante par rapport aux banques de données commerciales dont une pratique régulière permet d'appréhender le contenu. [Lardy 1996]*

En l'absence de toute gestion centralisée des ressources, le Web confronte l'usager à une conjoncture paradoxale où l'information se révèle à la fois directement accessible et fort difficile à atteindre. Cette situation a engendré l'apparition de sites que l'on pourrait qualifier d'*aiguilleurs du Web* : les outils de recherche.

Le premier outil de recherche développé pour Internet fut ARCHIE. Basé à l'Université McGill (Montréal), il permettait des fouilles par mots clés dans une base de données de noms de fichiers disponibles par FTP[2]. Depuis, les outils de recherche ne cessent de se multiplier sur Internet. Dans le contexte du Web[3], ils se présentent comme des services de repérage constitués d'une ou plusieurs base(s) de

[2] Pour *File Transfer Protocol*.
[3] Dans la mesure où seule une petite minorité des ressources d'Internet est inaccessible à partir du Web, le terme *outil de recherche du Web* est désormais synonyme, pour ainsi dire, du terme *outil de recherche d'Internet*.

données décrivant essentiellement des ressources WWW, d'un logiciel de recherche et d'une interface usager également accessible via le Web [Poulter 1997].

Ces outils peuvent être répartis en quelques catégories de base ; ils varient néanmoins énormément entre eux sur de nombreux points de détail. Il est souvent ardu d'en déterminer le fonctionnement exact puisque leurs concepteurs (qu'ils soient du milieu académique ou industriel) s'avèrent d'ordinaire fort peu diserts sur le sujet, soucieux de protéger des procédés de nature propriétaire.

Pourtant, cette connaissance approfondie des divers systèmes de repérage de l'information (SRI) sur Internet et de leur comportement réel face à une requête est absolument indispensable si l'on désire éviter les biais que les apparences et les idées préconçues introduisent trop fréquemment – souvent à l'insu même de l'internaute – dans les résultats d'une recherche. Elle permet également de faciliter autant que faire se peut un processus qui demeure intrinsèquement hasardeux, comme le fait remarquer S. Feldman :

> *Searching is a language game. Find just the right combination of words and you have the key to the black box of answers that we call a database. Guess wrong, and the box remains mum, or worse, it spews back non-sense. [Feldman 1999]*

C'est dans cette optique que s'inscrit notre mémoire. Nous nous proposons, dans un premier temps, de passer en revue les notions générales relatives au processus de repérage de l'information, puis, en second lieu, d'introduire les caractéristiques plus spécifiques de la recherche d'information sur Internet, notamment en résumant la typologie des outils actuellement disponibles dédiés à cette fin. Suite à cette mise en contexte théorique, nous présenterons et discuterons les résultats que nous avons obtenus en procédant à plusieurs requêtes exemplaires destinées à analyser le comportement réel des SRI sur Internet face à certaines thématiques d'importance (couverture des ressources, traitement des caractères accentués, efficacité respective des modes de repérage de type booléen ou statistique et des requêtes en langue naturelle, etc.). Nous terminerons par quelques considérations sur les implications pour l'usager des différences – souvent subtiles et inattendues – constatées entre les

divers outils de recherche testés au niveau du fonctionnement et du traitement de l'information. Des suggestions d'investigations futures seront également formulées.

Première partie

le repérage de l'information

1. Les principaux types de systèmes de repérage de l'information (SRI)

1.1. Les systèmes booléens ou traditionnels

Comme leur nom l'indique, ces systèmes se basent sur la logique développée par le mathématicien britannique George Boole. Ils utilisent des opérateurs pour combiner des termes de recherche entre eux, comme s'il s'agissait d'énoncés mathématiques. DIALOG et LEXIS-NEXIS sont des exemples.

Ces systèmes appréhendent un texte comme une suite aléatoire de mots délimités entre eux par des signes de ponctuation, des espaces typographiques ou d'autres caractères tels $%&-/#_~. Ils apparient requêtes et documents via le principe de concordance de modèle (*pattern matching*), et plus particulièrement la recherche de concordances exactes (*exact matches*). Lorsque la requête de l'usager est confrontée au contenu de la base de données, les entrées qui apparaissent sur la liste de résultats sont celles qui contiennent la ou les chaîne(s) recherchée(s), soit dans le texte même du document, soit dans d'autres champs de l'enregistrement (par exemple, les balises META d'un fichier HTML[4] ou encore, s'il y a lieu, les rubriques de classification). Les résultats ne font l'objet d'aucun tri.

[4] Les balises META, comme leur nom le suggère, sont des «informations sur l'information» : elles fournissent aux outils de recherche des renseignements spécifiques, par exemple un résumé ou une suite de mots clés relatifs au contenu d'une page Web. Ces codes appartiennent au langage HTML et ne sont pas visibles pour l'utilisateur. Ils s'inspirent du travail effectué pour les documents en sciences humaines dans le cadre de la TEI (*Text* Encoding Initiative), qui visait à spécifier des descripteurs de contenu à l'usage des auteurs et des éditeurs pour différents types de documents.

Les principaux opérateurs utilisés sont les suivants :

- **Les opérateurs dits booléens**

  - L'opérateur ET

Il permet de rendre la présence de mots obligatoire. Il est également symbolisé par son équivalent anglais AND ou par l'espace lorsqu'il est pris par défaut.
Exemple : *commerce ET électronique* repérera toutes les entrées où ces deux mots figurent.

  - L'opérateur OU

Il permet de rendre la présence de mots optionnelle. Il est également symbolisé par son équivalent anglais OR ou par l'espace lorsqu'il est pris par défaut.
Exemple : *commerce OU électronique* repérera toutes les entrées qui comprennent au minimum un de ces deux mots.

  - L'opérateur SAUF

Il permet d'exclure la présence de mots. Il est également symbolisé par ses équivalents anglais NOT, BUT NOT ou AND NOT.
Exemple : *commerce SAUF électronique* repérera toutes les entrées où figure le mot *commerce* mais sans qu'y apparaisse le terme *électronique*.

- Les parenthèses ( )

Elles permettent de limiter la portée des opérateurs booléens et/ou d'introduire un ordre de priorité entre les différentes parties d'une requête.

Exemple : *(commerce OU paiement) ET électronique* repérera les entrées qui contiennent à la fois *électronique* et soit *commerce* soit *paiement* soit ces deux termes.

- La troncature

Elle consiste à recourir à l'emploi de masques (*jokers* ou *wild cards*). Généralement symbolisée par * ou ? ou $, la troncature permet d'effectuer des recherches sur des parties de mots. Notons qu'elle est moins flexible dans le contexte de la recherche d'information sur le Web qu'en ce qui a trait aux logiciels documentaires traditionnels (impossibilité de l'appliquer en début de mot[5], nécessité fréquente de saisir un nombre minimum de caractères, etc.). Elle est toutefois intéressante en ce qu'elle permet de faire des recherches sur des mots de même famille et sur les variations de genre et de nombre.

Exemples : *biblio** repérera *bibliothèque*, *bibliothèques*, *bibliothécaire*, *bibliophile*, etc. La troncature peut aussi s'utiliser à l'intérieur d'un mot, pour remplacer un ou plusieurs caractère(s) : *coll$sion* repérera *collision* et *collusion*.

- La recherche de locutions[6]

Elle fonctionne habituellement à l'aide des guillemets " " et permet la recherche exacte d'une séquence ordonnée de mots adjacents.

Exemple : *"commerce électronique"* repérera toutes les entrées où ces deux mots figurent l'un à côté de l'autre et dans cet ordre.

---

[5] Il existe toutefois quelques exceptions, entre autres l'outil de recherche HOTBOT.

[6] Pour l'ensemble de ce travail, nous emploierons le terme *locution* pour désigner les groupes de mots qui fonctionnent comme un mot simple et dont le sens global diffère souvent sensiblement du sens initial des composantes (par exemple, *pot de vin*). On pourrait également parler de *syntagmes*, de *synapsies*, de *mots composés*, de *polytermes*, etc.

- **L'opérateur de proximité**

La recherche sur la proximité est considérée comme une extension du modèle booléen. L'opérateur de proximité permet de rechercher des entrées où les mots désirés apparaissent à l'intérieur d'une «fenêtre» de voisinage dont l'ampleur varie selon les outils (généralement entre 10 et 100 mots, parfois beaucoup plus). Les formulations les plus habituelles sont anglophones : NEAR ou FOLLOWED BY (dans ce dernier cas, on tient également compte de la linéarité, c'est-à-dire de l'ordre d'apparition des termes). Pour rechercher des termes côte à côte (un peu comme une recherche de locution, mais sans souci de linéarité), on emploie parfois également un opérateur de proximité spécifique, dit *opérateur d'adjacence*. Il est généralement symbolisé par ADJ.

Exemples : *commerce NEAR électronique* repérera les entrées où ces deux termes figurent près l'un de l'autre. *Commerce FOLLOWED BY électronique* exigera, de plus, que l'ordre de saisie des mots soit respecté. *Commerce ADJ électronique*, pour sa part, recherchera les entrées où ces deux termes apparaissent immédiatement l'un à côté de l'autre, peu importe l'ordre d'apparition.

1.2. Les systèmes statistiques ou probabilistes

Les systèmes statistiques ou probabilistes sont une application des recherches menées aux États-Unis par G. Salton à partir du milieu des années 1960[7]. Ils vont au-delà des approches booléennes par mots clés, dont ils tentent d'améliorer les performances. Leur but est de permettre le repérage des documents qui s'avèrent similaires à un ensemble de mots. Grâce à des technologies algorithmiques qui exploitent probabilités et statistiques inférentielles, ils repèrent et trient les réponses selon leur degré de correspondance avec la requête de l'usager, c'est-à-dire selon leur chance d'être jugées pertinentes par ce dernier. Ce type de recherche fournit donc non seulement les concordances exactes (*exact matches*) d'une requête, mais aussi celles qui s'en rapprochent (*close matches*). La plupart des outils de recherche

[7] Voir notamment [Salton & McGill 1983].

sur Internet dont il sera question dans la suite de ce travail soit relèvent de cette catégorie, soit sont des systèmes booléens augmentés de ce type de capacités statistiques, en particulier de fonctions d'évaluation de pertinence (*relevancy ranking*).

De manière très schématique, les systèmes statistiques basent leur fonctionnement sur le dénombrement des occurrences totales de chaque terme (sauf, éventuellement, les mots vides) dans un document, de même que dans l'ensemble de la base de données de l'outil. Toutefois, ceci ne veut pas dire nécessairement que les outils statistiques se bornent à compter les mots de la requête présents dans la base de données et que le document avec le plus d'occurrences «gagne», car une tactique aussi simpliste tendrait à favoriser exagérément les documents de taille importante. Un mécanisme supplémentaire d'assignation de poids différenciés aux divers mots existe donc généralement, la formule la plus fréquente consistant à affecter à ces derniers un poids inversement proportionnel à leur fréquence totale d'apparition dans la base de données : un mot relativement «rare» est ainsi doté d'un poids plus considérable qu'un mot très commun. Le principe sous-jacent est que le contenu informationnel d'un terme est inversement proportionnel à sa fréquence d'apparition : autrement dit, plus un mot figure souvent dans un texte ou un ensemble de textes, moins il est discriminant et véhicule en soi d'information.

D'autres facteurs peuvent être considérés dans le procédé de pondération des résultats, par exemple la *densité*, qui tient compte de la fréquence d'apparition d'un mot dans un document et de la taille de ce dernier. Une méthode reliée consiste à appliquer une courbe de pondération déclinante où la première occurrence d'un mot dans un document reçoit plus de poids que la seconde, elle-même supérieure à la troisième, etc. En ce qui concerne l'évaluation des documents, les critères suivants sont également susceptibles d'être utilisés :

- la proximité des mots clés entre eux ;
- l'emplacement des mots clés dans le document.

Depuis peu, dans le cas particulier des SRI sur Internet, on recourt en outre aux indicateurs suivants :

- le nombre de liens dans la base de données pointant vers une page (un peu à la manière d'une étude scientométrique de citations) ou la présence d'un lien en provenance d'un site «important» ;
- le nombre de fois qu'une page est visitée à partir d'une liste de résultats[8] ;
- pour les outils qui incorporent un annuaire, la présence dans l'annuaire de la page concernée[9].

Par ailleurs, pour tous les systèmes statistiques, la présence de l'ensemble des mots clés de la requête dans un document assure toujours à ce dernier l'émergence en tête de liste des résultats : ainsi, pour une requête comportant à la fois *bananes* et *pommes*, un document avec une occurrence de *bananes* et une occurrence de *pommes* précédera immanquablement un document avec seulement trois occurrences de *bananes*. (La pondération selon le rang d'apparition expliquée plus haut est l'une des techniques employées pour garantir ce résultat.)

L'approche développée par les systèmes statistiques rend possible l'identification automatique de «termes reliés» aux mots d'une requête (*related terms*), c'est-à-dire de termes qui co-occurrent dans la base de données avec les termes de la requête et contribuent ainsi à définir efficacement un concept. Par exemple, une recherche sur le SIDA à l'aide de cette fonctionnalité peut permettre de repérer des documents pertinents mais où cette expression n'est jamais explicitement mentionnée (parce qu'on y parle uniquement de VIH, etc.). Proposant des «termes reliés» pour une «recherche conceptuelle», ces outils peuvent parfois donner l'impression de réaliser une analyse linguistique, mais il ne faut donc pas perdre de vue qu'en réalité, ils se bornent à afficher une liste des termes qui apparaissent fréquemment dans les documents du corpus où figurent les termes de la requête. Par ailleurs, ce type de

[8] Sur Internet, DIRECT HIT est un système de ce genre. Il mesure quelles sont les pages visitées par les usagers à partir d'une liste de résultats de recherche : les pages réellement visitées obtiennent une augmentation de leur cote de pertinence ; celles qui sont laissées de côté voient la leur baisser. DIRECT HIT est notamment incorporé à HOTBOT.

[9] Cette stratégie est employée, par exemple, par INFOSEEK. Son utilité apparaîtra plus évidente à la lecture de la seconde partie de ce travail.

systèmes de repérage présente également l'avantage de permettre d'utiliser un document entier en tant que requête (c'est la fonction *More like this* ou *Plus de réponses comme celle-ci* que l'on trouve sur de nombreux outils de recherche sur Internet). Mais, là encore, une recherche de «documents similaires» à partir d'un document jugé pertinent s'appuie uniquement sur des comparaisons statistiques – et non sur la mise en correspondance de mots (*word matches*) ni, *a fortiori*, sur des analyses sémantiques ou pragmatiques.

Ajoutons enfin qu'avec ces systèmes, les mots de la requête peuvent être indiqués tels quels, sans être nécessairement joints par de quelconques opérateurs ou modificateurs. Du reste, le tri de pertinence redéfinit de fait les opérateurs booléens : les OU, par exemple, ne sont plus seulement des OU, ils fonctionnent aussi comme des ET flous (*fuzzy ANDs*) :

> *«The combination of these ranking algorithms makes an OR more than just an OR. It also is a fuzzy AND, because the more of the user's terms there are in a retrieved document, the higher it will rank.»* [Evans 1994].

1.3. Les systèmes de traitement du langage naturel (T.L.N.)

Le T.L.N. peut être considéré comme un sous-champ du secteur de l'intelligence artificielle. Les recherches qui y sont menées s'appuient sur des disciplines comme la linguistique, l'informatique et les sciences cognitives.

La recherche sur le langage naturel vise la compréhension et la modélisation de la façon dont l'être humain construit le sens d'une phrase ou d'un document, notamment via l'identification des indices exploités pour bâtir cette signification. Puisque l'acquisition du langage chez l'être humain se fait par le biais de l'assimilation progressive des règles et modèles (*patterns and templates*) qui le structurent – les enfants apprenant ainsi, par exemple, à exprimer l'opposition singulier/pluriel ou à construire une phrase, une question ou un ordre –, le T.L.N.

pose comme principe que, si nous arrivons à définir ces patrons et à les décrire à un ordinateur, alors nous pourrons enseigner à la machine une partie de la manière dont nous parlons et nous comprenons entre nous. L'experte américaine E. Liddy définit ainsi le T.L.N. :

> *"Natural language processing is a range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of particular tasks or applications"*. [Liddy 1998]

Les systèmes de T.L.N. sont, dans les faits, des systèmes statistiques auxquels l'on adjoint des bases conceptuelles, des bases de connaissances ou des thesaurus, et que l'on dote d'une interface en langue naturelle. La tâche centrale du T.L.N. en ce qui a trait aux SRI concerne la traduction de requêtes et de documents en langage naturel, donc potentiellement ambigus, en représentations internes non ambiguës pouvant être utilisées pour la mise en correspondance et le repérage. Idéalement, ce type de SRI pourrait permettre aux usagers de faire part de leurs requêtes de manière naturelle et avec tous les détails requis (exactement comme ils le feraient avec un bibliothécaire de référence...) et «comprendrait» le sens sous-jacent de la requête dans toute sa subtilité et sa complexité. Ce système permettant une analyse identique des documents de la base de données – peu importe leur nature –, il serait dès lors possible d'effectuer une mise en correspondance conceptuelle à part entière des requêtes et des documents.

La recherche sur le T.L.N. est actuellement en plein essor, car l'interrogation en langue naturelle de bases de données en texte intégral est depuis longtemps considérée comme l'une des clés possibles au «problème de l'usager final» dans le domaine de l'information électronique. (Comme nous le verrons plus loin, les SRI actuels conviennent surtout à leurs concepteurs et aux spécialistes formés aux procédures d'interrogation...). Les systèmes de T.L.N., s'ils peuvent éventuellement

manier les requêtes de type booléen ou statistique, fonctionnent en effet particulièrement bien sur des demandes en «langage ordinaire».

Il existe sept niveaux linguistiques (au moins) à partir desquels les humains extraient le sens d'un texte oral ou écrit et qui sont donc susceptibles d'être incorporés à un système de T.L.N. :

1.3.1. Niveau phonétique/phonologique

Ce niveau réfère à la façon dont les mots sont prononcés. Il n'est pas important en ce qui concerne le repérage de textes écrits, mais s'avère crucial pour la compréhension du langage oral et dans les systèmes de reconnaissance vocale.

1.3.2. Niveau morphologique

En linguistique, le *morphème* désigne la plus petite partie d'un mot porteuse de sens. Ce niveau concerne donc l'analyse componentielle des mots, par exemple l'étude des racines (*chanson* pour *chansonnier*, *chansonnette* ; en anglais *child* pour *childlike, childish, children*) ou des préfixes et suffixes (*poly-, in-, -ation, -s*).
Sous forme de troncature automatique (*stemming*), c'est le niveau le plus communément incorporé dans les SRI, et depuis le plus longtemps. Il est à noter que, plus les langues ont une morphologie riche (ce qui n'est pas le cas de l'anglais), plus l'attention portée dans un SRI à ce niveau linguistique s'avère payante.

1.3.3. Niveau lexical

Le niveau lexical concerne l'analyse du sens des mots (uniquement le sens «du dictionnaire», hors de tout contexte). C'est à ce niveau qu'un SRI peut opérer un étiquetage grammatical des parties du discours.

1.3.4. Niveau syntaxique

Ce niveau identifie le rôle joué par chacun des mots à l'intérieur d'une phrase et les relations des termes entre eux (le marquage des parties du discours réalisé à l'étape précédente est exploité à cette fin). La structure d'une phrase véhicule, en effet, ce genre d'informations, y compris dans les cas où le sens des mots eux-mêmes demeure inconnu. À titre d'exemple, *Paul frappe Jean* et *Jean frappe Paul* sont des énoncés formés des mêmes mots, mais dont les sens sont bien différents. La position des mots permet ici de déterminer qui est le sujet et qui est l'objet de l'action.

Les systèmes avancés de T.L.N. arrivent à exploiter cette information structurelle, notamment en emmagasinant des représentations de chaque phrase ou en caractérisant les genres de relations (par exemple, en identifiant comme des définitions les énoncés où des mots sont joints par des expressions comme *est un*).

1.3.5. Niveau sémantique

Ce niveau concerne l'analyse des sens possibles d'une phrase. Les mots à sens multiples y sont désambiguïsés. Puisque, vue de l'extérieur, une chaîne de caractères utilisée dans différents contextes demeure identique, la prise en compte des mots qui l'entourent se révèle nécessaire afin d'identifier le sens en jeu. Dans les SRI, il peut également y avoir expansion des requêtes (*query expansion*) par ajout de synonymes et développement des lieux géographiques (par exemple, *New England* se développera en *Maine, Massachusetts, New Hampshire, Vermont, Rhode Island* et *Connecticut*).

1.3.6. Niveau discursif

Le niveau discursif exploite la structure documentaire des différents genres de textes et de requêtes en vue d'une extraction additionnelle de sens. On peut ainsi tirer parti, par exemple, des traits structurels caractéristiques d'un article de journal, d'un article scientifique, d'un roman policier, etc. En profitant de cette structure prévisible, le T.L.N. peut déterminer le rôle d'une pièce d'information spécifique

dans un document (opinion, fait, prédiction, conclusion, etc.). La résolution des anaphores se fait également à ce niveau.

1.3.7. Niveau pragmatique

Ce niveau réfère au substrat sémantique formé par l'ensemble des connaissances du locuteur sur le monde, connaissances extérieures aux documents ou aux requêtes eux-mêmes mais nécessaires à leur bonne compréhension.

Pour inclure ce niveau dans les systèmes de T.L.N., il s'avère nécessaire de leur adjoindre de gigantesques bases de connaissances où des chercheurs ont recensé patiemment «tout» leur savoir sur le monde. Cette technique est longue et coûteuse ; elle présente, en outre, le désavantage de ne pas toujours refléter rapidement les dernières évolutions des connaissances humaines.

La taille de l'objet d'analyse augmente donc au fur et à mesure que l'on avance vers les niveaux supérieurs de compréhension, de même que les difficultés rencontrées par le traitement automatique :

> *The [...] levels of linguistic processing reflect an increasing size of unit of analysis as well as increasing complexity and difficulty as we move from top to bottom. The larger the unit of analysis becomes (i.e., from morpheme to word to sentence to paragraph to full document), the less precise the language phenomena and the greater the free choice and variability.* [Liddy 1998].

Bien sûr, tous les systèmes de T.L.N. n'opèrent pas sur l'ensemble de ces niveaux. Les produits qui prennent en charge les niveaux linguistiques élevés sont rares, surtout quand on s'intéresse à la fois au traitement des documents et à celui des requêtes. À titre d'exemple, on peut citer CONQUEST, INQUERY et DR-LINK[10]. En réalité, la plupart des systèmes contemporains dits de T.L.N. se limitent aux plus bas niveaux de compréhension, et ce, uniquement du côté des requêtes.

[10] Pour plus de détails, voir par exemple www.textwise.com (présentation de DR-LINK).

En ce qui concerne les SRI sur Internet, la majorité d'entre eux sont actuellement capables de tronquer sur le pluriel/singulier les termes de la requête, ou même d'ajouter/soustraire certaines autres formes d'un mot – essentiellement grâce à la manipulation de suffixes. Certains (INFOSEEK, ASKJEEVES) peuvent, en outre, interpréter quelque peu la syntaxe en «parsant» les éléments de la requête, mais ils n'appliquent pas cette technique au traitement des documents. On commence également à voir apparaître des procédés comme l'identification automatique des noms propres (basée sur la reconnaissance des majuscules et non sur une méthode plus motivée linguistiquement) et celle des locutions (qui semble s'appuyer surtout sur la proximité des mots entre eux).

2. Comment fonctionne un SRI

Le fonctionnement d'un SRI peut être divisé en quatre grandes étapes :

2.1. Le traitement des documents

C'est l'étape de l'ajout des documents au système et de la construction du *fichier inversé*, soit la liste alphabétique de tous les mots présents dans la base de données (les mots vides étant laissés de côté) avec les adresses de chacune de leurs occurrences. Pour les systèmes statistiques, il y a aussi établissement de poids différenciés pour les mots présents dans les documents.

D'autres opérations peuvent éventuellement avoir lieu à cette étape :

- l'ajout ou la création de bases de connaissances avec des lexiques internes ; des réseaux sémantiques ; des listes de syntagmes, de synonymes, de pronoms personnels ;
- l'extraction additionnelle d'information ou la réalisation d'opérations diverses sur les mots lors du stockage : lemmatisation ; identification des catégories du discours identification des noms propres et/ou communs ;

identification du rôle des mots et de leurs relations avec les autres mots de la phrase, du paragraphe, du document ;

- ♦ l'assignation automatique de termes d'indexation ou de larges catégories thématiques ;
- ♦ le stockage de représentations formelles de chacune des phrases.

2.2. Le traitement des requêtes

Cette étape concerne surtout les systèmes statistiques et de T.L.N., qui doivent accomplir en aval un travail qui est partiellement accompli en amont par le chercheur en ce qui concerne les systèmes booléens : rendre les requêtes compréhensibles par la machine.

Les systèmes statistiques peuvent éventuellement procéder à :

- ♦ l'identification des termes importants de la requête ;
- ♦ l'identification des racines et des variations de genre et de nombre ;
- ♦ l'assignation d'une pondération à chacun des termes de la requête.

Dans leur forme la plus achevée, les systèmes de T.L.N. peuvent mener à bien :

- ♦ l'étiquetage de toutes les parties du discours ;
- ♦ l'identification des sujets, objets, agents, verbes ;
- ♦ le développement des termes géographiques ;
- ♦ l'ajout de synonymes et de formes alternatives pour les noms propres.

Les systèmes de T.L.N. moins développés, pour leur part, se contentent habituellement d'effectuer l'identification des racines et une analyse syntaxique de base.

2.3. L'appariement des requêtes (*query matching*)

Cette étape concerne la mise en correspondance des requêtes avec le fichier inversé et, le cas échéant, la base de connaissances.

2.4. La présentation des résultats

Elle peut se faire par date, par champ ou par pertinence présumée par rapport à la requête.

3. Les critères d'évaluation d'un SRI : la précision et le rappel

La pertinence des résultats obtenus suite à une requête est le critère que l'on utilise habituellement lorsque l'on désire jauger l'efficacité et la qualité d'un SRI. Cette pertinence fait appel au jugement de l'usager final (ce dernier ayant toujours raison...) et on la mesure à l'aide de deux grands indicateurs : la *précision* et le *rappel*.

La précision se rapporte au pourcentage des documents repérés qui sont jugés pertinents par l'utilisateur final. Le rappel, quant à lui, concerne le pourcentage de documents, parmi tous ceux de la base de données qui seraient jugés pertinents par l'utilisateur final *s'ils étaient repérés*, qui sont effectivement rapatriés dans les faits. Les expressions *taux de bruit* et *taux de silence* sont également utilisées pour désigner ces phénomènes respectifs.

Le SRI accompli serait donc celui qui parviendrait à retrouver tout ce qui intéresse l'utilisateur tout en ne repêchant rien de ce qui ne l'intéresse pas – en d'autres termes, à atteindre à la fois 100% de rappel et 100% de précision. Il s'agit actuellement d'un idéal purement théorique puisque, dans les faits, ces deux taux ont plutôt tendance à être inversement proportionnels et à atteindre ensemble un total de 100% au lieu des 200% de la recherche idéale : un système qui favorise la précision voit d'ordinaire son taux de rappel baisser et vice-versa (le plus souvent, c'est la précision qui est privilégiée). La figure ci-dessous résume cette situation.

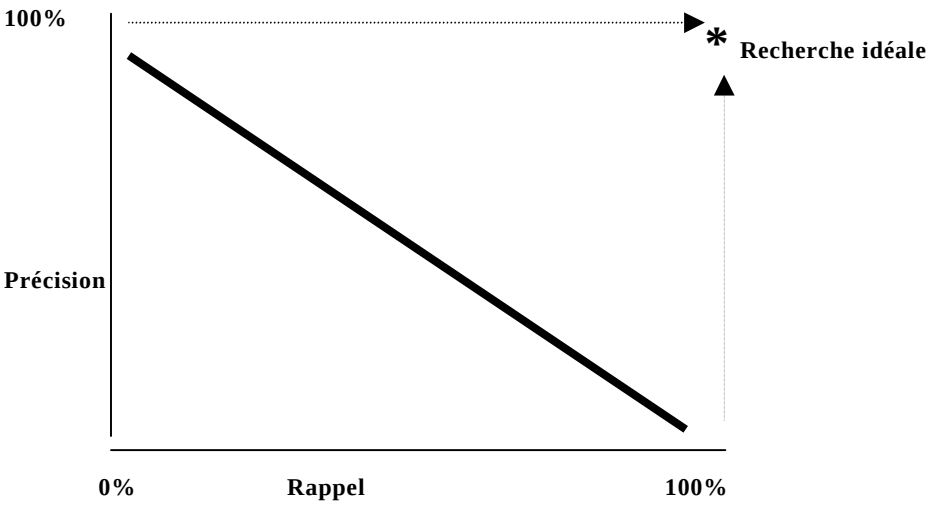

Figure 1 : la précision et le rappel

4. Conclusion de la première partie

Les performances des SRI en termes de précision et de rappel sont fort variables. Les systèmes booléens et statistiques peuvent ainsi être disposés aux deux extrémités d'un même spectre. mplacables, les systèmes booléens repêchent exactement ce qu'on leur a demandé. Si l'on a bien formulé sa requête, on obtient ce que l'on cherchait ; sinon, on risque de ne rien repérer qui soit utile. Selon certaines études, les systèmes booléens, même en présence de spécialistes chevronnés de l'information, n'atteignent guère que 20% de taux de rappel [Addison *& al.* 1993].

Les systèmes booléens présentent généralement des interfaces peu conviviales, ce qui, conjugué à leur mode d'interrogation à base de mots clés et d'opérateurs logiques et de proximité, contribue à les rendre difficiles à maîtriser pour les usagers non spécialistes. En fait, ils s'avèrent souvent frustrants même pour les experts, qui doivent mémoriser les différences subtiles entre les diverses interfaces booléennes.

Comme le souligne R. Evans :

> *The strength of exact-match Boolean searching is precision; its weakness is recall. It is a very good tool for finding a specific document that the user already knows is in the database, because the user knows specific query terms to use. But for finding documents about a general topic that is defined subjectively in the end-user's mind, traditional Boolean falls short.* [Evans 1994]

Les systèmes statistiques, au contraire, misent sur le rappel (ils se soucient toutefois également de la précision, puisque leur classement présente les résultats les plus pertinents en premier). À ce niveau, ils font un peu mieux que les systèmes booléens, atteignant un taux de rappel de près de 50% [Addison *&* *al.* 1993]. Avec ces systèmes, on obtient non seulement ce que l'on a demandé, mais aussi, éventuellement, ce que l'on *aurait dû* demander... de même que, bien souvent, des documents qui renferment les termes de la requête, mais pas l'information recherchée. Leur façon de procéder comporte néanmoins l'avantage de suggérer à l'usager des façons de modifier sa question s'il n'a pas trouvé ce qu'il cherchait : ainsi, en regardant les termes co-occurrents avec les mots de sa requête dans des documents partiellement pertinents, il peut décider d'élargir ou de préciser la requête initiale.

Le T.L.N., se superposant aux systèmes booléens ou statistiques, engendre pour sa part un accroissement à la fois du rappel et de la précision. Utilisé au niveau du traitement du document, il permet une extraction et un stockage plus riche de l'information ; utilisé au niveau du traitement de la requête, il facilite l'expression des besoins d'information grâce à la puissance du langage réel ; utilisé au niveau de l'évaluation des réponses, il simplifie la mise en correspondance avec le sens et l'intention de la requête, améliorant du même coup l'évaluation de pertinence. On peut prévoir que, à l'avenir, les interfaces en langue naturelle s'imposeront comme les préférées de la plupart des utilisateurs.

Selon S. Feldman :

> *Without NLP, we have gone about as far as we can go. Text databases are getting bigger. Search engines are returning larger and larger sets of documents. While Boolean search techniques allow us to narrow down our retrieval to a manageable size, they eliminate too many potentially valuable documents. statistical search techniques overwhelm us with documents, even with relevance ranking. NLP presents new tools for honing a search query so that it states our information need fully and then matches that query with an elaborate knowledge base built with NLP techniques. My prediction is that the best systems in the future will be those that combine useful features from several information retrieval technologies*. [Feldman 1999]

Les approches booléennes, statistiques et de T.L.N. doivent donc être vues comme étant complémentaires plutôt que concurrentes ou mutuellement exclusives. À titre d'exemple, dans le domaine légal, un «cas exemplaire» au niveau de la pertinence n'a que peu de valeur – en admettant même qu'il en ait aucune – si un statut ultérieur a été promulgué, ou si une instance supérieure a renversé par la suite le jugement concerné. Dans cette conjoncture, la précision booléenne s'avère préférable au tri de pertinence statistique. Les systèmes booléens, outre leur efficacité pour repérer un objet connu d'avance, performent habituellement bien lors de recherches par champ ou bibliographiques. Les systèmes statistiques et de T.L.N. offrent de meilleures performances pour les recherches plus générales, celles en texte intégral ou celles menées par des usagers inexpérimentés.

L'avenir est sans aucun doute à la combinaison d'éléments en provenance de ces diverses méthodes. De plus en plus de SRI amalgament ainsi le meilleur de deux (ou trois) mondes : des systèmes booléens offrent un classement de pertinence ; des systèmes statistiques recourent à des techniques de T.L.N. comme la lemmatisation

automatique ou l'identification des noms propres. De même, sans le tri de pertinence, les requêtes en langage naturel ne se révèlent guère exploitables : d'ordinaire, dans une telle requête, les termes sont analysés comme étant séparés par des OU, ce qui résulte en un mode de recherche très large. Grâce au tri de pertinence, les documents les plus pertinents se retrouvent en tête de liste.

Les stratégies de recherche évoluent également : grâce au tri de pertinence qui permet la gestion de larges ensembles de réponses, il devient possible de «tailler» les requêtes booléennes de manière à les rendre beaucoup plus larges qu'il n'était praticable autrefois, en incluant des synonymes ou des termes de recherche supplémentaires. D'où la nécessité pour les usagers – surtout professionnels – de revoir leurs pratiques :

> *Taking full advantage of these advanced technologies [i.e. relevance ranking, natural-language searching, document-as-query feedback, and automatic concept construction] requires experienced searchers to rethink their strategies. The old techniques of narrowing a Boolean set blindly until it is small enough to cope with – discarding relevant documents with each step – are the wrong way to search with these new tools. Rather, the user must learn to expand the search net, then browse from the top down.* [Evans 1994]

Ajoutons, enfin, que, dans le contexte d'Internet, l'application rigoureuse des critères de rappel et de précision pose problème. D'une part, la croissance exponentielle de la taille du Web et la couverture limitée que les outils de recherche en effectuent font du rappel une mesure difficile à exploiter telle quelle, car il s'avère impossible, dans les faits, de déterminer combien de réponses potentiellement pertinentes existent pour une requête. D'autre part, l'évaluation de précision doit aussi être adaptée puisque les SRI sur Internet retournent habituellement une très grande quantité de résultats, qui ne peut faire l'objet d'une évaluation de pertinence dans sa totalité.

Seconde partie

les SRI sur Internet[11]

[11] Cette section s'inspire de notre note de synthèse (*Les Outils de recherche sur Internet : typologie et principales caractéristiques*), que l'on pourra consulter pour plus de détails.

1. Les obstacles au repérage de l'information sur Internet

Avant de présenter plus en détail les différents types de SRI sur Internet, il peut être utile de rappeler succinctement le contexte général dans lequel évolue la recherche d'information sur le Web. La liste qui suit résume les principales difficultés qu'elle doit gérer (plusieurs sont de portée plus globale et s'appliquent à tous les systèmes d'information) :

1.1. Le manque d'habileté et de formation à la recherche des usagers

Une étude de 1986 portant sur le comportement d'usagers inexpérimentés face à un SRI a démontré que le quart d'entre eux n'atteignaient même pas le seuil défini comme minimal d'habileté de recherche[12]. On peut raisonnablement supposer qu'une situation semblable prévaut actuellement sur Internet où les usagers spécialistes des systèmes d'information ne constituent plus qu'une minorité, appelée sans doute à devenir encore plus infime dans les prochaines années.

1.2. La couverture limitée des SRI

Sur Internet, il est évident que, plus la base de données d'un outil est imposante et complète, plus ce dernier est susceptible de trouver des réponses à une requête, en particulier pour les sujets obscurs ou très précis. Toutefois, selon un article publié dans la revue *Science*[13], le meilleur outil au niveau de la couverture du Web, HOTBOT, n'indexait que 34% des 320 millions de pages estimées disponibles au moment de l'étude. Le pire, LYCOS, ne dépasserait pas les 3%. Cette situation inquiétante ne semble pas destinée à s'améliorer – bien au contraire – en regard de la croissance incontrôlée du Web, et également du fait que l'augmentation de la

[12] Borgman, C.L. «The user's mental model of an information retrieval system: an experiment on a prototype online catalog». *International Journal of Man-Machine Studies*, 1986, (24) : 47-64.
[13] Lawrence, S.R. et C.L. Giles. «Searching the World Wide Web». *Science* 280, 1998 : 98-100.

taille de leur base de données ne semble guère être une priorité chez la plupart des concepteurs d'outils[14].

Un autre problème considérable, à ce niveau, concerne l'augmentation du «Web invisible», c'est-à-dire des pages dont le traitement pose d'importantes difficultés aux outils de recherche. Les cadres[15] (*frames*) en sont un exemple typique : il n'est pas rare de voir des sites de ce genre de 100 pages ou plus uniquement représentés dans l'index d'un outil par leur page d'accueil (où, en outre, la seule information indexée est souvent l'inscription *This site requires frames*). Les *pages dynamiques*[16] sont également problématiques pour la plupart des outils de recherche, de même que celles qui emploient la technologie XML[17].

1.3. L'instabilité des ressources

Internet est d'une mouvance intrinsèque : chaque jour, des ressources apparaissent, disparaissent ou déménagent. Comme, par ailleurs, le Web descend en droite ligne du Gopher[18], il en a hérité plusieurs des caractéristiques, notamment le fait que les liens entre documents ne soient pas bidirectionnels : une ressource vers laquelle pointe un lien n'est pas au courant de cet état de fait. Si cette ressource change, est déplacée ou cesse d'exister, les liens URL ne font donc pas l'objet d'une mise à jour automatique, et demeurent bien souvent «pendants» (d'où le fameux code «Erreur 404»). C'est ce qui explique que les bases de données des outils de

---

[14] Ceci s'explique probablement par le fait que le grand public (soit la majorité des internautes) privilégie la précision aux dépens de l'exhaustivité : un usager qui fait une requête sur un mot clé comme *cars* ou *travels* n'a vraiment pas besoin de repêcher toutes les références potentiellement pertinentes qui existent sur le Web... Du reste, malgré sa piètre performance au niveau du rappel, les consommateurs continuent de plébisciter LYCOS qui demeure un des services majeurs de recherche d'information sur le Web.

[15] Les cadres permettent de disposer de plusieurs fenêtres sur une page Web.

[16] Les *pages dynamiques* sont des pages qui «résident» en pièces détachées dans une base de données : corps de la page, en-têtes, pieds de page, etc. Suite à une demande d'accès, la base de données assemble les divers éléments et délivre en temps réel la page Web. Les pages dynamiques sont reconnaissables à la présence d'un point d'interrogation dans leur URL (exemple : *http://www.website.com/cgi-bin/getpage.cgi?name=sitemap*). C'est précisément ce symbole qui pose problème, car la plupart des outils de type moteur ne lisent pas l'URL au-delà, d'où l'impossibilité d'indexer la page (il s'agit là d'un choix délibéré, destiné à éviter les «pièges à robot» où une même page peut être soumise des milliers de fois sous des URL légèrement différentes).

[17] XML (*eXtended Mark-up Language*) est un format de document électronique destiné, à terme, à remplacer HTML (*Hypertext Mark-up Language*), actuellement le standard pour les documents diffusés sur le Web.

[18] Pratiquement tombé en désuétude aujourd'hui, Gopher, développé à l'Université du Minnesota, fut un précurseur du Web qui permettait la navigation sur le Réseau par des choix dans des menus. Un outil de recherche nommé VERONICA fut

recherche comportent inévitablement des liens invalides, en quantité plus ou moins importante selon les cas.

1.4. L'ambiguïté linguistique

1.4.1. La surabondance de synonymes

Cette situation s'explique par la valorisation de la paraphrase dans les textes autres que purement techniques, pour des raisons d'élégance et de style. Elle est également tributaire des différences linguistiques diachroniques, régionales ou professionnelles.

1.4.2. La polysémie

Les données suivantes concernent l'anglais, mais sont intéressantes à titre indicatif. Le *Webster's Seventh Dictionary* recense quelque 60 000 entrées ; or, de celles-ci, 21 488 (soit presque 40%) ont deux sens ou plus [Wacholder *& al*. 1994]. En fait, dans la langue de Shakespeare, un mot aurait en moyenne sept acceptions différentes... La situation est d'autant plus préoccupante que ce sont les mots les plus courants qui ont le plus de sens distincts : à titre d'exemple, *run* a 29 sens, qui se subdivisent en près de 125 sous-sens.

À cette polysémie fondamentale des langues naturelles s'ajoutent, en outre, les usages métaphoriques et les comparaisons.

Nous l'avons vu, beaucoup des pièges du traitement automatique sont occasionnés par cette ambiguïté du langage. Les problèmes d'ambiguïté lors du repérage d'information sur Internet sont d'autant plus critiques que les SRI qu'on y trouve actuellement ne parviennent pas à extraire l'information contextuelle contenue dans les documents et les requêtes ; ils ne disposent pas non plus de la masse d'informations sur le monde emmagasinée dans le cerveau des usagers : *«To humans, disambiguation and paraphrasing are second-nature, to the point that they find it hard to conceive of the inherent complexity of linguistic expressions.»* [Wacholder & al. 1994]

---

développé pour Gopher ; il autorisait la recherche par mots clés dans une base de données de descriptions d'items de

1.4.3. Les variations orthographiques et les erreurs d'orthographe et de frappe

Un certain nombre de noms communs sont d'orthographe fluctuante, par exemple *clé/clef* ou *fantasme/phantasme* (en anglais, on pourrait citer *gray/grey*, *theatre/theater*, *aluminium/aluminum*, etc.). Les noms propres présentent eux aussi des variantes : ainsi, dans un texte, ils peuvent figurer en version abrégée dans le titre (pour sauver de l'espace), apparaître en version complète dans le premier paragraphe afin d'établir clairement la référence, puis revenir par la suite sous des formes plus courtes, l'entité ayant déjà été introduite. Outre les synonymes, le chercheur doit donc penser aux diverses variantes orthographiques possibles quand vient le moment d'imaginer les différentes manières dont un concept peut être exprimé (l'emploi de la troncature peut éventuellement lui faciliter un peu la tâche).

Le problème des fautes, pour sa part, est aggravé par l'incorporation de plus en plus fréquente, dans les bases de données, de textes numérisés à l'aide de techniques de reconnaissance optique de caractères (ROC[19]). Selon certaines études, en effet, ces textes, sans une relecture attentive des épreuves, peuvent facilement comporter jusqu'à 30 erreurs par page [Feldman 1999].

1.4.4. Les pertes d'information lors du traitement

Comme nous le verrons dans la troisième partie de ce travail, des phénomènes comme l'ordre des mots, la distinction minuscules/majuscules ou la présence de signes diacritiques et de caractères spéciaux ne sont pas toujours gérés de manière cohérente et efficace par les outils de recherche. Des subtilités comme les distinctions entre *AIDS* et *aids* (*SIDA* et *assistants*), *school library* et *library school* (*bibliothèque scolaire* et *école de bibliothéconomie*), *tache* et *tâche* leur échappent donc souvent, de même que la nécessité de mettre en correspondance des formes comme *online* et *on-line*. Il y a là une perte d'information importante, puisque la prise en compte des majuscules, par exemple, peut favoriser le traitement des abréviations[20], des acronymes et des noms propres.

---

menus Gopher du monde entier.

[19] OCR sous sa forme anglaise.

[20] Quoique ces dernières soient aussi souvent écrites en minuscules (comme *modem*, pour *modulation/demodulation*).

1.4.5. L'inconstance de l'indexation humaine

Selon certaines études, l'homogénéité de l'indexation humaine est au mieux de 50%, y compris en ce qui concerne le travail accompli par une seule et même personne [Feldman 1999]. On peut supposer que cette situation s'observe également dans les outils de recherche de type annuaire[21].

1.4.6. La difficulté de formulation de certains concepts

Sur Internet, les SRI ne permettent pas toujours la formulation de requêtes en langue naturelle, comme le souligne E. Liddy : *«The engines expect minimal one-word or two-word queries and are optimized for them, rather than for sentences, which would enable the user to fully present their information need.»* [Liddy 1998] Cela augmente la difficulté éprouvée à définir des concepts importants mais vagues : «*Speaking in code is difficult, and it leaves out important aspects of thought.*» [Feldman 1999]. Actuellement, la seule solution consiste bien souvent à administrer aux SRI de longues chaînes de synonymes et d'adjectifs.

1.4.7. Les «false drops»[22]

Il faut entendre par là les documents qui sont repêchés suite à une requête, mais qui sont sans rapport aucun avec le sujet. Ce phénomène de bruit est dû en bonne partie au tandem bon mot/mauvais sens ; il concerne en particulier les systèmes booléens, qui ne vérifient pas automatiquement la proximité et la fréquence des mots. À un moindre niveau, il affecte également les systèmes statistiques. Le problème des «false drops» illustre avec acuité, d'ailleurs, les limites de ces deux modes de repérage :

> *Most of today's commercial and Web search technologies retrieve information without knowing what it means. They do this by matching strings of letters (words) in the query to the documents in*

[21] Voir la suite de cette section.
[22] Malgré nos essais, nous n'avons pu formuler d'équivalent français pour cette expression anglo-saxonne fort répandue.

> *the database in order to find exact or best matches. This is like trying to carry on a conversation with a parrot. The parrot can mimic speech, but it ties words to, at most, a treat or a curse, not to their inherent meaning.* [Feldman 1999]

2. Les annuaires

Nous retiendrons comme premier type d'outils de recherche sur Internet les *annuaires* – que l'on appelle également *guides*, *répertoires* ou *catalogues*. Le prototype en fut la WORLD WIDE WEB VIRTUAL LIBRARY, localisée au CERN[23]. Les annuaires sont des regroupements par sujet des ressources d'Internet. Ils consistent en des classements arborescents où l'accès au thème souhaité s'effectue en parcourant une série de rubriques et de sous-rubriques. Comme on peut le lire dans l'aide en ligne de l'annuaire YAHOO!, *«l'analogie avec un arbre s'impose clairement : chaque catégorie du guide, ou branche de l'arbre, abrite plusieurs sous-catégories, d'autres branches qui, elles-mêmes, vous donnent le choix entre plusieurs chemins possibles au fur et à mesure de votre balade, etc.»*. En fait, les annuaires, dont les ramifications successives conduisent à des sujets de plus en plus pointus, pratiquent ce que l'on pourrait appeler le «principe de l'entonnoir». D'ordinaire, ils incorporent également un moteur de recherche par mot clé, ce qui permet d'effectuer directement une requête sur le sujet souhaité.

Ces listes thématiques de sites constituent en quelque sorte l'équivalent cybernétique (et moins élaboré) du plan de classification que l'on applique traditionnellement dans les bibliothèques et centres de documentation. Elles présentent également des similitudes avec les bibliographies thématiques, info-guides et autres listes imprimées de ressources que les bibliothécaires mettent à la disposition de leur clientèle, et avec ces pages Web personnelles qui proposent en

[23] Également connu sous l'appellation *Laboratoire européen pour la physique des particules*, le CERN, situé en Suisse, est à l'origine du concept de *World Wide Web*.

compilation les «meilleures» ressources d'Internet ou, tout simplement, les sites préférés de leur auteur.

Le consultant Internet français Olivier Andrieu propose la définition suivante des annuaires :

> *« Un annuaire est un outil de recherche qui recense un certain nombre de sites (et non de pages) Web au travers de fiches descriptives comprenant, en règle générale, le titre, l'adresse (l'URL) et un bref commentaire d'une longueur allant le plus souvent de 15 à 25 mots au maximum. Chaque site est inscrit dans une ou plusieurs catégorie(s) – on parle également de rubrique(s) –. Ces outils peuvent ainsi être considérés comme les pages jaunes du Web. Lorsqu'un mot clé est saisi dans le formulaire proposé, l'annuaire effectue une recherche sur les occurrences de ce terme dans ses fiches descriptives de site, et non pas dans le contenu des pages du site en question. Il s'agit là de la différence la plus notable avec les moteurs de recherche ».* [www.abondance.com]

On peut résumer ainsi les principales caractéristiques des annuaires :

- ils recensent des *sites* et non des *pages* individuelles ;
- ils structurent leur inventaire selon une classification en général propre à l'outil (certains ont recours aux classifications documentaires traditionnelles comme celle de la Bibliothèque du Congrès de Washington – utilisée notamment par la WORLD WIDE WEB VIRTUAL LIBRARY – ou celle de Dewey, mais le cas demeure rare) ;
- le repérage et la catégorisation des ressources s'effectuent souvent manuellement, au moins en partie. Les annuaires recourent, à cette fin, soit à des professionnels de la documentation (bibliothécaires, documentalistes), soit à des spécialistes des diverses thématiques concernées (par exemple, des médecins pour la rubrique *Santé*), soit encore à des volontaires (rémunérés ou non) ;
- les annuaires incorporent parfois directement des sites Web dans leur base de données (suite à une décision de l'équipe éditoriale ou à une suggestion en provenance des usagers du service) ; toutefois, il est généralement

> nécessaire d'entreprendre une démarche délibérée d'inscription : le responsable du site à enregistrer doit soumettre ce dernier, qui est alors visité, évalué et – si accepté – inclus dans l'arborescence de l'outil.

Le principe des annuaires présente plusieurs avantages. Tout d'abord, ces instruments permettent de guider l'utilisateur dans ses investigations ; ils s'avèrent donc moins «intimidants» que la ligne de commande vide des autres outils de recherche. Grâce à la catégorisation effectuée sur l'information, il s'avère aisé pour l'usager de «butiner» entre sites traitant d'un même sujet, un peu comme l'on bouquine devant les rayons d'une bibliothèque. La philosophie des annuaires permet également de limiter le taux de bruit, et s'accompagne d'une substantielle valeur ajoutée due à l'activité humaine de sélection, d'évaluation et de hiérarchisation des ressources. On note également, bien sûr, certains inconvénients : augmentation du taux de silence (en supposant qu'un document soit classé dans une seule catégorie), couverture relativement restreinte d'un bassin potentiel de millions de sites Web, mise à jour moins rapide que pour les autres outils, dépendance par rapport aux choix éditoriaux des réalisateurs (il n'y a souvent qu'un pas entre l'évaluation des ressources et la censure...). En outre, même si les requêtes de recherche sont possibles, elles offrent en général moins de souplesse et de précision que celles permises dans les outils de type moteur.

De manière globale, on peut donc dire que les annuaires, favorisant le repérage de sites généraux sur un sujet donné, s'avèrent surtout utiles pour des fouilles vastes et thématiques ou pour débuter une recherche d'information encore mal définie. Leur convivialité en faisant par ailleurs les outils de recherche les plus simples d'utilisation, ils sont également tout indiqués pour les débutants.

Il convient, enfin, de souligner que les annuaires disponibles en plusieurs versions linguistiques ne constituent pas autant de copies d'une même base de données simplement coiffées d'interfaces différentes. Il s'agit bien, dans les faits, de bases totalement dissociées ; il importe donc de les interroger successivement et d'effectuer les requêtes dans la langue de l'interface (par exemple, en anglais dans YAHOO! INTERNATIONAL et en français dans YAHOO! FRANCE).

Quelques annuaires en langue anglaise :

Nom	**URL**
Galaxy	http://galaxy.einet.net/
Jassan	http://www.jassan.com/
Looksmart	http://www.looksmart.com/
Magellan	http://magellan.excite.com/
Open Directory Project	http://dmoz.org/
Snap	http://www.snap.com/
Yahoo! International	http://www.yahoo.com/

Quelques annuaires en langue française :

Nom	**URL**
Carrefour	http://www.carrefour.net/
CTrouvé	http://www.ctrouve.com/
Francité	http://www.i3d.qc.ca/
Nomade	http://www.nomade.fr/
Yahoo! France	http://www.yahoo.fr/

3. Les moteurs

Le second type d'outils de recherche sur Internet est constitué par ce que l'on appelle des *moteurs*. WEBCRAWLER fut le premier instrument de ce genre, en ligne depuis avril 1994. Si les annuaires évoquent le plan de classification des bibliothèques traditionnelles, les moteurs, pour leur part, ressemblent un peu à ces programmes qui produisent automatiquement des index primitifs en associant, à chaque mot d'un document, la ou les page(s) où il figure – du reste, on les appelle aussi parfois des *index*. Pour reprendre une comparaison communément admise : *«If we regard the World Wide Web as a huge, disorganized book, then a subject*

directory is like a table of contents and search engines are like the book's indexes.» [Dong & Su 1997].

Les moteurs permettent à l'usager de repérer l'information non suite à une navigation thématique, mais via l'interrogation à l'aide de mots clés et de commandes logiques d'une base de données indexée ; leur fonctionnement rejoint ainsi celui des logiciels de gestion documentaire usuels. En général, deux modes de recherche sont disponibles : *recherche simple* (proposée par défaut à partir de la page d'accueil de l'outil, avec plus ou moins de possibilités de recherche) et *recherche avancée* (accessible en option et où des possibilités de recherche variées et approfondies, souvent paramétrables, sont offertes).

Voici ce que dit O. Andrieu à propos des moteurs :

> *Lorsque l'internaute saisit un mot clé dans le formulaire proposé, le moteur va en rechercher les occurrences dans son index, c'est-à-dire dans le contenu (le texte) des pages Web sauvegardées au préalable. Une fois identifié le «lot» de pages contenant le terme demandé, le moteur classe les pages par ordre de pertinence, selon un ordre et un algorithme (basé sur certains critères de tri) qui lui est spécifique. Le moteur de recherche effectue donc ses recherches sur des pages Web, alors que l'annuaire, pour sa part, vous proposera des sites Web. Là est toute la différence qui explique qu'il est absolument impossible de comparer les résultats fournis par les deux types d'outils.* [http://www.abondance.com/]

Le fonctionnement des moteurs s'appuie sur la collecte de données par des *robots*, lesquelles sont ensuite indexées directement à l'aide des mots qui les constituent. De gigantesques bases de données – autrement plus imposantes que celles des annuaires – sont ainsi élaborées ; elles opèrent *grosso modo* sur le mode des *fichiers inversés* en établissant des correspondances entre des mots et des URL. Les utilisateurs sondent la base à l'aide d'un module d'interrogation qui recourt à un

langage de requête plus ou moins standard ; des interfaces conviviales sont généralement mises en place afin de faciliter l'interaction. L'activité des moteurs de recherche, contrairement à celle des annuaires, est entièrement automatisée.
Les robots – qui connaissent diverses autres appellations évocatrices, notamment *spider* («araignée»), *ant* («fourmi»), *worm* («ver de terre» ou «se faufiler»), *wanderer* («vagabond»), *crawler* («nageur»), etc. – sont tout simplement des programmes informatiques qui tournent sur un ordinateur relié au Réseau et qui explorent systématiquement celui-ci de manière à collecter l'information présente. Les robots procèdent en repérant les liens hypertextuels d'un document pour ensuite aller visiter les pages vers lesquelles pointe ce dernier. Ils parcourent ainsi rapidement un site, puis d'autres sites qui lui sont liés, et ainsi de suite. Comme le fait remarquer J.-N. Plourde, *«c'est l'automatisation et la systématisation de ce que l'on fait de chez soi en se baladant dans le Web*» [Plourde 1996]. Une fois le site indexé, le robot revient régulièrement «capturer» une version plus récente des différentes pages. Il n'est pas rare que le même robot soit utilisé par plusieurs moteurs différents, avec seulement quelques différences de paramétrage.

Généralement, seuls les fichiers ASCII et HTML sont indexés (et non, par exemple, les fichiers compressés ou de type *.pdf*). Le fonctionnement mécanique des robots fait en sorte qu'il est fort difficile de contrôler quelles pages sont récupérées pour être indexées. Le contenu de la base d'un moteur demeure donc essentiellement tributaire des sites utilisés comme points de départ et de la stratégie privilégiée pour la visite des liens (ce peut être une stratégie en largeur, où tous les liens immédiats dans l'ensemble des pages rapatriées sont visités, ou une stratégie en profondeur où, pour une sélection de documents, le robot descend de page en page jusqu'au dernier lien existant) : *«Each robot will [...] produce a different view of resources on the WWW, according to its page retrieval strategy.»* [Poulter 1997]

Mentionnons que, contrairement aux annuaires, les moteurs qui se déclinent en plusieurs versions linguistiques ne proposent, en général, que des versions localisées d'une même base de données (EXCITE FRANCE, LYCOS FRANCE). Beaucoup des

grands moteurs internationaux ne se donnent, d'ailleurs, pas cette peine et se contentent de doter leur interface anglophone d'une option de recherche de restriction linguistique (ALTAVISTA, HOTBOT, NORTHERN LIGHT).

Un des avantages de la démarche de type moteur réside dans le fait que l'utilisateur n'a pas à connaître la catégorie (et la structure hiérarchique) dans laquelle pourrait se trouver l'information recherchée, puisque cette dernière n'est pas compartimentée de la sorte et que la recherche s'opère principalement par concordance avec un modèle (*pattern matching*). Par ailleurs, comme l'absence d'intervention humaine équivaut souvent à une absence de déontologie, les moteurs sont en principe plus performants que les annuaires pour repérer des documents à contenu sensible (violence, pornographie) ou carrément sujets à controverse (sites haineux, terroristes, pédophiles, etc.), une caractéristique que l'on peut ou non applaudir mais qui est conforme à l'esprit libertaire et anarchiste du Net.

Le taux de rappel obtenu par les moteurs est souvent bon, mais il s'accompagne malheureusement d'une grande quantité de bruit, c'est-à-dire d'une baisse du taux de précision : les moteurs suscitent des réponses très hétérogènes, où les doublons abondent parfois. La (non-)mise à jour des index constitue souvent également une source de problèmes. Autre inconvénient : contrairement aux annuaires, les moteurs abandonnent l'usager à lui-même (rien ne guide ni ne balise la recherche) et ne fonctionnent habituellement pas sur le mode d'un ensemble de réponses qu'il est possible de restreindre et d'affiner successivement : la recherche se fait en un coup et un seul. Enfin, leur maniement demeure délicat et les recherches peuvent prendre beaucoup de temps.

Généralement plus appréciés des internautes aguerris que des débutants, les moteurs, en un certain sens, sont plus «puissants» que les annuaires. Ils sont donc tout indiqués pour des recherches qui portent sur des sujets fins et précis ou sur un objet dont l'existence est connue d'avance, mais ils risquent de générer des milliers de réponses d'intérêt inégal si la requête s'avère trop vague ou trop commune.

Comme on le voit, les moteurs se différencient des annuaires à de nombreux points de vue. On peut résumer ainsi leurs principales caractéristiques :

- Ils recensent des pages individuelles et non des sites en tant qu'entités ;

- Aucune structuration, classification ou hiérarchisation de l'information n'est effectuée ;
- Leur fonctionnement ne comporte aucune intervention humaine ;
- Il n'est pas absolument nécessaire d'inscrire les pages d'un site auprès des divers moteurs : on peut tout simplement choisir d'attendre que les robots débusquent le site concerné au détour d'un lien, le visitent et en indexent les différentes pages[24]. Cette méthode demeure néanmoins aléatoire et requiert habituellement l'écoulement d'un certain laps de temps. Il est donc nettement préférable d'opter pour la soumission manuelle des URL que l'on désire faire connaître. Pratiquement tous les moteurs offrent, en effet, une fonction de type Add a site ou Add URL, qui sert à signaler au robot l'adresse de pages à visiter[25].

Enfin, si les annuaires et les moteurs sont des outils bien distincts, il convient de signaler que de plus en plus de sites de recherche combinent l'accès aux deux genres d'instruments, selon des formules qui privilégient l'un ou l'autre type : moteur agrémenté d'un annuaire (par exemple, VOILA) ou annuaire complété d'un moteur de recherche externe (par exemple, FRANCITE). Une autre tactique consiste à conclure des accords de partenariat avec des sociétés concurrentes : l'annuaire YAHOO!, par exemple, dirige l'internaute sur le moteur INKTOMI en cas de recherche infructueuse. Les moteurs INFOSEEK FRANCE et EXCITE FRANCE, pour leur part, affichent les catégories et les descriptions de sites de l'annuaire NOMADE.

Quelques moteurs en langue anglaise :

Nom	**URL**
ALTAVISTA	http://www.altavista.com/

[24] Il n'est habituellement pas possible de retirer ou de modifier manuellement les références ainsi incluses péremptoirement dans la base de données d'un moteur. Toutefois, on peut empêcher l'aspiration d'une page grâce à l'emploi de la balise HTML <ROBOTS> ou à l'insertion d'un fichier spécial (*robots.txt*). Du moins en théorie, car les moteurs de recherche ne prennent pas toujours en compte la présence de ces éléments...

	http://www.av.com/
	http://altavista.digital.com/
THE ELECTRIC MONK	http://www.electricmonk.com
EXCITE	http://www.excite.com/
EXCITE version française	http://www.fr.excite.com
HOTBOT	http://www.hotbot.com/
INFOSEEK	http://infoseek.go.com/
LYCOS	http://www-english.lycos.com/
LYCOS version française	http://www.lycos.fr/
NORTHERN LIGHT	http://www.northernlight.com/
	http://www.nlsearch.com/
WEBCRAWLER	http://www.webcrawler.com/

Quelques moteurs en langue française :

Nom	**URL**
ÉCILA	http://www.ecila.fr/
LOKACE	http://www.lokace.com/
VOILA	http://www.voila.fr/
VOILA version mondiale	http://www.voila.com/

4. Les métamoteurs

Le troisième grand groupe d'outils de recherche est celui des *métamoteurs*. Ce sont des instruments qui visent à faciliter la transmission d'une même requête vers différents moteurs et annuaires.

Les métamoteurs se subdivisent en deux catégories. La première rassemble les *Configurable Unified Search Interfaces* (*CUSI*), que l'on appelle également – de manière plus prosaïque – les *bibliothèques de moteurs* ou les *All in One*. Ce genre

[25] Il n'est, du reste, pas requis d'être responsable d'un site pour proposer son inclusion à un moteur. Chaque internaute est libre de suggérer ce qui lui plaît, situation qui contribue sans doute au caractère fortement hétéroclite des bases de données

d'instrument recense habituellement un grand nombre d'outils de recherche en fournissant un accès direct, sur une même page, à la ligne de commande de chacun d'eux. Utiles dans la mesure où ils permettent la consultation de plusieurs services à partir d'un même site et disposent souvent d'une interface astucieuse qui évite à l'usager d'avoir à retaper continuellement sa requête, ces métamoteurs de première génération demeurent, toutefois, assez primitifs et ne rendent que peu de services supplémentaires. Ils se chargent tout simplement de communiquer la requête concernée aux différents outils de recherche, généralement de façon séquentielle.

Quelques CUSI, parmi bien d'autres :

Nom	**URL**
ALL-IN-ONE SEARCH PAGE	http://www.allonesearch.com/
Easy Search (japonais)	http://www.aist.go.jp/NIBH/~honda/EasySEARCH/index.cgi
GOLDENBRICK (francophone)	http://www.goldenbrick.fr/goldensearch/recherche.html
THE SEARCH PLACE	http://users.isaac.net/duane/search/

Ceci dit, le terme *métamoteur* renvoie la plupart du temps à une seconde catégorie d'outils, à valeur ajoutée ceux-là : les *Simultaneous Unified Search Interfaces* (*SUSI*). Ces métamoteurs fonctionnent en transmettant simultanément la requête de l'usager à plusieurs outils de recherche, principalement des moteurs. La quantité d'outils ainsi «interpellés» est très variable ; elle se situe d'ordinaire entre 5 et 150. Les métamoteurs récupèrent par la suite les différentes listes de résultats et les façonnent en un document unique. Certains procèdent, en outre, à un classement de pertinence supplémentaire et à l'élimination des doublons. Plusieurs d'entre eux permettent également de configurer la liste des sources à interroger.

---

de ce type d'outil.

Les principaux avantages liés à l'emploi des SUSI ont trait au gain de temps (il n'est plus requis de visiter les outils un à un) et au fait qu'ils dispensent l'usager de la nécessité de s'initier aux modalités d'utilisation de chacun des sites à interroger – entreprise qui s'avère parfois laborieuse en ce qui concerne les modes de recherche avancée. L'utilisation de ces métamoteurs fait face, toutefois, à certains problèmes pratiques. Tout d'abord, il s'avère impossible pour ces logiciels d'exploiter les fonctionnalités avancées des outils de recherche, précisément parce que la syntaxe en est très variable. Leurs requêtes doivent demeurer suffisamment «basiques» pour être acceptées par tous les outils auxquelles elles sont envoyées, ce qui en diminue la puissance. Ensuite, comme le fait remarquer O. Andrieu :

> *[...] les métamoteurs font la synthèse de résultats fournis par plusieurs moteurs différents, classant chacun leurs résultats de façons différentes, sans utiliser les mêmes critères de pertinence. Une synthèse de documents classés de façons ainsi disparates est-elle si simple que cela à effectuer, et surtout, est-elle plus pertinente ? La question reste posée...* [www.abondance.com]

Par ailleurs, il souligne à juste titre les problèmes moraux que suscite l'apparition de ce genre d'outils[26] :

> *L'utilisation de ce type de métamoteurs engendre un autre problème de fond : quasiment tous les moteurs de recherche sur lesquels ils s'appuient se financent grâce aux bandeaux publicitaires qu'ils affichent. Or, les promoteurs de cette couche logicielle supplémentaire que sont les métamoteurs ne répercutent pas systématiquement (ou pas du tout) ces bandeaux, préférant même parfois proposer leurs propres annonces. Le recours à ces métamoteurs réduit donc de façon substantielle le nombre d'accès au moteur de recherche traditionnel, ce qui compromet ses recettes publicitaires et risque, à terme, de signer son arrêt de mort. D'autre part, se pose un problème d'éthique : est-il juste d'utiliser pour son propre compte les technologies et investissements mis en œuvre par d'autres sociétés, sans contrepartie financière ?* [www.abondance.com]

[26] Certains outils de recherche importants, par exemple NORTHERN LIGHT, empêchent d'ailleurs les métamoteurs d'accéder à leur site en raison du caractère parasitaire de ces derniers.

Ajoutons – ce qui n'étonnera personne – que les métamoteurs anglophones font généralement preuve de ce que l'on pourrait qualifier de «myopie anglo-saxonne» en ce qui concerne la liste des outils à sonder... Le concept de métamoteur, tout en étant intéressant en soi, demeure donc l'objet d'un certain nombre de réserves. Pris pour ce qu'il est, toutefois, et utilisé un peu à la manière d'un annuaire (pour des recherches larges et thématiques), ce type d'outil peut tout de même s'avérer d'une utilité non négligeable.

Quelques SUSI :

Nom	**URL**
ARI@NE	http://www.espace2001.com/moteur
COPERNIC	http://www.copernic.com/fr/
DEBRIEFING	http://www.debriefing.com/france/
DOGPILE	http://www.dogpile.com/
INFERENCE FIND	http://www.infind.com/
INFERENCE FIND version française	http://www.infind.com/infind_fr/
METACRAWLER	http://www.metacrawler.com
SAVYSEARCH	http://www.savvysearch.com/

5. Les agents intelligents

Le concept d'agent intelligent recouvre des réalités nombreuses et diverses. Au sens large, les agents intelligents peuvent être définis comme des outils *«permettant d'automatiser, périodiquement ou à la demande, des tâches de façon transparente pour l'utilisateur qui bénéficie des résultats»* [Philippe Courtot, CEO de Verity, cité dans Careil et de Frémont, s.d.]. Dans le contexte plus spécifique de la recherche d'information, ces logiciels sont généralement dotés, à des degrés divers, des caractéristiques de base suivantes :

- L'automatisation et l'autonomie du fonctionnement ;
- La mobilité, c'est-à-dire l'aptitude à voyager sur les réseaux ;

- La capacité d'interaction avec des interlocuteurs humains ou mécaniques ;
- La capacité dynamique d'apprentissage.

Contrairement aux annuaires, aux moteurs et aux métamoteurs, les agents intelligents ne forment pas une classe clairement délimitée de SRI sur Internet. D'une part, ils sont souvent incorporés aux outils des autres groupes : les robots que nous avons évoqués précédemment constituent, en fait, un type élémentaire d'agent intelligent[27], tout comme les métamoteurs sont une application de cette technologie. D'autre part, les agents varient énormément entre eux au niveau de leurs caractéristiques spécifiques. P. Nygren [http://perso.club-internet.fr/nygren/] propose un classement des agents intelligents susceptibles d'être rencontrés sur Internet en fonction de leur mission, c'est-à-dire de leur capacité à exécuter des tâches particulières :

5.1. Agents de recherche d'information

5.1.1. Fédérateurs de recherche

Ces outils accomplissent de nombreuses tâches : recherche d'information simultanée sur plusieurs outils ; rapatriement et indexation des pages en local ; classement et gestion des informations ; élimination des doublons ; création de résumés ; surveillance des modifications de sites selon une périodicité paramétrable, etc. Les métamoteurs s'inscrivent dans cette catégorie.

5.1.2. Agents sectoriels

Ce sont des fédérateurs de recherche spécialisés dans un domaine précis, par exemple les sciences et techniques, la finance ou la littérature. Les agents sectoriels consultent des outils de recherche spécialisés dans les domaines concernés.

[27] D'ailleurs, le terme *robot* sert parfois aussi à désigner les agents intelligents.

5.2. Agents pour la consultation hors ligne

Ces outils permettent d'aspirer un site Web (texte et images) pour le recopier sur un poste local, en respectant l'arborescence du site d'origine. Il est habituellement possible de spécifier le niveau de profondeur des pages à inclure. Exemple : WEBWHACKER.

5.3. Agents autonomes

Ces agents ont pour mission de dépister «toutes» les pages susceptibles de répondre à une requête donnée (ils peuvent éventuellement prendre l'initiative d'enrichir cette dernière). Ils filtrent et analysent les documents trouvés, ne rapatriant que ceux qui sont réellement pertinents. Ils permettent souvent l'emploi du langage naturel. Exemple : DIGOUT4U.

5.4. Agents pour le commerce électronique

5.4.1. Assistants d'achat (*shopbots*)

Destinés aux consommateurs, ils enregistrent les préférences de ces derniers et visent à faciliter la sélection de boutiques virtuelles, de marques ou de produits. Ils peuvent ainsi parcourir les galeries marchandes du Web à la recherche d'un produit ou service spécifique ; comparer les prix ; dresser un tableau récapitulatif des offres disponibles ; recommander des produits ou même procéder directement à l'achat. Ces assistants peuvent être généralistes (exemple : SHOPPING EXPLORER) ou porter sur un domaine d'activité précis (exemple : PRICELINE pour les billets d'avion, chambres d'hôtel, etc.).

5.4.2. Agents d'analyse de la demande

Destinés aux commerçants, ils permettent de mieux connaître la demande et les consommateurs, pour une meilleure gestion des profils clients et la personnalisation de l'offre. Exemple : SELECTCAST.

Certes, pour le moment, les agents intelligents ont quelque peu usurpé leur nom… Ils deviennent, toutefois, de plus en plus efficaces, et on commence à voir se réaliser les prédictions formulées à leur sujet par J. de Rosnay en 1995 :

> *Les agents vont rapidement constituer une nouvelle population d'êtres virtuels. Comme des virus informatiques contrôlés, ils vont se reproduire, constituer des groupes, des «cultures». Représentants de la vie artificielle, ils vont progressivement coloniser des continents entiers du cyberespace. Des agents travailleront en équipe. Munis de «permis» et «d'autorisations» (d'achat, de négociation), ils pourront se partager un travail et comparer des informations ; leurs compétences s'accroissant au fur et à mesure de leurs travaux de recherche ou de préparation de dossiers. Circulant sur les réseaux, ces «intra-terrestres» d'un nouveau genre offriront leurs services. Grâce aux algorithmes génétiques, des programmes d'agents pourront muter, s'autosélectionner, évoluer pour résoudre des problèmes de plus en plus complexes. Leur valeur augmentera à la bourse des emplois électroniques. Mais les agents représenteront aussi des dangers potentiels. Sachant tout sur les habitudes, préférences ou secrets de leurs patrons, ils pourront être kidnappés sur les réseaux et utilisés contre leurs employeurs.* [de Rosnay 1995]

Les agents intelligents peuvent rencontrer les suffrages de nombreuses clientèles. Pour les particuliers, ils peuvent agir comme guides vers les informations recherchées sur le Web, comme assistants d'achat ou encore pour la gestion documentaire personnelle (lorsque l'agent est configuré pour effectuer des recherches sur le poste même de l'utilisateur). Beaucoup d'agents évoluent au fil du temps, s'adaptent aux circonstances, prennent des décisions et enrichissent eux-mêmes leur comportement sur la base des observations qu'ils effectuent : ils peuvent donc étudier les réactions de leur «propriétaire» face aux premiers résultats de leur travail et modifier leurs activités en conséquence afin de mieux coller aux attentes de ce dernier.

Pour les entreprises, les agents intelligents s'avèrent également d'une utilité appréciable dans un contexte de veille concurrentielle et technologique sur Internet :

«L'agent intelligent est l'outil de prédilection du cyber-veilleur. De façon transparente ou active, il est obligé de passer par lui pour retrouver l'information pertinente au milieu de ce cyber-fatras.» [Careil et de Frémont, s.d.] Les agents intelligents permettent, en effet, aux veilleurs d'économiser du temps tout en effectuant un parcours exhaustif des sources d'information. Il devient aussi possible pour les entreprises de mettre en place des pratiques de surveillance systématique de l'environnement : en maintenant des agents en recherche permanente sur le site d'un concurrent, aucun des mouvements économiques et stratégiques de celui-ci n'échappera aux utilisateurs desdits agents. Les agents intelligents pourront également être utilisés, enfin, pour élaborer des bases de données thématiques ou pour analyser des serveurs hors ligne.

Quelques agents intelligents :

Nom	**URL**
AURESYS	http://ms161u06.u-3mrs.fr/hom.html
DIGOUT4U	http://www.arisem.com/index_fr.html
INFORIAN QUEST 98	http://www.inforian.com
MATA HARI	http://www.thewebtools.com
NEARSITE	http://www.nearsite.com
PRICELINE	http://www.priceline.com
SELECTCAST	http://www.aptex.com/products-selectcast.htm
SHOPPING EXPLORER	http://www.shoppingexplorer.com
WEBWHACKER	http://www.bluesquirrel.com/products/whacker/whacker.html
WEBZINGER	http://www.webzinger.com

6. Conclusion de la seconde partie

Sur Internet, le processus de repérage de l'information est confronté à de multiples difficultés, certaines spécifiques (comme l'instabilité des ressources), d'autres

communes à tous les systèmes d'information (par exemple, les problèmes découlant des ambiguïtés langagières). Les outils de recherche développés pour tenter de gérer cette situation sont très nombreux à l'heure actuelle ; ils continuent de se multiplier à un rythme effréné et il n'est sans doute pas exagéré de prétendre qu'il en apparaît de nouveaux presque tous les jours.

Face à un tel foisonnement, l'internaute moyen est souvent tenté de s'en tenir à la consultation d'un service ou deux parmi les plus connus, tels ALTAVISTA ou YAHOO!. Pourtant, il est au contraire impératif, lorsque l'on mène des recherches d'information sur Internet, de ne pas se cantonner à un seul outil ni même à un seul genre d'outils. Ici aussi, la complémentarité est le maître mot : aucun outil de recherche n'offre de couverture parfaitement exhaustive ; en outre, il semble que les recoupements entre les portions d'Internet couvertes par les différentes bases d'outils de même type demeurent assez minimes, bien qu'il soit fort difficile d'évaluer la situation à ce niveau. Comme, par ailleurs, les différents types d'outils ont été conçus pour répondre à des besoins distincts (recherches simples, générales ou thématiques pour les annuaires et métamoteurs ; recherches complexes ou pointues pour les moteurs), il s'avère beaucoup plus judicieux d'employer en parallèle plusieurs outils, sans pour autant tomber dans la surenchère : deux ou trois outils de chaque type suffisent généralement.

Comme les outils sont de valeur parfois très inégale, les choix de l'usager se révèlent lourds de conséquences. Outre les mesures de rappel et de précision (malgré les réserves précédemment évoquées), les indicateurs suivants peuvent s'avérer utiles lorsque vient le moment d'évaluer un outil de recherche :

- La crédibilité de l'organisme de maintenance ;
- La taille de la base de données, l'ampleur et l'objectivité de la couverture, les modalités d'ajout (notamment la possibilité de soumission de sites) ;
- La fréquence de mise à jour ;
- Les fonctionnalités disponibles pour les modes de recherche simple et avancée. Il convient de vérifier tout particulièrement :

- les domaines de recherche : possibilité d'effectuer des recherches limitées par champ (titre, intitulé de l'URL, liens hypertextuels, etc.), par genre de ressources (forums Usenet, adresses de courriel, pages personnelles, etc.) ou par type de fichiers (texte, image, fichiers audio ou vidéo), possibilité de faire des requêtes sur des intervalles de dates ou des noms de personne.
- les modes d'interrogation : possibilité de requêtes booléennes ou en langage naturel, sensibilité à la casse et aux caractères diacritiques, prise en compte de la proximité et de l'ordre des mots.

- l'affichage des résultats : facilité de consultation, possibilité de configurer la quantité de résultats à afficher et le format d'affichage, présence d'un pourcentage de pertinence par rapport à la requête, critères de tri disponibles.
- éventuellement, la possibilité d'accéder à l'historique des recherches et celle de pratiquer des interrogations récurrentes (c'est-à-dire d'effectuer une nouvelle recherche à l'intérieur des résultats d'une requête précédente).

- La rapidité de fonctionnement ;
- La facilité globale d'utilisation et la convivialité ;
- La présence de procédures d'aide claires et détaillées. En effet, cet aspect ne doit pas être négligé, comme le souligne J.-N. Plourde :

> *La documentation pour les services de repérage aide les utilisateurs à atteindre deux objectifs. Le premier est d'évaluer la pertinence de la base, c'est-à-dire sa nature (objets répertoriés), ses objectifs, son autorité, etc. Le second est la maîtrise et l'utilisation efficace des services de repérage et la vérification du comportement de ces services (obtient-on les résultats escomptés ?).* [Plourde 1996]

- L'originalité de l'outil ;
- Les services complémentaires : par exemple, les différentes ressources inhérentes aux sites de type portail[28], la possibilité de traduire

[28] Les *portails* sont des sites qui tentent de se positionner comme point d'entrée de l'internaute sur le Web. La plupart des outils de recherche sur Internet évoluent actuellement vers ce type de services. En addition à la fonction de recherche d'information proprement dite (et parfois au détriment de la qualité de celle-ci...), ils proposent désormais tout un éventail de services supplémenaires : actualités, dépêches d'agence, météo, cours de la Bourse, résultats sportifs, horoscope,

automatiquement les documents repérés (comme chez ALTAVISTA ou INFOSEEK), etc.

-

Il faut rappeler que les outils de recherche, même pris dans leur ensemble, ne peuvent rendre compte de tout ce qui se trouve sur Internet. Comme nous l'avons mentionné, plusieurs facteurs expliquent cette situation : immensité et métamorphoses du Réseau, sites non trouvés ou non explorés en profondeur, difficultés d'accès (présence de *firewalls*[29], sites interdits aux robots) et de traitement, censure, etc.

Une ressource non recensée demeurant quasi impossible à découvrir (à moins de suivre un lien ou d'en connaître l'URL d'avance...), les outils de recherche, en dépit de leurs lacunes actuelles, restent toutefois la meilleure façon d'exploiter l'information disponible sur Internet. Les conseils suivants sont susceptibles d'en optimiser l'emploi :

- Résumer au préalable son besoin d'information sous forme d'une phrase, puis identifier les principaux concepts ayant trait à la requête, en déterminant les termes les plus significatifs (plusieurs, de préférence). Les mots clés retenus doivent, dans la mesure du possible, s'avérer «discriminants», c'est-à-dire être rares ou inhabituels. Les mots trop communs sont à éviter absolument, de même, naturellement, que les fautes d'orthographe et de frappe. Il faut également songer à d'éventuels synonymes et traductions.
- Il est impératif de bien connaître le fonctionnement des outils employés en ce qui a trait au type d'indexation effectuée, aux domaines de

---

téléchargement de logiciels, petites annonces, dossiers spéciaux (ex. : crise du Kosovo, rapport Starr), Pages Jaunes et Blanches, bavardage en direct, adresse gratuite de courriel et hébergement de pages Web, offres d'emploi, calendriers et agendas en ligne, dictionnaires électroniques, enchères en ligne, envoi de cartes «postales», mise en place de «filtres familiaux» (tels l'*AV Family Filter* d'ALTAVISTA) destinés à bloquer – en principe – l'accès aux pages à contenu disgracieux, horaires des programmes de télévision, état du trafic sur les routes, personnalisation du site (c'est-à-dire l'occasion pour l'usager de configurer ses préférences d'interface), possibilité d'installer en local l'outil pour les recherches internes sur un site, etc. Les choix sont quasi infinis et la créativité des concepteurs semble sans bornes...

[29] L'Office de la langue française du Québec définit le *firewall* ou *coupe-feu* comme un «dispositif informatique qui permet le passage sélectif des flux d'information entre un réseau interne et un réseau public, ainsi que la neutralisation des tentatives de pénétration en provenance du réseau public» [http://www.olf.gouv.qc.ca/].

recherche, à la formulation des requêtes (notamment les opérateurs à utiliser et les questions de majuscules et de diacritiques), aux options d'affichage, etc.

- Pour une requête en français, il est souvent préférable de s'adresser en priorité à des outils disponibles en langue française, particulièrement en ce qui concerne les annuaires.
- Si une requête donnée demeure infructueuse avec les outils usuels :

- recourir aux métamoteurs pour croiser les recherches ;
- se servir d'agents intelligents ;
- rechercher dans les *foires aux questions* (FAQ) ;
- utiliser des *newsgroups* judicieusement choisis pour poser la question ;
- tenter une nouvelle requête à l'aide de mots clés plus génériques.

- Si, au contraire, une requête particulière s'avère trop fructueuse :

- ajouter un ou plusieurs mot(s) clé(s) supplémentaire(s) ;
- pour les outils qui le permettent, recourir aux opérateurs booléens, notamment à la recherche de locutions ;
- exploiter les possibilités de recherche avancée ;
- tenter une nouvelle requête à l'aide de mots clés plus spécifiques.

- Il ne faut pas négliger les outils de recherche spécialisés, dont on a souvent intérêt à se faire des signets.
- On peut également recourir avec profit à ce que P. Nygren [http://perso.club-internet.fr/nygren/] appelle la reverse psychology : cette technique consiste, à partir d'un site jugé pertinent, à rechercher systématiquement toutes les pages possédant un lien hypertextuel pointant vers son URL. Ce résultat peut être atteint soit par l'emploi des commandes avancées des moteurs de recherche soit par le recours aux annuaires, en retrouvant la ou les catégorie(s) où le site en question a été référencé.

•

Troisième partie :

du comportement des SRI sur Internet lors de quelques requêtes-test

1. Les outils retenus

Nous avons examiné 12 outils lors de nos investigations :

- Trois annuaires : Ctrouve.com, Nomade, Yahoo!
- Sept moteurs : AltaVista, Ecila, Excite, HotBot, Infoseek, Lycos, Voila
- Un métamoteur : COPERNIC 99
- Un agent intelligent : DIGOUT4U

La sélection des annuaires et des moteurs s'est faite sur la base de leur notoriété ; COPERNIC 99 et DIGOUT4U, pour leur part, ont été choisis parce qu'une version gratuite était disponible en téléchargement[30]. Dans le cas des outils internationaux, nous avons utilisé la version de langue française lorsqu'elle était disponible, que ce soit sous la forme d'une base de données complètement distincte (YAHOO! FRANCE) ou uniquement d'une interface adaptée (VOILA, INFOSEEK FRANCE, LYCOS FRANCE). Des fiches signalétiques détaillées sont disponibles en annexe pour chacun de ces outils. Les 11 premiers correspondant globalement aux caractéristiques de leurs classes respectives telles que nous les avons esquissées dans la seconde partie de ce travail, nous nous bornerons ici à apporter quelques précisions supplémentaires sur le fonctionnement de l'agent intelligent DIGOUT4U.

DIGOUT4U est un produit développé par la compagnie ARISEM[31]. Ce logiciel est dédié à la recherche de documents sur Internet. Pour ce faire, il s'appuie sur la technologie L4U (*Language4U*), relative à la constitution de bases de connaissances lexico-sémantiques multilingues. Grâce à une analyse sémantique et pragmatique des textes, ce système de compréhension du langage naturel fait en sorte que l'idée sous-jacente à une requête est reconnue sous toutes ses formes d'expression (le système gère indifféremment le français et l'anglais, ce qui permet, par exemple, de

[30] À l'origine, nous avions prévu d'inclure dans nos examens plusieurs autres outils de type agent, notamment INFOSCAN (pour le filtrage du courriel : http://www.machinasapiens.com/francais/produits/infoscan/infoscan.html) et NOMINO (SRI dont le fonctionnement est basé sur une analyse morpho-syntaxique très poussée : http://www.ling.uqam.ca/nomino/). Nous avons malheureusement dû y renoncer à cause de problèmes d'équipement informatique.

[31] Pour *ARtificial Intelligence & SEMantics*.

formuler une requête dans une langue et de repérer, si on le désire, des documents des deux langues).

Cette capacité se fonde sur des bases de connaissances qui regroupent les termes des langues cibles ainsi que la sémantique qui s'y rapporte. DIGOUT4U exploite ainsi une base de connaissances généraliste («*The Genus*»), qui regroupe environ 35 000 termes français, 17 000 termes anglais et 70 000 règles (formes fléchies, rattachement des termes aux idées, désambiguïsation, idées associées, hyperonymie sémantique, etc.). Il est possible d'y incorporer des bases de connaissances spécialisées supplémentaires, relatives par exemple à un métier ou à un domaine d'activité. En outre, lorsqu'un élément n'est pas compris par le système, l'utilisateur peut illico en préciser simplement et rapidement la définition. Cet ajout sera ensuite conservé.

Une recherche avec DIGOUT4U se déroule ainsi :

- L'usager formule une requête en langue naturelle (la saisie d'une simple suite de mots clés est déconseillée) ;
- Le logiciel identifie les concepts de la requête et envoie ses agents interroger les outils de recherche du Web ;
- Les premières pages de résultats rapatriées sont «lues» par DIGOUT4U qui les classe et leur attribue une note de pertinence en comparant la requête au contenu sémantique de la page ;
- Les agents poursuivent leur investigation en profondeur en suivant les liens hypertextuels contenus dans les documents ayant obtenu un score de pertinence élevé.

Quand la recherche est terminée et après des fonctions d'édition classiques, on peut exporter comme résultat final une liste de références classées par pertinence et incluant ou non des résumés.

Soulignons, pour clore cette brève présentation, que le logiciel dispose d'une fonction particulièrement intéressante qui permet, pour chaque document repéré, de demander l'affichage d'une fenêtre spéciale contenant, d'une part, une courbe de pertinence (représentant la «distribution» de l'information pertinente au fil du texte)

et, d'autre part, les extraits (phrases) pertinents du document (sélectionnés par rapport à la requête initiale). Comme on le voit dans la figure ci-après, l'usager est libre de configurer le niveau de restitution de l'information. Quand le curseur est en haut, l'extrait est très court et concerne le ou les «pic(s) de pertinence». Plus le curseur est abaissé, et plus les extraits sont longs et nombreux.

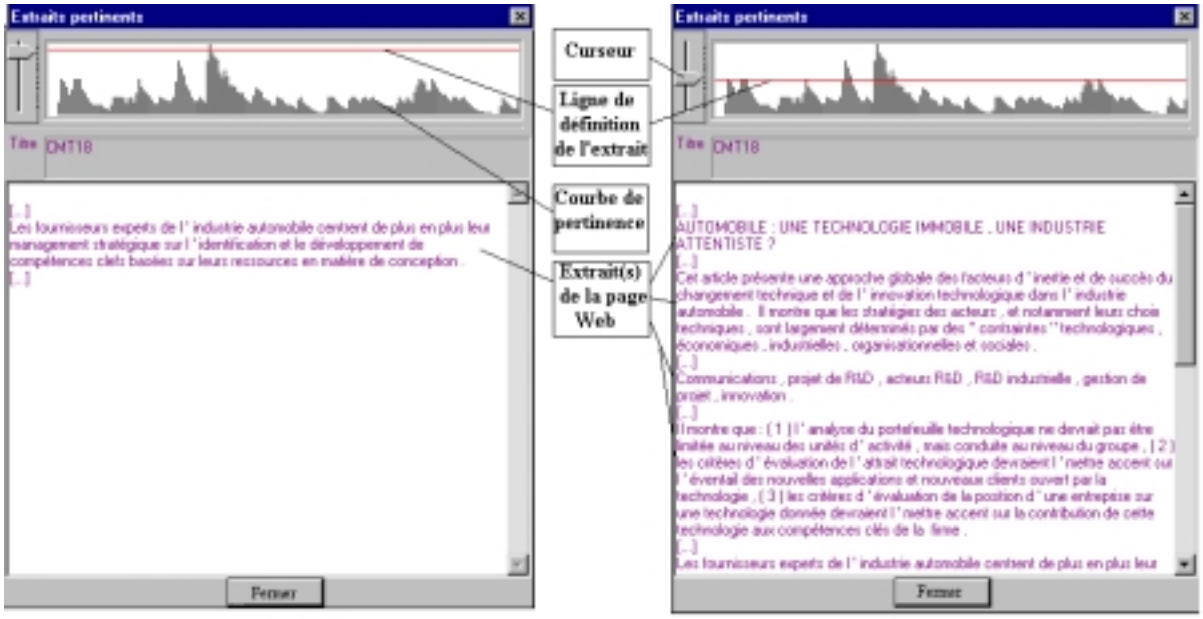

Figure 2 : la fonction «Extraits pertinents» dans DIGOUT4U

Source : guide d'utilisation de DIGOUT4U

2. Les tests effectués

Les tests présentés dans cette section ont été réalisés entre le 27 juillet et le 23 août 1999.

Quelques remarques préliminaires :

- L'annuaire NOMADE et le moteur ECILA ne précisent pas le nombre total de réponses à une requête lorsque ce dernier dépasse un certain plafond (respectivement fixé à 150 et 200). Les résultats sont donc parfois plus flous en ce qui les concerne.

- Les annuaires YAHOO! et NOMADE redirigeant les requêtes infructueuses vers des outils de type moteur (respectivement INKTOMI et ALTAVISTA), nous incluons les résultats en provenance de ces outils lorsque nécessaire.
- Pour toutes les requêtes adressées à l'annuaire CTROUVE.COM, nous avons dû nous en tenir au mode d'interrogation par défaut (soit un OU implicite entre les mots), aucune autre option de recherche n'étant disponible.
- Afin de confirmer les résultats recueillis, chaque requête effectuée a été systématiquement resoumise une seconde fois, à quelques minutes d'intervalle.

Pour chacun des tests présentés, les résultats intégraux sont fournis à la suite de nos commentaires d'introduction et d'analyse. Les requêtes effectuées avec DIGOUT4U font l'objet d'une section ultérieure.

2.1. Les modes de formulation d'une requête

À l'exception de l'annuaire CTROUVE.COM, tous les outils de notre échantillon proposent à l'internaute des alternatives au mode de recherche «par défaut» (qui relie les mots de la requête par ET ou OU, selon les cas). Certains SRI permettent ainsi l'emploi d'opérateurs et de modificateurs directement sur la ligne de commande ; d'autres présentent les options disponibles sous forme de menus déroulants, de cases à cocher, de boutons radio, etc. ; d'autres enfin (soit la plupart) choisissent de mêler ces deux approches.

Deux possibilités se révèlent particulièrement répandues : la recherche de locutions et l'emploi des opérateurs booléens. Disponible dans tous les outils – via l'utilisation de guillemets ou de tirets, ou encore le choix d'options comme *le segment*, *la phrase exacte* ou *rechercher avec l'expression* –, la recherche de locutions, ainsi que nous l'avons expliqué dans la première partie de ce travail, permet la recherche exacte d'une série ordonnée de mots adjacents. Les résultats de recherche sont ainsi épurés des documents où les mots de la requête apparaissent uniquement séparés les uns des autres.

Outre son utilité pour délimiter les mots composés, cette méthode s'avère particulièrement efficace pour les recherches sur des noms propres. Par exemple, en recherchant «*Ronald Reagan*», on évite les références relatives à des particuliers nommés *Ronald* durant la *période Reagan*. En anglais, l'emploi de guillemets autour d'un nom comme *River Phoenix* restreindra la recherche aux documents consacrés à cet acteur américain, en excluant tout ce qui traite des rivières près de la ville de Phœnix, en Arizona. En français, un exemple du même type pourrait être une recherche consacrée à l'acteur *Pierre Brasseur*, où la définition comme locution empêchera le repêchage de sites consacrés à la bière ou à la gemmologie. Selon l'aide en ligne du moteur HOTBOT, la recherche sous forme de locution permettrait, en fait, de diviser par 15 le volume des réponses obtenues.

L'utilisation des opérateurs booléens, pour sa part, s'effectue à travers des choix dans des formulaires (*tous les mots*, *au moins un mot*, etc.) ou via la saisie sur la ligne de commande des classiques AND, OR et NOT, avec quelques nuances selon les outils : certains exigent, en effet, l'emploi de majuscules ou la réécriture de NOT en AND NOT. Les équivalents français sont utilisés dans certains SRI francophones, par exemple ECILA.

Une stratégie reliée consiste à permettre le recours aux signes + – pour respectivement forcer et exclure la présence d'un mot dans un document. Ces fonctions booléennes simplifiées peuvent s'avérer utiles pour mieux «filtrer» les réponses : par exemple, une requête sur le mot *pêche* peut judicieusement être enrichie par – *fruit* lorsqu'on ne s'intéresse qu'aux poissons... Elles présentent d'ordinaire l'avantage de ne pas désactiver l'algorithme de pertinence de l'outil de recherche, comme cela peut être le cas avec l'utilisation des opérateurs AND, OR et (AND) NOT.

Selon certains experts, tels D. Sullivan et S. Feldman, l'emploi des signes + – et des guillemets serait d'ailleurs plus approprié que la construction de requêtes

traditionnelles avec des opérateurs booléens pour la recherche sur les SRI du Web[32] :

> *For example, enter a string of words and most search engines will naturally try to find them in close proximity to each other. This eliminates a need to specify a proximity command like NEAR. Likewise, do a search at AltaVista or Google, and they will automatically try to detect phrases in your queries and give you pages that contain those phrases. By entering a complex Boolean command, you are searching in a way the search engines are not designed for.* [Sullivan 1999]

Néanmoins, comme le font remarquer X. Dong et L.T. Su, *«there is no consensus on which way of submitting a query can improve precision in searching and what kind of search technique should be selected when using different search engines.»* [Dong & Su 1997] Selon ces auteurs, *«a good search engine should be able to retrieve highly relevant results no matter what kind of query is submitted.»* [Idem] C'est la raison pour laquelle, pour chacun des outils de notre échantillon, nous avons choisi de mettre en parallèle les différentes manières d'exprimer une requête, pour ensuite comparer entre elles les quantités de résultats obtenues[33].

Nous avons retenu à cette fin la requête *beau dommage*, qui nous semblait intéressante parce que, tout en étant constituée de mots individuels très courants, elle peut référer, en tant qu'expression, à la fois à une locution du lexique commun de la langue et à un nom propre (groupe de musique québécois). Pour chaque outil, les différentes manières de repérer des documents contenant simultanément ces deux mots ont été testées, en plus du mode de recherche par défaut (qui, comme nous l'avons mentionné, est parfois un OU). Nous avons ainsi recouru aux guillemets, à l'opérateur AND ou ET, de même qu'à certaines options propres à chaque outil (du genre *laisser le moteur décider* pour YAHOO!). Lorsqu'un outil permettait deux procédures apparemment équivalentes, par exemple l'emploi de guillemets et

---

[32] Selon S. Feldman, cette situation tient au fait que les requêtes de type booléen ont été développées pour des SRI qui ne disposaient pas de la puissance de traitement exigée par des requêtes formulées davantage en langage naturel, ce qui n'est plus le cas des outils contemporains.

[33] Étant donné l'ampleur limitée de ce travail, nous avons choisi d'analyser les données de ce test uniquement en ce qui concerne le rappel, mais il aurait été tout aussi intéressant de comparer les différents modes de requête au point de vue de la précision.

l'option *phrase exacte*, les deux procédures ont été vérifiées. Enfin, nous avons testé au passage les alternatives à l'emploi de guillemets pour définir manuellement une locution lorsque les outils en proposaient (par exemple, l'utilisation de tirets entre les mots pour ALTAVISTA et INFOSEEK).

Au vu des disparités et des contradictions – pour ne pas dire des aberrations... – relevées lors de l'analyse des données, il s'avère bien ardu de tirer des conclusions d'ensemble des résultats de ce test. Ainsi, d'un côté, nous avons constaté que des modes de recherche équivalents à première vue et considérés comme tels par certains outils produisent avec d'autres outils des sommes différenciées :

INFOSEEK

+beau +dommage	860
beau AND dommage	860

10/08/1999 sur *tout le Web*

LYCOS

+ beau + dommage	442
beau AND dommage	442

04/08/1999 sur le *Web mondial*

ALTAVISTA

+beau +dommage	15 960
beau AND dommage (recherche avancée)	3 295

23/08/1999 (*any language*)

EXCITE

"beau dommage"	194
beau dommage (*segment*)	194

27/07/1999 en *français* sur le *Web mondial*

LYCOS

"beau dommage"	442
beau dommage (recherche avancée : *la phrase exacte*)	182
beau dommage (recherche avancée : *tous les mots adjacents*)	183

04/08/1999 sur le *Web mondial*

Dans le cas de l'opposition +/AND, peut-être faut-il voir dans ces chiffres un reflet du fait que l'emploi d'opérateurs booléens désactive chez certains outils, ainsi que nous l'avons précisé plus haut, l'algorithme de pertinence qui permet l'affichage d'un très grand nombre de réponses avec les «meilleures» en tête de liste.

D'un autre côté, à l'inverse, des modes de recherche de prime abord bien distincts ont aussi produit à la fois des sommes identiques et des sommes différenciées (les modes de recherche par défaut et par locution, entre autres, ont souvent tendance à se confondre) :

EXCITE

beau dommage	41 717
"beau dommage"	194
beau dommage (*segment*)	194
beau dommage (*le(s) mot(s)*)	1 935

27/07/1999 en *français* sur le *Web mondial*

VOILA

beau dommage	62 121
beau dommage (recherche avancée : *la phrase*)	382
beau dommage (recherche avancée : *les mots*)	1 152

28/07/1999 sur le *Web mondial*

NOMADE

beau dommage	2
"beau dommage"	2
beau dommage (recherche avancée : *tous les mots*)	2

30/07/1999 dans *Tout Nomade*

ALTAVISTA

beau dommage	431
"beau dommage"	431

23/08/1999 (*any language*)

INFOSEEK

beau dommage	155
"beau dommage"	155

10/08/1999 sur *tout le Web*

LYCOS

beau dommage	442
"beau dommage"	442
+ beau + dommage	442
beau AND dommage	442
beau WITH dommage[34]	442
beau ADJ dommage	442
beau BEFORE dommage	442
beau OADJ dommage	442

[34] Selon l'aide en ligne de LYCOS, l'opérateur WITH est équivalent à l'opérateur AND. OADJ, pour sa part, équivaut à l'opérateur d'adjacence ADJ, mais introduit une notion additionnelle de prise en compte de l'ordre d'apparition des termes.

En outre, on constate également des différences entre les quelques outils sensibles à la casse : HOTBOT se conforme au comportement attendu[35], de même qu'ALTAVISTA – sauf en ce qui concerne le mode de recherche par défaut, où *beau dommage* obtient moins de réponses que *Beau Dommage* (nous supposons ici que la recherche par locution respecte la casse de manière absolue, ce qui explique que *"Beau Dommage"* repère plus d'occurrences que *"beau dommage"*). Par contre, INFOSEEK semble ne tenir compte de la casse que lors de requêtes impliquant des signes + ou l'opérateur AND :

HOTBOT

beau dommage (*all the words* par défaut)	1 250
Beau Dommage (*all the words* par défaut)	290
"beau dommage"	390
"Beau Dommage"	220
+beau +dommage	1 250
+Beau +Dommage	290
beau dommage (*exact phrase*)	390
Beau Dommage (*exact phrase*)	220
beau AND dommage (*Boolean phrase*)	1 270
Beau AND Dommage (*Boolean phrase*)	290
beau dommage (recherche avancée : *must contain the words*)	1 250
Beau Dommage (recherche avancée : *must contain the words*)	290
beau dommage (recherche avancée : *must contain the phrase*)	390
Beau Dommage (recherche avancée : *must contain the phrase*)	220

23/08/1999 (*any language*)

ALTAVISTA

beau dommage	431
Beau Dommage	467
"beau dommage"	431
"Beau Dommage"	467

[35] La sensibilité à la casse se rapporte à la prise en compte de la distinction majuscules/minuscules. Cette question est abordée plus en détail dans un test subséquent. Pour le moment, mentionnons simplement que, en principe, une requête entièrement en minuscules est censée repêcher toutes les occurrences du motif (peu importe la casse), alors qu'une requête comportant des majuscules n'est repérée que telle quelle.

+beau +dommage	15 960
+Beau +Dommage	596
beau AND dommage (recherche avancée)	3 295
Beau AND Dommage (recherche avancée)	596
beau NEAR dommage (recherche avancée)	549
Beau NEAR Dommage (recherche avancée)	470

23/08/1999 (*any language*)

INFOSEEK

beau dommage	155
Beau Dommage	155
"beau dommage"	155
"Beau Dommage"	155
+beau +dommage	860
+Beau +Dommage	176
beau AND dommage	860
Beau AND Dommage	176

10/08/1999 sur *tout le Web*

Par ailleurs, comme on pouvait s'y attendre, le mode de recherche par locution abaisse sensiblement le nombre des résultats obtenus. Cette tendance s'observe pour les moteurs ECILA, EXCITE, VOILA et HOTBOT (quoique dans une moindre mesure pour les requêtes impliquant des majuscules, dans ce dernier cas). On la retrouve également, de manière moins marquée, chez COPERNIC. Les résultats fournis par le moteur LYCOS à cet égard sont surprenants : l'emploi de guillemets ne modifie pas le nombre de résultats obtenus par la requête par défaut, alors que les options *la phrase exacte* et *tous les mots adjacents* diminuent effectivement ce nombre, et que le choix *tous les mots dans l'ordre* l'augmente :

LYCOS

beau dommage	442
"beau dommage"	442
beau dommage (recherche avancée : *la phrase exacte*)	182
beau dommage (recherche avancée : *tous les mots dans l'ordre*)	452
beau dommage (recherche avancée : *tous les mots adjacents*)	183

04/08/1999 sur le *Web mondial*

Quant aux différentes façons de formuler manuellement des locutions chez INFOSEEK et ALTAVISTA, elles se sont bien avérées équivalentes entre elles.

Ajoutons, à titre d'information, que les diverses possibilités de recherche dont il a été question ici demeurent l'apanage d'une minorité d'internautes : une étude menée sur plus de 50 000 requêtes soumises par quelque 18 000 usagers d'EXCITE a démontré que l'opérateur AND était utilisé dans moins de 7% des requêtes et les guillemets et signes + – dans moins de 6% des requêtes[36].

CTROUVE.COM

beau dommage	110

28/07/1999

NOMADE

beau dommage	2
"beau dommage"	2
beau dommage (recherche avancée : *tous les mots*)	2

30/07/1999 dans *Tout Nomade*

YAHOO!

beau dommage	62 catégories / 720 sites
"beau dommage"	0 (173 avec INKTOMI)
+beau +dommage	0 (1 744 avec INKTOMI)
beau dommage (recherche avancée : *rechercher avec l'expression*)	0 (173 avec INKTOMI)
beau dommage (recherche avancée : *rechercher avec tous les mots*)	0 (1 744 avec INKTOMI)
beau dommage (recherche avancée : *laisser le moteur décider*)	62 catégories / 720 sites

29/07/1999

ALTAVISTA

beau dommage	431
Beau Dommage	467
"beau dommage"	431
"Beau Dommage"	467
beau-dommage	431
Beau-Dommage	467
beau,dommage	431
Beau,Dommage	467
beau.dommage	431
Beau.Dommage	467
beau/dommage	431
Beau/Dommage	467
beau_dommage	431

[36] Jansen, B.J.; Spink, A.; Bateman, J. et T. Saracevic. «Real life information retrieval: a study of user queries on the Web». *SIGIR Forum*, 1998, 32 (1) : 5-17.

Beau_Dommage	467
+beau +dommage	15 960
+Beau +Dommage	596
beau AND dommage (recherche avancée)	3 295
Beau AND Dommage (recherche avancée)	596
beau NEAR dommage (recherche avancée)	549
Beau NEAR Dommage (recherche avancée)	470

23/08/1999 (*any language*)

ECILA

beau ET dommage	plus de 200
beau dommage (*phrase exacte*)	82
beau dommage (*tous les mots*)	plus de 200

29/07/1999

EXCITE

beau dommage	41 717
"beau dommage"	194
+beau +dommage	1 935
beau AND dommage	1 935
beau dommage (*segment*)	194
beau dommage (*le(s) mot(s)*)	1 935

27/07/1999 en *français* sur le *Web mondial*

HOTBOT

beau dommage (*all the words* par défaut)	1 250
Beau Dommage (*all the words* par défaut)	290
"beau dommage"	390
"Beau Dommage"	220
+beau +dommage	1 250
+Beau +Dommage	290
beau dommage (*exact phrase*)	390
Beau Dommage (*exact phrase*)	220
beau AND dommage (*Boolean phrase*)	1 270
Beau AND Dommage (*Boolean phrase*)	290
beau dommage (recherche avancée : *must contain the words*)	1 250
Beau Dommage (recherche avancée : *must contain the words*)	290
beau dommage (recherche avancée : *must contain the phrase*)	390
Beau Dommage (recherche avancée : *must contain the phrase*)	220

23/08/1999 (*any language*)

INFOSEEK

beau dommage	155

Beau Dommage	155
"beau dommage"	155
"Beau Dommage"	155
beau-dommage	155
Beau-Dommage	155
+beau +dommage	860
+Beau +Dommage	176
beau AND dommage	860
Beau AND Dommage	176

10/08/1999 sur *tout le Web*

LYCOS

beau dommage	442
"beau dommage"	442
+ beau + dommage	442
beau AND dommage	442
beau WITH dommage	442
beau ADJ dommage	442
beau BEFORE dommage	442
beau OADJ dommage	442
beau dommage (recherche avancée : *la phrase exacte*)	182
beau dommage (recherche avancée : *tous les mots*)	589
beau dommage (recherche avancée : *tous les mots dans l'ordre*)	452
beau dommage (recherche avancée : *tous les mots adjacents*)	183

04/08/1999 sur le *Web mondial*

VOILA

beau dommage	62 121
+beau +dommage (recherche avancée)	1 152
beau AND dommage	61 130
beau dommage (recherche avancée : *la phrase*)	382
beau dommage (recherche avancée : *les mots*)	1 152

28/07/1999 sur le *Web mondial*

COPERNIC

"beau dommage" (recherche rapide)	76
beau dommage (recherche rapide : *expression exacte*)	76
beau dommage (recherche rapide : *tous les mots*)	87

30/07/1999 sur le *Web*

2.2. Les ressources francophones

Nous avons ensuite vérifié la couverture spécifique du Web francophone chez les différents outils de notre échantillon. Ces derniers étant soit exclusivement de langue française (le moteur ECILA et les annuaires CTROUVE.COM, NOMADE et

YAHOO!), soit dotés d'options de recherche de restriction linguistique et/ou géographique permettant d'isoler des portions francophones d'Internet, nous avons voulu établir si des options comme *le Web français*, *le Web francophone* ou *les sites de langue française* prenaient en compte la «totalité» du contenu disponible en français ou si elles ne limitaient pas plutôt la recherche de l'usager aux URL en provenance de France, d'Europe francophone ou, au mieux, d'un ensemble prédéfini de pays de la Francophonie.

Pour ce faire, nous avons utilisé à nouveau la requête portant sur *beau dommage* en tant que locution, l'hypothèse sous-jacente étant qu'une grande quantité des URL potentiellement pertinentes pour cette recherche provient du Québec. Pour les outils qui le permettaient, nous avons cette fois fait porter la requête, selon les cas, soit sur une option prédéfinie (du type *Web francophone* par opposition à *Web mondial*), soit sur une restriction linguistique en français, soit enfin sur une restriction géographique (France, Québec, Canada). Nous avons ensuite soumis à tous les SRI une requête visant à repérer l'URL d'un organisme québécois, le Centre d'expertise et de veille Inforoutes et Langues (CEVEIL)[37].

Ici aussi, les données obtenues sont surprenantes à plusieurs égards. Tout d'abord, on peut remarquer que, pour deux des outils (les moteurs LYCOS et VOILA), les résultats repérés pour *beau dommage* en isolant les sites de langue française sont supérieurs en nombre à ceux générés en faisant porter la requête sur l'ensemble du Web :

LYCOS

"beau dommage"	442
	04/08/1999 sur le *Web mondial*
"beau dommage"	948
	04/08/1999 sur le *Web français*

VOILA

beau dommage (recherche avancée : *la phrase*)	382
	28/07/1999 sur le *Web mondial*
beau dommage (recherche avancée : *la phrase*)	545
	28/07/1999 sur le *Web francophone*

[37] http://www.ceveil.qc.ca .

Faut-il en déduire que ces outils ont élaboré des bases de données distinctes ou qu'ils appliquent un quota sur le nombre de pages en français sondées lors de recherches sur le Web mondial ? Par ailleurs, les résultats fournis par le moteur INFOSEEK soulèvent également, à première vue, quelques interrogations : si *tout le Web* affiche 155 occurrences de la locution *beau dommage* alors que *la France* n'en fournit aucune et *le Canada* 2 seulement, on peut se demander d'où proviennent les 153 autres... On pourrait, tout d'abord, être tenté de conclure soit que cette expression est sur-représentée dans les pages belges et suisses, soit qu'elle a fait l'objet d'un emprunt par quelque langue étrangère. Mais, après inspection des résultats, il appert que cette situation découle du fait qu'INFOSEEK affiche pour *tout le Web* des URL qui répondent à la requête sans égard à leur terminaison de nom de domaine (*.com*, *.edu*, *.net*, *.fr*, *.ca*, etc.), alors que les résultats sont limités aux URL en *.fr* pour *la France* et à ceux en *.ca* pour *le Canada*. Une situation similaire se présente pour le moteur EXCITE, qui propose à l'internaute le choix entre le *Web mondial* et le *Web français* dans sa page d'accueil. Puisqu'elle limite la recherche aux sites en *.fr*, l'expression *Web français* signifie ici «Web de France» et non «Web de la Francophonie», ce qui n'est pas évident de prime abord (à tout le moins pour les francophones hors de l'Hexagone...). On voit ici quel silence important peut découler d'une simple restriction géographique appliquée en toute bonne conscience par un internaute non averti. À noter aussi que l'on constate, pour ce test, des incohérences entre nos résultats et les affirmations officielles des concepteurs des divers outils : ainsi, l'opposition *Web mondial / Web français* existe également chez LYCOS, où l'option *Web français* est censée limiter la recherche aux URL en *.fr*, *.be* et *.ch*. Ceci n'a pas empêché cet outil de repérer l'URL du CEVEIL avec une recherche sur le *Web français* (ce qui n'a pas été le cas pour EXCITE).

Lors de l'examen des résultats pour la requête *ceveil*, nous avons pris soin de distinguer les SRI qui identifiaient le site lui-même (ou des sites québécois ou canadiens reliés) de ceux qui ne proposaient que des liens externes éloignés (par exemple, uniquement des sites français mentionnant le CEVEIL). Tous les outils de notre échantillon permettent de retrouver cette URL à partir de leur contenu

francophone – qu'il soit le seul disponible ou qu'il ait été cerné à travers la requête –, à l'exception d'EXCITE et des annuaires CTROUVE.COM et YAHOO!. Évidemment, dans ces deux derniers cas, la présence d'un site est conditionnelle à son inscription préalable, ce qui explique probablement ces résultats négatifs (nous y avons constaté, par ailleurs, l'inclusion de nombreux sites québécois). Ces résultats permettent de conclure globalement à un bon recensement de l'ensemble du contenu français d'Internet, bien que certains outils présentent le défaut important de baser leur classification linguistique sur les suffixes de domaine[38].

Parmi les autres curiosités constatées, mentionnons que, pour certains annuaires, l'attribution d'un site à un pays semble se baser uniquement sur les données indiquées par le webmestre lors de l'inscription – y compris lorsque ces dernières sont contredites par la terminaison de l'URL. Par exemple, dans CTROUVE.COM, plusieurs sites se terminant par *.ch* (soit la Suisse) mais comportant la mention *France* à l'indication du pays sont classés parmi les sites français... Un peu dans le même ordre d'idées, il est étonnant que l'annuaire NOMADE permette de repérer l'URL du CEVEIL pour une recherche sur le *Canada* mais non pour une recherche sur le *Québec* (d'autant plus que cette URL, comme beaucoup de sites québécois, présente la double terminaison *.qc.ca*).

Afin de faciliter les comparaisons, nous présentons ci-dessous à la fois les résultats de ce test et les données pertinentes recopiées du test précédent. Pour les outils ne permettant aucune des restrictions mentionnées plus haut, nous indiquons uniquement les résultats de la requête *ceveil*. Les chiffres fournis pour cette dernière requête incluent au moins une page en provenance du site lui-même, à moins d'indication contraire.

---

[38] Cette tactique est d'autant plus condamnable que les risques de bruit et de silence sont multiples : outre le fait que tous les sites de France (par exemple) ne se terminent pas par *.fr*, bon nombre de pages en français proviennent de pays officiellement non francophones. De plus, un site en *.ca* a plus de chance d'être en anglais qu'en français, un site en *.be* peut fort bien être en néerlandais et un site en *.ch* en allemand, etc. Puisque des technologies fort efficaces existent déjà qui permettent d'identifier automatiquement la langue d'une page Web (voir par exemple SILC – *Système d'Identification de la Langue et du Codage* – à l'URL http://www-rali.iro.umontreal.ca/ProjetSILC.fr.html), le recours à des méthodes de ce genre nous paraît nettement plus sûr et performant.

CTROUVE.COM

ceveil	0

28/07/1999

NOMADE

"beau dommage"	2

30/07/1999 dans *Tout Nomade*

"beau dommage"	0

30/07/1999 dans *la France*

"beau dommage"	1

30/07/1999 dans *le Québec*

"beau dommage"	2

30/07/1999 dans *le Canada*

ceveil	1

30/07/1999 dans *Tout Nomade*

ceveil	0

30/07/1999 dans *la France*

ceveil	0

30/07/1999 dans *le Québec*

ceveil	1

30/07/1999 dans *le Canada*

YAHOO!

ceveil	2 (sites français uniquement)

29/07/1999

ALTAVISTA

"beau dommage"	431

23/08/1999 (*any language*)

"beau dommage"	299
ceveil	343

23/08/1999 (*French*)

ECILA

ceveil	43

29/07/1999

EXCITE

"beau dommage"	194

27/07/1999 en *français* sur le *Web mondial*

"beau dommage"	8
ceveil	5 (sites français uniquement)

27/07/1999 en *français* sur le *Web français*

HOTBOT

"beau dommage"	390

23/08/1999 (*any language*)

"beau dommage"	72
ceveil	54

23/08/1999 (*French*)

INFOSEEK

"beau dommage"	155

10/08/1999 sur *tout le Web*

"beau dommage"	0
ceveil	1 (site français)

10/08/1999 sur *la France*

"beau dommage"	2
ceveil	3

10/08/1999 sur *le Canada*

LYCOS

"beau dommage"	442

04/08/1999 sur le *Web mondial*

"beau dommage"	948
ceveil	31

04/08/1999 sur le *Web français*

VOILA

beau dommage (recherche avancée : *la phrase*)	382

28/07/1999 sur le *Web mondial*

beau dommage (recherche avancée : *la phrase*)	545
ceveil	317

28/07/1999 sur le *Web francophone*

COPERNIC

beau dommage (recherche rapide : *expression exacte*)	76

29/07/1999 sur le *Web*

beau dommage (recherche rapide : *expression exacte*)	29
ceveil	38

29/07/1999 sur le *Web en français*

2.3. La casse, les caractères diacritiques et les caractères spéciaux

Le traitement réservé à la casse – soit la distinction entre lettres majuscules et minuscules – varie beaucoup d'un outil de recherche à l'autre. Certains SRI ignorent complètement ce phénomène, ramenant toute séquence à une suite de lettres minuscules. D'autres optent pour un traitement différencié, généralement selon la formule suivante : une requête entièrement en minuscules repêche toutes les

occurrences du motif concerné (majuscules et minuscules confondues), tandis qu'une requête comportant des majuscules ne repère que le motif exact soumis. Par exemple, *paris* repère *paris*, *Paris*, *PARIS*, etc., mais *Paris* ne repère que *Paris*.
De la même manière, la gestion des caractères diacritiques (accents, cédilles, etc.) divise les SRI sur Internet en deux factions : d'une part, ceux qui éliminent complètement les accents ; d'autre part, ceux qui optent pour une prise en compte non stricte (où *tache* repère *tache* et *tâche*, tandis que *tâche* ne repère que *tâche*). Enfin, les caractères non alphanumériques tels – ou / font, eux aussi, l'objet de traitements variables selon les cas : leur présence dans une requête est parfois ignorée et parfois strictement respectée.

Nous avons soumis aux outils de recherche diverses variantes de la requête *côte d'azur* : *côte d'azur*, *cote d'azur*, *côté d'azur*, *coté d'azur*, *Côte d'Azur*, *Cote d'Azur*, *Côté d'Azur*, *Côte d'azur*, *côte d'Azur*, etc. Les requêtes *tire-bouchon* et *tire bouchon* ont également été mises en parallèle afin de vérifier le traitement des caractères spéciaux. Toutes les requêtes retenues ont été définies comme locutions, afin de favoriser leurs chances d'être recherchées telles quelles.
Le tableau ci-dessous résume nos conclusions. Les mentions «différencié» et «pris en compte» ont été attribuées dans tous les cas où les multiples versions d'une requête ont entraîné une variation – ne serait-ce que minimale – dans les résultats. Nous avons signalé par un point d'interrogation (?) les cas où les données recueillies ne permettent pas de trancher.

	Casse	**Caractères diacritiques**	**Caractères spéciaux**
CTROUVE.COM	indifférencié	indifférencié	pris en compte
NOMADE	indifférencié	différencié	?
YAHOO!	indifférencié	différencié	ignorés
ALTAVISTA	différencié	différencié	ignorés
ECILA	?	?	pris en compte
EXCITE	indifférencié	différencié	ignorés
HOTBOT	différencié	différencié	ignorés
INFOSEEK	indifférencié	différencié	ignorés
LYCOS	indifférencié	différencié	pris en compte
VOILA	indifférencié	indifférencié	pris en compte
COPERNIC	différencié	différencié	pris en compte

Tableau 1 : Traitement de la casse, des caractères diacritiques et des caractères spéciaux

On constate, au vu de ces résultats, que les outils de notre échantillon sont majoritairement insensibles à la casse mais sensibles aux caractères diacritiques. En ce qui concerne le traitement appliqué aux caractères spéciaux, ils se répartissent à parts égales entre la prise en compte et la non-prise en compte.

Accessoirement, les données obtenues ont mis en évidence un inconvénient potentiel du mode de recherche par locution, souvent le préféré des spécialistes. Le repérage exact censé le caractériser ne peut pas être intégralement mis en œuvre chez les outils qui opèrent au départ un traitement indifférencié de la casse ou des caractères diacritiques, comme le montrent les résultats suivants :

Lycos

"côte d'azur"	1 402
"Côte d'Azur"	1 402
"Côte d'azur"	1 402
"côte d'Azur"	1 402

04/08/1999 sur le *Web mondial*

VOILA

NB : le traitement de la casse est indifférencié.

côte d'azur (recherche avancée : *la phrase*)	19 638
cote d'azur (recherche avancée : *la phrase*)	19 638
côté d'azur (recherche avancée : *la phrase*)	19 638
coté d'azur (recherche avancée : *la phrase*)	19 638

28/07/1999 sur le *Web mondial*

Une telle perte de précision, certes, est déplorable. On peut toutefois se demander si elle n'est pas préférable à la situation que provoque au niveau du rappel, chez les autres outils, la conjonction du traitement différencié et de la recherche par locution. Considérons les résultats suivants :

HOTBOT

"côte d'azur"	7 890
"Côte d'Azur"	7 750
"Côte d'azur"	7 750
"côte d'Azur"	6 990

23/08/1999 (*any language*)

EXCITE

NB : le traitement de la casse est indifférencié.

"côte d'azur"	5 827
"cote d'azur"	1 393
"côté d'azur"	1
"coté d'azur"	0

27/07/1999 en *français* sur le *Web mondial*

On voit que les motifs exacts sont recherchés lors de chacune des requêtes, puisque le total des occurrences obtenues par les requêtes à la «syntaxe pauvre» (respectivement *côte d'azur* et *cote d'azur*) demeure inférieur à la somme des totaux de leurs variantes respectives. Ceci revient à dire qu'un internaute désireux d'obtenir des renseignements sur la Côte d'Azur et qui recourt à la recherche par locution dans un outil comme EXCITE ou HOTBOT passe inévitablement à côté d'une quantité non négligeable de résultats, peu importe la manière dont il saisit sa requête. Pour réaliser toutes les implications de ce phénomène, il ne faut pas perdre de vue que beaucoup de textes sur Internet sont inaccentués ou rédigés entièrement en majuscules. Une autre source de problèmes, à ce niveau, peut venir de l'emploi des majuscules accentuées : les usages variant selon les régions de la Francophonie, des recherches sur des mots comme *État* / *Etat* ou *École* / *Ecole* risquent fort d'être hasardeuses.

Les résultats de ce test, enfin, ont à nouveau contredit, à l'occasion, ce que l'on peut lire dans le mode d'emploi des différents outils. C'est le cas notamment pour NOMADE (qui affirme ne distinguer ni casse ni caractères diacritiques) et pour INFOSEEK (qui a produit des résultats indifférenciés au niveau de la casse, alors que cet outil prétend baser son identification des noms propres sur les distinctions de majuscules[39]).

Les résultats ci-dessous n'incluent pas les variantes avec majuscules pour les outils insensibles à la casse, à moins qu'elles n'aient fait l'objet d'un commentaire, comme pour INFOSEEK.

[39] Plus précisément, INFOSEEK considère deux mots qui se suivent et qui commencent chacun par une majuscule (ce qui n'est pas tout à fait le cas de *Côte d'Azur*) comme un titre ou un nom propre. Peut-être cet outil limite-t-il la reconnaissance de la casse à ces cas particuliers...

CTROUVE.COM

NB : le traitement de la casse est indifférencié.

côte d'azur	987
cote d'azur	987
côté d'azur	987
coté d'azur	987
tire-bouchon	0
tire bouchon	42

28/07/1999

NOMADE

NB : le traitement de la casse est indifférencié.

"côte d'azur"	124 catégories / plus de 150 sites
"cote d'azur"	124 catégories / plus de 150 sites
"côté d'azur"	0 (16 avec ALTAVISTA)
"coté d'azur"	0
"tire-bouchon"	0 (537 avec ALTAVISTA)
"tire bouchon"	0 (537 avec ALTAVISTA)

30/07/1999 sur *Tout Nomade*

YAHOO!

NB : le traitement de la casse est indifférencié

"côte d'azur"	134
"cote d'azur"	135
"côté d'azur"	0 (6 avec INKTOMI)
"coté d'azur"	0
"tire-bouchon"	3
"tire bouchon"	3

29/07/1999

AltaVista

"côte d'azur"	23 835
"cote d'azur"	30 319
"côté d'azur"	16
"coté d'azur"	12
"Côte d'Azur"	15 236
"Côte d'azur"	15 999
"côte d'Azur"	20 765
"Cote d'Azur"	30 751
"tire-bouchon"	537
"tire bouchon"	537

23/08/1999 (*any language*)

ECILA

côte d'azur (*phrase exacte*)	plus de 200
cote d'azur (*phrase exacte*)	plus de 200

côté d'azur (*phrase exacte*)	plus de 200
coté d'azur (*phrase exacte*)	plus de 200
tire-bouchon (*phrase exacte*)	116
tire bouchon (*phrase exacte*)	31

28/07/1999

EXCITE

NB : le traitement de la casse est indifférencié.

"côte d'azur"	5 827
"cote d'azur"	1 393
"côté d'azur"	1
"coté d'azur"	0
"tire-bouchon"	174
"tire bouchon"	174

27/07/1999 en *français* sur le *Web mondial*

HOTBOT

"côte d'azur"	7 890
"cote d'azur"	5 680
"côté d'azur"	9
"coté d'azur"	4
"Côte d'Azur"	7 750
"Côte d'azur"	7 750
"côte d'Azur"	6 990
"Cote d'Azur"	64
"tire-bouchon"	140
"tire bouchon"	140

23/08/1999 (*any language*)

INFOSEEK

NB : le traitement de la casse est indifférencié.

"côte d'azur"	5 573
"cote d'azur"	10 047
"côté d'azur"	25
"coté d'azur"	1
"Côte d'Azur"	5 573
"Côte d'azur"	5 573
"côte d'Azur"	5 573
"Cote d'Azur"	10 047
"tire-bouchon"	144
"tire bouchon"	144

10/08/1999 sur *tout le Web*

Lycos

NB : le traitement de la casse est indifférencié.

"côte d'azur"	1 402
"cote d'azur"	402

"côté d'azur"	123
"coté d'azur"	19
"tire-bouchon"	42
"tire bouchon"	64

04/08/1999 sur le *Web mondial*

VOILA

NB : le traitement de la casse est indifférencié.

côte d'azur (recherche avancée : *la phrase*)	19 638
cote d'azur (recherche avancée : *la phrase*)	19 638
côté d'azur (recherche avancée : *la phrase*)	19 638
coté d'azur (recherche avancée : *la phrase*)	19 638
tire-bouchon (recherche avancée : *la phrase*)	280
tire bouchon (recherche avancée : *la phrase*)	64

28/07/1999 sur le *Web mondial*

COPERNIC

côte d'azur (recherche rapide : *expression exacte*)	79
cote d'azur (recherche rapide : *expression exacte*)	73
côté d'azur (recherche rapide : *expression exacte*)	32
coté d'azur (recherche rapide : *expression exacte*)	21
Côte d'Azur (recherche rapide : *expression exacte*)	80
Cote d'Azur (recherche rapide : *expression exacte*)	70
Côté d'Azur (recherche rapide : *expression exacte*)	34
Coté d'Azur (recherche rapide : *expression exacte*)	15
tire-bouchon (recherche rapide : *expression exacte*)	58
tire bouchon (recherche rapide : *expression exacte*)	53

29/07/1999 sur le *Web*

2.4. L'ordre des mots

Il nous a paru intéressant de vérifier si l'ordre des mots était pris en compte en ce qui a trait aux requêtes de base[40]. Nous avons donc administré aux SRI de notre échantillon deux requêtes successives, l'une portant sur les mots *course* et *voiture*, l'autre sur les mots *voiture* et *course*. Afin de rendre les deux termes obligatoires, nous avons utilisé le symbole + devant chaque mot pour tous les outils qui le

[40] Nous entendons par là des requêtes ne comportant ni opérateurs, ni modificateurs, ni définition d'options de recherche autres que celles disponibles par défaut.

permettaient (soit la majorité) ; dans les autres cas, nous avons choisi l'option *tous les mots*[41].

Nous escomptions une réponse négative. Effectivement, seuls le moteur HOTBOT et le métamoteur COPERNIC ont produit des résultats différenciés lors de ce test, encore que cette différence soit fort minime dans le cas de COPERNIC (89 versus 88). Les autres outils ne font aucun cas de l'ordre d'apparition des termes d'une requête de base.

On peut, certes, justifier cette situation en argumentant que ce genre de prise en compte automatique engendrerait inutilement beaucoup de silence. L'ordre des mots, en effet, s'avère non pertinent pour les requêtes en OU et pour la majorité des requêtes en ET. Il est, cependant, plusieurs cas où une telle distinction peut s'avérer utile. Nous n'en développerons qu'un, à titre d'exemple : le traitement des locutions.

En anglais, les noms composés ou *noun phrases* formés par la juxtaposition de noms simples sont monnaie courante, par exemple *box office* à partir de *box* (boîte) et *office* (bureau) ou *dress circle* (premier balcon) à partir de *dress* (robe) et *circle* (cercle). Dans la mesure où, comme nous l'avons mentionné, la majorité des internautes entrent leurs requêtes sans y incorporer d'opérateurs et sans identifier explicitement les locutions (par l'emploi de guillemets, de cases à cocher ou de choix dans un formulaire), une suite de mots clés en langue anglaise peut aussi bien être formée de termes totalement disjoints que comporter un ou plusieurs groupe(s) de mots fonctionnant comme un tout.

La situation est différente en d'autres langues. En allemand, les noms composés demeurent habituellement des unitermes, obtenus en «agglutinant» les mots initiaux, par exemple *Regenschirm* (parapluie) à partir de *Regen* (pluie) et *Schirm* (bâton). En français, les noms composés obtenus par juxtaposition de noms simples, comme *cheval vapeur*, sont relativement rares. D'ordinaire, ce type de syntagme comporte

[41] Comme nous l'avons constaté, ces deux modes de requête ne sont pas forcément équivalents. Mais ces nuances opératoires n'altèrent en rien les conclusions de ce test particulier, dans la mesure où il s'agit ici d'une comparaison intra-outil.

l'inclusion minimale d'une préposition entre les différents éléments, comme dans *course de voitures* et *voiture de course*, structure qui peut servir à les repérer et empêche la confusion avec une simple suite de mots clés individuels.

Dans la mesure où l'emploi de mots composés constitue une excellente technique pour l'interrogation de SRI, il nous semble que l'ordre d'apparition des mots dans les requêtes de base pourrait être exploité pour faciliter l'identification de ces derniers – y compris en français, puisque la plupart des SRI éliminent de ces requêtes les soi-disant «mots vides» (articles, conjonctions, prépositions, etc.), ramenant ainsi *course de voitures* et *voiture de course* à *course voiture* versus *voiture course*[42]. Un tel processus pourrait s'effectuer via la confrontation de la requête avec des bases de connaissances lexicales (idéalement multilingues) recensant les mots composés. Plus simplement, une solution mitoyenne pour limiter le silence tout en augmentant la précision des résultats serait de proposer en tête des classements de pertinence les documents où l'ordre d'apparition des mots est respecté[43].

CTROUVE.COM

course voiture	1 012
voiture course	1 012

28/07/1999

NOMADE

course voiture (recherche avancée : *tous les mots*)	39
voiture course (recherche avancée : *tous les mots*)	39

30/07/1999 sur *Tout Nomade*

YAHOO!

+course +voiture	15
+voiture +course	15

29/07/1999

ALTAVISTA

[42] La troncature en *s* est automatique sur la plupart des outils de recherche.

[43] Ce qui n'était pas le cas lors de notre test : *course voiture*, par exemple, a souvent repêché en priorité des documents consacrés aux voitures de course (sites de fabricants, petites annonces, etc.) ou entremêlé dans les premières places ces documents avec ceux traitant de courses de voitures. Puisque la distinction fondée sur l'ordre des mots est particulièrement susceptible d'être pertinente en anglais, il est d'ailleurs curieux que les outils d'origine anglo-saxonne en tiennent si peu compte ; toutefois, peut-être effectuent-ils ce genre de traitement sur les requêtes clairement composées de mots clés de langue anglaise, ce que nous n'avons pas vérifié.

+course +voiture	10 535
+voiture +course	10 535

23/08/1999 (*any language*)

ECILA

course voiture (*tous les mots*)	plus de 200
voiture course (*tous les mots*)	plus de 200

29/07/1999

EXCITE

+course +voiture	849
+voiture +course	849

27/07/1999 en *français* sur le *Web mondial*

HOTBOT

+course +voiture	2 060
+voiture +course	2 080

23/08/1999 (*any language*)

INFOSEEK

+course +voiture	53
+voiture +course	53

10/08/1999 sur *tout le Web*

LYCOS

+ course + voiture	207
+ voiture + course	207

04/08/1999 sur le *Web mondial*

VOILA

+course +voiture (recherche avancée)	1 369
+voiture +course (recherche avancée)	1 369

28/07/1999 sur le *Web mondial*

COPERNIC

course voiture (recherche rapide : *tous les mots*)	89
voiture course (recherche rapide : *tous les mots*)	88

29/07/1999 sur le *Web*

2.5. Résultats obtenus avec DIGOUT4U

Toutes les requêtes soumises à l'agent intelligent DIGOUT4U ont été réalisées le 30 juillet 1999. Le logiciel a été configuré de manière à fonctionner 20 minutes par requête et à ne repérer que des documents en français. Suite à chaque recherche, les

résultats ont été filtrés à l'aide de l'option «*les 100 meilleures pages*» et exportés avec résumés.

Nous avons conservé les requêtes *Beau Dommage*, *CEVEIL* et *Côte d'Azur* ; nous avons également soumis les énoncés *bœuf aux hormones* et *maladie de la vache folle*. Nous présentons ci-dessous les mots clés français et anglais[44] suggérés par le logiciel pour chaque requête.

- Beau Dommage

Mots clés français choisis par DIGOUT4U : *beau, dommage*

Mots clés anglais choisis par DIGOUT4U : *beautiful, dommage*

- CEVEIL

Mots clés français choisis par DIGOUT4U : *ceveil*

Mots clés anglais choisis par DIGOUT4U : *ceveil*

- Côte d'Azur

Mots clés français choisis par DIGOUT4U : *côte ciel*

Mots clés anglais choisis par DIGOUT4U : *côte, sky*

- bœuf aux hormones

Mots clés français choisis par DIGOUT4U : *bœuf, hormone*

Mots clés anglais choisis par DIGOUT4U : *ox, hormone*

- maladie de la vache folle

Mots clés français choisis par DIGOUT4U : *maladie, vache, folle*

Mots clés anglais choisis par DIGOUT4U : *disease, mad, cow*

On voit que le processus de décomposition en mots clés des requêtes fonctionne bien sur les noms communs (*bœuf aux hormones*, *maladie de la vache folle*), mais connaît des ratés en ce qui concerne les noms propres : il aurait été nettement préférable de conserver telles quelles des expressions comme *Beau Dommage* et *Côte d'Azur*, que le logiciel n'identifie visiblement pas comme étant des locutions[45]. On peut toujours lui pardonner sa méconnaissance de la réalité musicale québécoise, mais une telle ignorance est plus surprenante en ce qui concerne une expression aussi répandue que *Côte d'Azur*...

---

[44] DIGOUT4U affiche une «traduction» de la requête en mots clés français et anglais même lorsque l'emploi d'une des deux langues est désactivé comme c'était ici le cas. Nous incluons les mots clés en anglais obtenus suite à nos requêtes pour donner une idée de la performance du logiciel à ce niveau.

[45] Si cela avait été le cas, il aurait probablement proposé la traduction anglaise pertinente de *Côte d'Azur*, soit *French Riviera*.

Toutefois, en examinant les résultats obtenus, on constate que ces problèmes de traduction en mots clés n'ont pas empêché DIGOUT4U de produire des résultats tout à fait honorables en ce qui concerne les requêtes relatives à des noms propres. Parmi les réponses évaluées à 100% de pertinence par le logiciel pour la recherche *Côte d'Azur* (requête au demeurant fort vague), on trouve ainsi des titres comme :

- Azur.org, le site de la Côte d'Azur
- Centre de Ressources Côte d'Azur
- Chambre de Commerce et d'Industrie Nice Côte d'Azur
- Hôtels de la Côte d'Azur
- Informations pratiques Côte d'Azur French Riviera France
- etc.

De même, la quasi-totalité des documents repérés pour la requête *Beau Dommage* présentent un lien (plus ou moins substantiel selon les cas, puisqu'il s'agit parfois d'une simple mention) avec le groupe québécois : textes de chansons, discographies, biographies des membres du groupe, etc. Toutes les réponses obtenues pour la requête *CEVEIL* mentionnent également explicitement cet organisme, selon des niveaux de détail variables.

Pour ces requêtes, les «false drops» constituent donc l'exception. À titre d'exemple, mentionnons le repêchage cocasse, pour la requête *Côte d'Azur*, d'un site consacré aux truites arc-en-ciel et doté d'un taux de pertinence de 98% :

La truite arc-en-ciel

http://www.ncr.dfo.ca/COMMUNIC/ss-marin/rainbow/arc-ciel.htm

[...] Les grosses truites arc en ciel anadromes de la côte du Pacifique sont connues sous le nom de steelhead . [...][46]

Force est de constater, cependant, quelques problèmes d'identification linguistique pour la requête *Côte d'Azur*, où plusieurs des documents proposés (dont certains avec des taux de pertinence de 100%) sont en allemand ou en anglais. Par ailleurs, le regroupement des résultats et l'affichage d'un seul document par site (*clustering*) demeurant optionnels chez DigOut4U, nous avons remarqué, pour ces trois requêtes,

[46] Rappelons que l'extrait de texte affiché par DIGOUT4U correspond à la ou aux portion(s) du texte censée(s) correspondre le mieux à la requête initiale. On voit que figurent dans cette phrase les mots *côte* et *ciel* identifiés comme mots clés par le logiciel pour *Côte d'Azur*...

une forte concentration des résultats autour de quelques URL : ainsi, pour *Côte d'Azur*, le site de la *Chambre de Commerce et d'Industrie Nice Côte d'Azur* fournit plus de la moitié des résultats, tandis que plus des trois-quarts de ceux obtenus pour la requête *CEVEIL* proviennent du site lui-même.

Assez curieusement, par contre, les résultats obtenus pour *bœuf aux hormones* et *maladie de la vache folle* sont plutôt mitigés. Ils sont, tout d'abord, assez peu nombreux (respectivement 66 et 34 réponses au lieu des 100 autorisées, seuil que les autres requêtes ont atteint sans peine). Le taux de pertinence des réponses varie également beaucoup : alors que, pour les trois requêtes précédentes, il atteignait 100% pour les documents en tête de liste et descendait jusqu'à 95 ou 90% pour les documents les moins bien cotés, pour ces deux requêtes, on ne trouve aucun score parfait et les résultats s'échelonnent entre 98% et des taux aussi bas que 6 ou 7%. Pour *bœuf aux hormones*, par exemple, les derniers documents fournis sont des recettes de cuisine :

Saveurs du monde / Matambre - Roulade de bœuf aux œufs cuits durs
http://saveurs.sympatico.ca/ency_6/boeuf/matambre.htm
Saveurs du monde / Matambre Roulade de bœuf aux œufs cuits durs Matambre Roulade de bœuf aux œufs cuits durs [...] Temps de cuisson : 2 heures environ ou davantage selon la partie de bœuf choisie [...] 1 flanchet de bœuf de 1 , 3 kg ou autre partie de bœuf à braiser 1 oignon coupé grossièrement 60 ml de vinaigre de vin sel , poivre , thym et persil Farce 450 g d ' épinards 375 ml de chapelure fraîche [...]

Pour cette requête, du reste, les résultats s'avèrent décevants dans leur ensemble. Plus du tiers des réponses est constitué par la répétition incessante[47] de cet extrait d'une interrogation adressée à DEJA.COM, outil spécialisé dans la recherche à l'intérieur des groupes de discussion :

Deja.com: Discussion Search Results
http://r.hotbot.com/r/hb_res_sp_hlt_deja/http://www.deja.com/=hotbotad/dnquery.xp?query=boeuf+aux+hormones
[...] Discussion Search Results : " boeuf aux hormones " Help | Feedback Top Forums related to boeuf aux hormones : Up to 50 % off on books about boeuf aux hormones at Amazon . com [...] Get more forums related to boeuf aux hormones Messages related to bœuf aux hormones : Messages 1 25 of exactly 67 matches [...] Re : du boeuf aux hormones su fr . rec . sport . cyclisme [...] du boeuf aux hormones sur le fr . rec . sport . cyclisme [...] Re : du boeuf aux hormones su fr . rec . sport . cyclisme [...] Re : du boeuf aux hormones su fr . rec . sport . cyclisme [...]

Le moteur VOILA fournit également, à lui seul, un autre tiers des 66 réponses, sous la forme de références peu conviviales comme celle-ci :

Voila les réponses...
http://world.voila.com/search?dt=*&medor=web&kw=boeuf+hormone&an=1&dc=&ad=0&ap=8
[...] Tous les services de Voila ACTUALITE Journal AFP Programmes Télé Bourse Météo Programmes Sorties Horoscope ANNUAIRES Pages Jaunes Pages Blanches Rues Commerçantes Adresses E mail Paris en Photos COMMUNICATION E Mail gratuit Voila Club Newsgroups Chat INFOVILLE Plans Itineraires Tourisme RECHERCHE Web Francophone Web Mondial ACCUEIL [...] boeuf (7 . 210) hormone (89 . 843) Affiner la recherche [...] The Hormone Foundation [1 mot sur 2] Welcome to the Hormone Foundation web site . The Hormone Foundation is dedicated to improving the quality of life by promoting the prevention , diagnosis , and treatment of human disease in ... 8 Juil 1999 , 1kb , www . hormone . org / Plus de pages sur ce site ... Facts about human growth hormone [1 mot sur 2] Facts about human growth hormone What is human growth hormone ? Human growth hormone (hGH) is produced in the pituitary gland of humans , and the hormone is ... 18 Mai 1999 , 3kb , www . novo . dk / backgrou / backgrou / bahghuk . htm

Ce dernier extrait est non seulement de langue anglaise, mais se rapporte, comme on le voit, à un site consacré aux hormones de croissance humaine. Cela laisse entrevoir la stratégie de recherche de DIGOUT4U : les documents comportant tous les concepts définis par le logiciel sont favorisés, mais, lorsque les documents de ce type se font rares, les agents se rabattent rapidement sur ceux qui ne renferment que l'un ou l'autre de ces concepts.

En fait, pour cette requête, on peut compter sur les doigts d'une main les réponses «exploitables», comme celle-ci qui semble toutefois être davantage un article d'opinion qu'un texte informatif sur le sujet (taux de pertinence de 91%) :

la Baleine. Mars-Avril 1997
http://www.apro.fr/natcog/at/publications/baleine/b_1997_3.html
[...] Y a t il obligation pour l ' homme européen de manger du bœuf américain aux hormones ? Sanitairement , non . Commercialement, oui . [...] Rien ne prouverait que le bœuf «enrichi» aux hormones pose un problème de santé pour le consommateur . Rien ne prouve le contraire non plus , surtout à longue échéance Naturellement pour cette organisation le doute doit profiter au commerce et aux américains au nom de la libre concurrence ! [...]

La requête *maladie de la vache folle* a été un peu plus fructueuse. Elle a permis de repérer des documents comme ceux-ci (taux respectifs de 97 et 86% de pertinence) :

Dossier sur la vache folle - par G. Latzko-Toth
http://www.mlink.net/~glt/prions.htm
Dossier sur la vache folle par G . Latzko Toth Le point sur la maladie de la vache folle [...] Officiellement cantonnée à l ' espèce bovine jusqu ' en 1996 , l ' épidémie d ' encéphalopathie spongiforme qui a frappé l ' ensemble du cheptel anglais

[47] Seules de légères variations d'URL sont constatées, ce qui incline à penser que DIGOUT4U est tombé lors de cette requête dans ce que nous avons appelé précédemment un «piège à robot».

pourrait bien avoir franchi la " barrière des espèces " pour s ' en prendre à l ' homme . Aujourd ' hui , la plupart des chercheurs du monde entier , à la suite d ' un groupe de chercheurs britanniques , considèrent qu ' une vingtaine de cas de maladie de Creutzfeldt Jakob (une maladie humaine jusqu ' ici rarissime) sont très probablement liés à l ' ingestion de viande contaminée . De plus , la recherche accélérée sur cette maladie a vu triompher l ' hypothèse naguère très controversée selon laquelle ces maladies sont causées par un nouveau type d'agents infectieux , les prions , qui ont valu à leur " inventeur " , Stanley Prusiner , le prix Nobel de médecine 1997 . [...] un article de vulgarisation pour mieux comprendre de quoi il s ' agit : Prions : les microbes du 3e type arrivent

L'Encephalopathie Spongiforme Bovine
http://www.who.ch/inf/am/am113.html
[...] C ' est en novembre 1986 , lorsqu ' une forme jusque là inconnue de maladie neurologique est apparue chez des bovins au Royaume Uni , que l ' attention de la communauté scientifique a été attirée pour la première fois sur l ' encéphalopathie spongiforme bovine (ESB) . Entre novembre 1986 et le 31 mai 1996 , environ 160 000 cas de cette maladie des bovins nouvellement identifiée ont été confirmés au Royaume Uni . [...] Différentes hypothèses ont été avancées pour expliquer l ' apparition de cette maladie dans la chaîne alimentaire du bétail , parmi lesquelles sa présence spontanée chez des bovins dont les carcasses ont ensuite été introduites dans la chaîne alimentaire du bétail , ou encore son entrée dans cette chaîne à partir de carcasses de moutons atteints d ' une maladie similaire . [...] La maladie est mortelle pour les bovins en quelques ou quelques semaines mois . [...] Dans un groupe de pays constitué par la France , l ' Irlande , le Portugal et la Suisse , la maladie est apparue dans des troupeaux indigènes et le phénomène a été attribué en partie à l ' importation d ' aliments pour bétail en provenance du Royaume Uni.

Toutefois, la seconde moitié de la liste des résultats s'éloigne résolument du thème de la vache folle pour proposer des documents reliés de manière plus large à la santé (obésité, protection face au soleil) et à la maladie (cancer, SIDA, maladie d'Alzheimer, etc.).

3. Conclusion de la troisième partie

La méthodologie que nous avons appliquée pour examiner les SRI sur Internet demeure une ébauche et gagnerait à être retravaillée sur certains aspects. Il y aurait ainsi lieu d'accroître la taille de l'échantillon d'outils, de viser une représentation plus équitable des différentes catégories d'instruments et de sortir des sentiers battus pour évaluer des systèmes moins connus. Il conviendrait aussi de multiplier les points de comparaison et, surtout, d'augmenter considérablement le nombre des requêtes soumises pour chacun d'eux. Enfin, il serait sans doute intéressant de mener ce type d'analyse sur une base longitudinale, afin de suivre l'évolution dans le temps du comportement des outils.

Les données obtenues doivent donc être abordées avec une certaine circonspection. Elles ont, néanmoins, le mérite de démontrer que le fonctionnement des outils de recherche contribue sensiblement à rendre le processus de repérage de l'information sur Internet encore plus complexe et aléatoire.

Un des problèmes les plus fréquemment évoqués pour expliquer le relatif constat d'échec de la recherche d'information sur le Web – outre le gigantisme des bassins de ressources à recenser et les vicissitudes du traitement documentaire – tient à la difficulté d'appropriation, pour l'internaute moyen, des multiples systèmes de repérage aux syntaxes d'interrogation parfois sibyllines. Comme le souligne A. Poulter :

> *It is highly ironic that a unitary global information space (of networked computers of all types, various client/server applications and standard format data files) accessible via one freely-available software package (a WWW client browser) should be so balkanised by a plethora of search engines. It is the complete reverse of the traditional information world, of printed sources, CD-ROM and online databases, where a limited and comparatively stable range of well-known and trusted search tools attempt to homogenise a large number of physically separate and disparate collections.* [Poulter 1997]

Il est certain, du reste, que la performance d'un outil de recherche est tributaire en partie du comportement de l'usager, de son «background», de ses besoins informationnels. Ce dernier porte d'ordinaire une part de responsabilité dans l'insuccès de ses requêtes de recherche, comme Poulter le fait remarquer : *«This is a common failing of WWW search engines, in that although they are populist tools, they assume a great deal on the part of their searchers.»* [Idem] D'autres facteurs externes peuvent également jouer un rôle à ce niveau, tels le type de requête effectuée ou le sujet sur lequel porte la recherche.

Ce que nos résultats semblent suggérer, toutefois, c'est que les difficultés de repérage sont *aussi* causées par le fait que le monde des outils de recherche sur

Internet constitue un domaine de faux-semblants et d'apparences trompeuses, où un outil en dissimule souvent un autre. Un univers où, par exemple, des manières de formuler une requête présentées comme équivalentes peuvent ultimement se différencier au niveau des résultats, alors que, à l'inverse, des formulations apparemment distinctes vont se traduire dans les faits par des résultats identiques. Des fluctuations de ce type se rencontrant à la fois entre les outils et pour un même instrument, il devient impossible de définir des principes de recherche uniformément valables. De même, les étiquettes des choix disponibles sur les pages des outils de recherche sont équivoques : on ne sait jamais vraiment ce qu'elles recouvrent et ce qui est laissé de côté. Enfin, comme nous l'avons vu à propos du traitement des majuscules et des lettres accentuées, certains des choix logistiques opérés par les outils de recherche peuvent s'avérer très lourds de conséquences dans certains contextes ; or, ces implications sont rarement connues des usagers, même spécialistes. Couronnant le tout, la documentation en ligne des divers SRI ne se révèle souvent d'aucune utilité pour éclairer l'internaute, étant incomplète, confuse, voire même carrément inexacte.

Dans une telle conjoncture, blâmer principalement la négligence ou l'incompétence des utilisateurs pour justifier de l'incurie actuelle des outils de recherche relève de la mauvaise foi caractérisée. Si la solution au problème du repérage d'information sur Internet comporte effectivement un aspect relié à la formation des usagers, elle devra surtout s'appuyer, d'une part, sur un mouvement de normalisation intra- et inter-outils des caractéristiques de fonctionnement, et, d'autre part, sur le perfectionnement des instruments de seconde génération plus «intelligents» et conviviaux qui commencent à apparaître, comme DIGOUT4U.

Conclusion

Nous conclurons en discutant brièvement les implications, pour l'usager profane, de la réalité actuelle du repérage de l'information sur Internet, et en proposant quelques pistes pour des recherches futures.

Comme nous l'avons suggéré déjà, pour l'internaute type, la précision des résultats demeure le souci principal suite à une requête, loin devant le rappel qui préoccupe tant les spécialistes de l'information (80% des usagers du Web ne visionneraient que les deux premières pages d'une liste de résultats, et encore[48]). Dans ce contexte, on peut supposer que l'impact des problèmes de repérage qui ont été discutés tout au long de ce travail demeure relativement mineur pour ces usagers : une très grande quantité de résultats est d'ordinaire repérée suite à une requête sur Internet et, le tri de pertinence faisant son œuvre, il se trouve presque immanquablement, en tête de liste des résultats, quelques références susceptibles de satisfaire le besoin d'information – ou de divertissement – de ces internautes. Le fait de passer à côté d'une quantité importante de résultats potentiellement tout aussi intéressants leur demeure, à la limite, indifférent ; ils ne souffriront de cette situation que lors de recherches pointues.

En fait, nous inclinerions même à penser que les dysfonctionnements et limitations des outils de recherche sur Internet ne devraient pas constituer outre mesure un sujet de préoccupation pour les utilisateurs finaux non spécialistes (en admettant qu'ils en aient conscience), du moins dans la mesure où leurs besoins demeurent satisfaits. Tout au plus peut-on leur conseiller de ne jamais rien prendre pour acquis et d'effectuer à l'occasion quelques requêtes exemplaires afin de vérifier le comportement des SRI lorsque les résultats s'avèrent peu nombreux ou insatisfaisants (ceci vaut tout particulièrement pour les utilisateurs qui appliquent

[48] Jansen, B.J.; Spink, A.; Bateman, J. et T. Saracevic. «Real life information retrieval: a study of user queries on the Web». *SIGIR Forum*, 1998, 32 (1) : 5-17.

fréquemment des critères de restriction géographique ou linguistique, ou qui sont des adeptes de la recherche par locution).

De toute façon, la maîtrise absolue du processus de recherche sur le Web demeurera sans doute une chimère :

> *Information retrieval on the Web is rooted in an interactive graphical presentation and mouse-based point-and-shoot input [...] which is different from the traditional query-based search technique. The Web search procedures which follow the links between hypertexts involve large jumps between information subjects. Users, facing too many choices, might jump away from the original search target or get lost in cyberspace. It is almost impossible to formalize a search strategy or to repeat the same search procedure at a different time in a complicated search procedure. The interaction of Internet tools is limited compared with more established electronic sources [...]*. [Dong & Su 1997]

En ce qui concerne les orientations futures de recherche, il convient, tout d'abord, de mentionner l'intérêt que présenterait l'étude approfondie du mode de fonctionnement des outils en ce qui a trait à l'analyse linguistique effectuée sur les documents et les requêtes. Quelles sont précisément, par exemple, les règles mises en œuvre pour permettre l'élargissement d'une recherche sur les pluriels – réguliers et irréguliers – et les termes proches phonétiquement ? Les algorithmes utilisés à cette fin pour produire automatiquement les variantes morphologiques des mots (ainsi, étant donné un terme comme *travel*, les formes *travels*, *traveled*, *traveling*, etc., sont générées) se basaient exclusivement, à l'origine du moins, sur les règles de l'anglais. Cette remarque s'applique également aux processus de découpage d'une requête en mots clés (souvent sur la base de la ponctuation), à la troncature à droite sur les pluriels[49], aux listes de mots vides, etc. Comment la situation a-t-elle évolué à ce niveau ? Le traitement du français – et d'autres langues – est-il désormais pris en charge convenablement ? Comment les SRI parviennent-ils à gérer de façon

[49] On peut opposer ici les règles morphologiques motivées linguistiquement aux algorithmes de troncature, qui se contentent de «couper» un certain nombre de caractères à partir de la fin d'un mot.

simultanée la réalité de plus en plus multilingue d'Internet ? Constate-t-on une différence marquée à ce sujet entre les outils d'origine anglo-saxonne et les autres ?

Une autre thématique à surveiller concerne la manière dont les SRI sur Internet s'y prendront pour favoriser la recherche de l'information disponible sous des formats non textuels : les fichiers audio et vidéo, par exemple, se multiplient à l'heure de la «grande convergence» constatée entre les différents médias (Internet, radio, télévision, presse, etc.). Il serait fort utile de pouvoir repérer des informations plus précises que le simple nom de ces fichiers ou un éventuel résumé de leur contenu... Un peu comme les moteurs permettent de retrouver un renseignement qui figure dans l'énième page d'un document en texte intégral, il y aurait lieu de développer ce mode d'accès pour les autres types de fichiers, ce qui constitue à n'en pas douter un défi redoutable pour les concepteurs d'outils de recherche.

Enfin, il sera également intéressant de voir comment les outils de recherche sur Internet évolueront en ce qui concerne les nouvelles technologies mises au point afin de tenter de rendre la machine un peu plus conviviale pour l'être humain moyen. À l'heure où, par exemple, les fureteurs Web commencent à permettre la navigation sur la base de simples commandes prononcées à voix haute, verra-t-on bientôt des outils de recherche qui accepteront en direct des requêtes vocales ?

Bibliographie

Outre les sources mentionnées ci-après, nous avons eu recours, dans le cadre de ce travail, à l'information présente sur les sites mêmes des divers outils de recherche. Nous avons tenté, pour les références des ressources électroniques, de fournir une description aussi exhaustive que possible. Toutefois, certaines informations manquent parfois (date, nom du responsable, etc.), car elles sont demeurées introuvables.

RESSOURCES IMPRIMEES

Addison, E.R. ; Feder, J. et H.D. Wilson. «The impact of plain English searching on end users». In Martha E. Williams (éd.) : *Proceedings of the 14th National Online Meeting 1993*. Learned Information, Inc., New York, 4-6 May 1993. Medford (New Jersey) : Learned Information, Inc., 1993 : 5-9.

Allen, E.E. «Searching, naturally». *Internet Reference Services Quarterly*, 1998, 3 (2) : 75-81.

Allen, J. *Natural Language Understanding. 2nd Ed*. Redwood City (Californie) : Benjamin-Cummings, ©1995. XV + 654 pages.

Andrieu, O. *Créer du trafic sur son site Web*. Paris : Éditions Eyrolles, 1998. 500 pages.

Andrieu, O. *Trouver l'info sur Internet*. Paris : Éditions Eyrolles, 1998. 460 pages.

Balas, J.L. «Exploring some new search tools for librarians». *Computers in Libraries*, 1999, (19) : 34-37.

Basch, R. *Researching Online for Dummies*. Foster City (Californie) : IDG Books Worldwide, 1998. 328 pages.

Basch, R. «Searching in plain English». *Link-Up (USA)*, 1994, 11 (2) : 14-15.

Belkin, N.J. et W.B. Croft. «Information filtering and information retrieval : two sides of the same coin ?». *Communication of the ACM*, 1992, 35(12) : 29-38.

Blakeman, K. «Intelligent agents : search tools of the future ?» *Business Information Searcher*, 1997, 7 (1) : 16-18.

Brandt, S.D. «What flavor is your Internet search engine ?». *Computers in Libraries*, 1997, (17) : 47-50.

Centre d'expertise et de veille Inforoutes et Langues (CEVEIL). *Internet, intranet, extranet : comment en tirer profit*. Montréal : Les Éditions Transcontinental, 1998. 208 pages.

Courtois, M.P. et M.W. Berry. «Results ranking in Web search engines». *Online*, 1999, 23 (3) : 39-46.

Croft, W.B. «Approaches to intelligent information retrieval». *Information Processing & Management*, 1987, 23 (4) : 249-254.

Dalloz, X. «Les agents intelligents arrivent». *L'Atelier*, 1995, (46-47) : 24-27.

Desert, S.E. «WESTLAW is natural v. Boolean searching : a performance study». *Law Library Journal*, 1993, 85 (4) : 713-42.

Dong, X. et L.T. Su. «Search engines on the World Wide Web and information retrieval from the Internet : a review and evaluation». *Online & CDROM Review*, 1997, 21 (2) : 67-82

Doszkocs, T.E. «Natural language processing in information retrieval». *Journal of the American Society for Information Science*, 1986, 37 (4) : 191-196.

Evans, R. «Beyond Boolean : relevance ranking, natural language and the new search paradigm». In Martha E. Williams (éd.) : *Proceedings of the 15th National Online Meeting 1994*. Learned Information, Inc., New York, 10-12 May 1994. Medford (New Jersey) : Learned Information, Inc., 1994 : 121-128.

Feldman, S.E. «NLP meets the Jabberwocky : natural language processing in information retrieval». *Online*, 1999, 23 (3) : 62-72. Disponible sur le Web : *http://www.onlineinc.com/onlinemag/OL1999/feldman5.html*

Feldman, S.E. «Searching natural language systems : searchers know thy engine». *Searcher*, 1994, 2 (8) : 34-39.

Feldman, S.E. «Testing natural language : comparing DIALOG, TARGET, and DR-LINK». *Online*, 1996, 20 (6) : 71-79.

Gaizauskas, R. et Y. Wilks. «Information extraction : beyond document retrieval». *Journal of Documentation*, 1998, 54 (1) : 70-105.

Garman, N. «Meta search engines». *Online*, 1999, 23 (3) : 74-78.

Gauch, S. «Intelligent information retrieval : an introduction». *Journal of the American Society for Information Science*, 1992, 43 (2) : 175-182.

Gillaspie, D.L. «The role of linguistic phenomena in retrieval performance». *Proceedings of the 58th Annual Meeting of the American Society for Information Science,* 1995 : 90-96.

Green, R. «The expression of conceptual syntagmatic relationships : a comparative survey». *Journal of Documentation*, 1995, 51 (4) : 315-338.

Haas, S.W. «Natural language processing : toward large-scale, robust systems». *Annual Review of Information Science and Technology*, 1996, 31 : 83-119.

Hayes, P.J. et G. Koerner. «Intelligent text technologies and their successful use by the information industry». In Martha E. Williams (éd.) : *Proceedings of the 14th National Online Meeting 1993*. Learned Information, Inc., New York, 4-6 May 1993. Medford (New Jersey) : Learned Information, Inc., 1993 : 189-196.

Hersh, W.R. et D.H. Hickam. «An evaluation of interactive Boolean and natural language searching with an online medical textbook». *Journal of the American Society for Information Science*, 1995, 46 (7) : 478-489.

Hock, R. «Web search engines features and commands». *Online*, 1999, 23 (3) : 24-28.

Hyams, P. «Q. What creates no noise but isn't silent ?» *Information World Review*, 1997, (131) : 37-38.

Jacso, P. «Don't kiss Boolean goodbye. It's AND not OR, let alone XOR». *Information Today*, 1994, 11 (2) : 22-24.

Jones, K.S. «Artificial intelligence : what can it offer information retrieval ?». In Kevin P. Jones et Verina Horsnell (éd.) : *Informatics 3 Conference held by the Aslib Coordinate Indexing Group 2-4 Apr 75, Emmanuel College*. Londres : Aslib, 1978 : 3-10.

Kang, H-K. et K-S. Choi. «Two-level document ranking using mutual information in natural language information retrieval». *Information Processing & Management*, 1997, 33 (3) : 289-306.

Lalonde, L.-G. et A. Vuillet. *Chercher et trouver dans Internet*. Montréal : Éditions Logiques, 1998. 139 pages.

Lardy, J-P. «Les outils de recherche d'information sur Internet : guides, listes thématiques et index». *Documentaliste*, 1996, 33 (1) : 33-39.

Larouk, O. «Modeling users needs : schemas of interrogation and filtering of answers from the Web in co-operative mode». In Widad Mustafa el Hadi, Jacques Maniez et Steven A. Pollitt (éd.) : *Structures and Relations in Knowledge Organization : Proceedings of the Fifth International ISKO Conference*. Lille (France), 25-29 August 1998. Würzburg : Ergon Verlag, 1998 : 106-115.

Le Guern, M. «Un analyseur morpho-syntaxique pour l'indexation automatique». *Le Français Moderne*, 1991, 59 (1) : 22-35.

Leontyeva, N.N. «Stages of information analysis of natural language texts». *International Forum on Information and Documentation*, 1987, (12) : 8-14.

Liddy, E.D. «An alternative representation for documents and queries». In Martha E. Williams (éd.) : *Proceedings of the 14th National Online Meeting 1993*. Learned Information, Inc., New York, 4-6 May 1993. Medford (New Jersey) : Learned Information, Inc., 1993 : 279-284.

Liddy, E.D. «Enhanced text retrieval using natural language processing». *ASIS Bulletin*, 1998, 24 (4).
Disponible sur le Web : *http://www.asis.org/Bulletin/Apr-98/liddy.html*

Mauldin, M. ; Carbonell, J. et R. Thomason. «Beyond the keyword barrier : knowledge-based information retrieval». *Information Services & Use*, 1987, 7 (4-5) : 103-117.

Narasimhamurthi, N. «Intelligent information retrieval : an introduction». *Information Studies*, 1996, 2 (2) : 75-84.

Notess, G.R. «Rising relevance in search engines». *Online*, 1999 23 (3) : 84-86.

Notess, G.R. «Search engines in the Internet age». *Online*, 1999, 23 (3) : 20-22.

O'Donnell, R. et A.F. Smeaton. «A linguistic approach to information retrieval». In Ruben Leon (éd.) : *Proceedings of the 16th Research Colloquium of the British Computer Society Information Retrieval Specialist Group*. Drymen, Scotland, 22-23 March 1994. Londres : Taylor Graham, 1996 : 68-80.

O'Kane, K.C. «World Wide Web-based information storage and retrieval». *Online & CDROM Review*, 1996, 20(1) : 11-19.

Polity, Y. «Évaluation des modes de recherche en langage naturel». *Documentaliste*, 31 (3) : 136-142.

Poulter, A. «The design of World Wide Web search engines : a critical review». *Program*, 1997, 31 (2) : 131-145.

Pritchard-Schoch, T. «Comparing natural language retrieval : Win & Freestyle». *Online*, 1995, 19 (4) : 83-87.

Pritchard-Schoch, T. «Natural language comes of age». *Online*, 1993, 17 (3) : 33-43.

Repman, J. et R.D. Carlson. «Surviving the storm : using metasearch engines effectively». *Computers in Libraries*, 1999, (19) : 50-55.

Sabah, G. «Knowledge representation and natural language understanding». *AI Communications*, 1993, 6 (3-4) : 155-186.

Salton, G. et M.J. McGill. *Introduction to Modern Information Retrieval*. New York : McGraw-Hill, 1983. 400 pages.

Shukla, K.K. «Some AI techniques for information retrieval». *DESIDOC Bulletin of Information Technology*, 1996, 16 (4) : 13-18.

Smeaton, A.F. «Natural language processing and information retrieval (special issue)». *Information Processing & Management*, 1990, 26 (1) : 19-186.

Smeaton, A.F. «Prospects for intelligent, language-based information retrieval». *Online Review*, 1991, 15 (6) : 373-382.

Stock, O. «A third modality of natural language ?» *Artificial Intelligence Review*, 1995, 9 (2-3) : 129-146.

Strzalkowski, T. «Natural language information retrieval». *Information Processing & Management*, 1995, 31 (3) : 397-417.

Sullivan, D. «Crawling under the hood : an update on search engine technology». *Online*, 1999, 23 (3) : 30-38.

Tegenbos, J. et P. Nieuwenhuysen. «My kingdom for an agent ? Evaluation of Autonomy, an intelligent search agent for the Internet.» *Online & CDROM Review*, 1997, 21 (3) : 139-48.

Thil, J. «Outils "intelligents" de recherche d'informations : mythe ou réalité». *Technologies Internationales*, 1996, (26) : 7-10.

Tomaiuolo, N.G. et J.G. Packer. «An analysis of Internet search engines : assessment of over 200 search queries». *Computers in Libraries*, 1996, (16) : 58-62.

Tudor, J.D. «The new alchemy : using droids & agents to threat information overload». *Online*, 1997, 21 (6) : 50-58.

Vidmar, D.J. «Darwin on the Web : the evolution of search tools». *Computers in Libraries*, 1999, (19) : 22-28.

Wacholder, N. et R.J. Byrd. «Retrieving information from full text using linguistic knowledge». In Martha E. Williams (éd.) : *Proceedings of the 15th National Online Meeting 1994*. Learned Information, Inc., New York, 10-12 May 1994. Medford (New Jersey) : Learned Information, Inc., 1994 : 441-447.

Warner, A.J. «Natural language processing». *Annual Review of Information Science and Technology*, 1987, 22 : 79-108.

Warner, A.J. «Natural language processing in information retrieval». *Bulletin of the American Society for Information Science*, 1988, (14) : 18-19.

Watson, D. «Is this software after your job ?» *Library Association Record*, 1997, 99 (7) : 364-365.

Weinberg, B.H. «Levels of linguistic analysis and information processing». In Charles W. Husbands et Ruth L. Tighe (éd.) : *Information revolution : proceedings of the 38th ASIS Annual Meeting*. Boston, Massachusetts, October 26-30, 1975. Washington (DC) : American Society for Information Science, 1975, 12 : 71-72.

Young, C.W. ; Eastman, C.M. et R.L. Oakman. «An analysis of ill-formed input in natural language queries to document retrieval systems». *Information Processing & Management*, 1991, 27 (6) : 615-622.

RESSOURCES ELECTRONIQUES

Sites spécialisés

Abondance : recherche d'information, référencement et promotion de sites Web
http://www.abondance.com/
Maintenu par **Olivier Andrieu.**

Les agents intelligents
http://ms161u06.u-3mrs.fr/

Maintenu par **Bruno Mannina**.

La Loupe : guide de recherche sur Internet
http://laloupe.magnit.com

Meta News
http://www.metanews.net/
Maintenu par la société La Mine.

Les moteurs de recherche francophones
http://www.idf.net/mdr/
Maintenu par la société IDF.net.

Les outils de recherche : pour enfin s'y retrouver
http://pages.infinit.net/popnet/recherche/
Maintenu par la société Services Pop.net.

Search Engine Showdown (anglophone)
http://www.notess.com/search/
Maintenu par **Gregg R. Notess**.

Search Engine Watch (anglophone)
http://www.searchenginewatch.com/
Maintenu par **Danny Sullivan.**

Un outil de veille stratégique sur Internet
http://perso.club-internet.fr/nygren/
Maintenu par **Pierre Nygren**.

Forum de discussion

alt.internet.search (anglophone)

Listes de discussion/diffusion

Agents
Porte sur les agents intelligents.
Pour inscription : mailto:agents-subscribe@egroups.com

I-Search Digest (anglophone)
Porte sur les outils de recherche.
http://www.audettemedia.com/i-search/

Motrech
Porte sur les moteurs de recherche.
http://www.chez.com/jcharron/motrech/presentation.html
Pour inscription mailto:motrech-subscribe@egroups.com

Autres documents en ligne

Careil, J.M. et B. de Frémont. «Les agents intelligents». Présentation interactive disponible pour consultation sur le site de Bruno Mannina [http://ms161u06.u-3mrs.fr/].

Conférence des recteurs et des principaux des universités du Québec (CREPUQ), Sous-comité des bibliothèques, Groupe de travail sur l'accès aux ressources documentaires, Sous-groupe de travail sur Internet. *«GIRI – Guide d'initiation à la recherche dans Internet»*. Édition du 1er juin 1998.
http://www.bibl.ulaval.ca/vitrine/giri/index.htm

Conférence des recteurs et des principaux des universités du Québec (CREPUQ), Sous-comité des bibliothèques, Groupe de travail sur l'accès aux ressources documentaires, Sous-groupe de travail sur Internet. *«GIRI 2 – Guide des indispensables de la recherche dans Internet»*. Édition du 1er mars 1999.
http://www.bibl.ulaval.ca/vitrine/giri/giri2/index.html

de Rosnay, J. «Les agents intelligents : robots logiciels». 19 octobre 1995.
http://194.199.143.5/derosnay/agent.htm

Jakob, D. «Trouver des informations sur le Web». *Flash Réseau* (15), Bibliothèque nationale du Canada, Services de technologie de l'information. 10 octobre 1995 (révisé le 29 juillet 1997).
http://www.nlc-bnc.ca/pubs/netnotes/fnotes15.htm

Koster, M. «Robots in the Web : threat or treat ?». Avril 1995.
http://info.webcrawler.com/mak/projects/robots/threat-or-treat.html

Laublet, P. «Collecte d'information et recherche documentaire sur Internet». CAMS, Université de Paris-Sorbonne. S.d. (postérieur à 1997).
http://www.mpl.orstom.fr/CDROM/ch06/laublet/laublet.htm
[Ce document est désormais inaccessible]

Plourde, J.-N. «Définition et application de critères d'évaluation d'outils de recherche dans Internet». *Cursus*, 1996, 1 (2).
http://www.fas.umontreal.ca/EBSI/cursus/vol1no2/plourde.html

Annexes : Fiches signalétiques

Table des annexes

Annexe A : Ctrouve.com[50]

URL	http://www.ctrouve.com/
Catégorie	Annuaire
Versions localisées	Non
Version francophone	Outil francophone (sites en provenance de tous les pays)
Taille de la base de données	Plus de 60 000 sites
Possibilité de soumission manuelle de sites	Oui
Modes de recherche	- Navigation thématique - Recherche par mots clés
Prise en compte des méta-données	Oui
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Indifférencié
Mode de recherche par défaut	OU
Fonctions booléennes	Non
Emploi de +/-	Non
Recherche de locutions	Non
Requête à l'intérieur d'un premier groupe de résultats	Non (toutefois, une navigation géographique est disponible parmi les résultats d'une requête)
Classement des résultats	Pertinence présumée
Affichage par défaut	- Titre - Pays et ville, si connus - Nom du responsable - Date d'inscription - URL - Premières lignes du texte
Possibilité de modifier l'affichage par défaut	Non
Choix de la quantité de résultats à afficher	Non
Affichage d'un taux de pertinence	Non
Évaluation de la pertinence	Chaque mot reçoit un poids en fonction de sa position : le titre, le descriptif et certains éléments du contenu du site (balises META, liens hypertextuels, caractères gras) sont favorisés. On tient également compte de la distance entre les mots.
Particularités	Lors de l'enregistrement d'un site, une liste de mots clés est fournie par le responsable (ou constituée par les soins de l'équipe éditoriale). Le mode de classement de CTROUVE.COM est basé sur ces listes : la page d'accueil regroupe alphabétiquement l'ensemble des mots clés qui les composent. Chaque mot clé constitue ainsi un lien qui donne accès à la liste des sites où il figure, de même qu'à une liste des autres mots clés communs à ces mêmes sites. Une façon de procéder originale, mais qui pose d'évidents problèmes en termes d'homogénéité et de rigueur de classement...

[50] Cet outil s'appelait, jusqu'à tout récemment, EUREKA.

Annexe B : NOMADE

URL	http://www.nomade.fr
Catégorie	Annuaire
Versions localisées	Non
Version francophone	Outil francophone (sites en provenance de tous les pays)
Taille de la base de données	75 000 sites
Possibilité de soumission manuelle de sites	Oui
Modes de recherche	- Navigation thématique - Recherche par mots clés
Prise en compte des méta-données	Oui
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Indifférencié
Mode de recherche par défaut	OU
Options de restriction de recherche	- *Tout NOMADE* vs *communiqués de l'AFP* vs *sites récents* vs *sélections de NOMADE* - Sur la racine (recherche avancée) - Sur la nature du site (éducation, personnel, entreprise, etc.) (recherche avancée : cinq options) - Sur la région française ou le Québec (recherche avancée : 31 options) - Sur les pays d'Europe francophone et d'Amérique du Nord (recherche avancée : huit options) - Sur le public (adulte, enfant, professionnel, etc.) (recherche avancée : cinq options)
Fonctions booléennes	ET, OU (formulaire, recherche avancée)
Emploi de +/-	Non
Recherche de locutions	Oui (" ")
Requête à l'intérieur d'un premier groupe de résultats	Non
Classement des résultats	Pertinence présumée
Affichage par défaut	- Nature du site - Nom de l'éditeur - Ville (si connue) - Pays - Public(s) - Titre - Description - Catégorie(s) - URL
Possibilité de modifier l'affichage par défaut	Non
Choix de la quantité de résultats à afficher	Non
Affichage d'un taux de pertinence	Non
Particularités	- Une requête infructueuse est automatiquement redirigée vers ALTAVISTA (d'où la présence éventuelle de pages en d'autres langues dans les résultats). - Sauf si on emploie des guillemets, les mots non significatifs ne sont pas pris

	en compte et sont systématiquement éliminés (articles, conjonctions, prépositions, etc.).

Annexe C : YAHOO!

URL	http://www.yahoo.fr/
Catégorie	Annuaire
Versions localisées	Oui (20)
Version francophone	Oui (base de sites dédiée)
Taille de la base de données	Plus de 75 000 sites
Possibilité de soumission manuelle de sites	Oui
Modes de recherche	- Navigation thématique - Recherche par mots clés
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Indifférencié
Mode de recherche par défaut	OU
Options de restriction de recherche	- Sur le titre - Sur l'URL - *Sites Web* vs *Usenet* (recherche avancée) - *Catégories* vs *Sites Web* vs *Dépêches d'actualité* vs *tout* (recherche avancée) - Restriction chronologique sur la date d'indexation (recherche avancée : sept options)
Fonctions booléennes	ET, OU, troncature (formulaire, recherche avancée)
Emploi de +/-	Oui
Recherche de locutions	Oui (" " ou formulaire en recherche avancée)
Requête à l'intérieur d'un premier groupe de résultats	Non
Classement des résultats	Pertinence présumée
Affichage par défaut	- Catégories YAHOO! (s'il y en a qui contiennent les mots clés recherchés) - Sites Web répertoriés - Catégorie - Titre - Résumé - Pages Web indexées par INKTOMI (si aucun catégorie ou site; sinon, ce choix est accessible en option) - Titre - Premières lignes - URL - Dépêches d'actualité
Possibilité de modifier l'affichage par défaut	Oui
Choix de la quantité de résultats à afficher	Oui (recherche avancée)
Regroupement des résultats par site (*clustering*)	Oui (pour INKTOMI)
Affichage d'un taux de pertinence	Non
Évaluation de la pertinence	YAHOO! tient compte des critères suivants : nombre de mots clés repérés, exactitude des correspondances, emplacement des mots clés (une

	correspondance dans le titre d'un site a priorité sur la même correspondance dans le commentaire, dans le corps du texte ou dans l'URL), hiérarchie des catégories (parmi les catégories correspondantes, celles qui sont situées en haut de l'arborescence YAHOO! – donc plus générales – sont mieux classées que les catégories inférieures plus précises).
Particularités	- Une requête qui ne donne pas de résultats dans YAHOO! FRANCE est transférée automatiquement vers la base de données du moteur INKTOMI. - Seule l'équipe éditoriale décide des catégories où indexer un site.

Annexe D : ALTAVISTA

URL	http://www.altavista.com/ http://altavista.digital.com/ http://www.av.com/
Catégorie	Moteur
Versions localisées	Oui (3)
Version francophone	Non
Taille de la base de données	Plus de 150 millions d'URL
Possibilité de soumission manuelle d'URL	Oui
Indexation du texte intégral	Oui
Prise en compte d'un fichier *robots.txt* ou d'une balise <ROBOTS>	Oui
Prise en compte des méta-données	Oui
Indexation des cadres	Oui
Présence d'une section de type annuaire	Oui (*Directory* + *Channels*)
Traitement de la casse	Différencié
Traitement des caractères spéciaux et diacritiques	Différencié
Mode de recherche par défaut	OU
Options de restriction de recherche	- *Pages Web* vs *images* vs *fichiers vidéos* vs *fichiers audio* - Sur la langue (25 options) - Dans Usenet - Sur un intervalle de dates (recherche avancée) Les recherches suivantes se tapent en toutes lettres dans la ligne de requête : - Sur le titre - Sur l'URL - Sur le nom de domaine - Sur l'hôte (nom de l'ordinateur) - Sur les noms d'images - Sur les liens hypertextuels - Sur le texte introduisant les liens hypertextuels - Sur les applets Java - Sur le texte intégral d'une page (excluant les étiquettes d'images, les URL et les liens)
Fonctions booléennes	- Troncature - AND, OR, AND NOT, NEAR, parenthèses (recherche avancée)
Emploi de +/-	Oui (uniquement dans la recherche simple)
Recherche de locutions	Oui (" " ou emploi des signes suivants entre les termes : – , . / _)
Requête à l'intérieur d'un premier groupe de résultats	Non
Classement des résultats	Pertinence présumée (dans la recherche avancée, il est possible de désamorcer le tri par pertinence présumée)
Affichage par défaut	- Titre ou la mention *No Title*

	- Balise <DESCRIPTION> ou, à défaut, premières lignes du document - URL - Date de dernière modification - Taille du fichier en KO - Langue du document - Lien offrant la possibilité d'une traduction automatique
Possibilité de modifier l'affichage par défaut	Non
Choix de la quantité de résultats à afficher	Non
Regroupement des résultats par site (*clustering*)	Non
Affichage d'un taux de pertinence	Non
Évaluation de la pertinence	ALTAVISTA tient compte des critères suivants : fréquence d'apparition des mots clés, emplacement des mots clés (une importance spécifique est accordée aux balises META et au titre), nombre de termes trouvés, proximité des mots clés entre eux, liens qui pointent vers une page.
Particularités	- L'usager est invité à formuler ses requêtes en langage naturel. - Le logiciel est équipé d'une technologie de filtrage permettant d'identifier automatiquement les «pages offensantes» (c'est-à-dire celles reliées aux thèmes suivants : drogues/tabac/alcool, discours haineux, jeux d'argent, violence, sexualité explicite). Les internautes sont également invités à participer à cet effort d'épuration en soumettant des URL à bannir. Par défaut, seules les requêtes sur les images et les fichiers sons et vidéos sont filtrées (il est possible de faire filtrer en outre les pages Web, ou de désactiver le tout). Cette technologie ne fonctionne toutefois que pour le contenu d'Internet diffusé en langue anglaise. - ALTAVISTA offre un service de traduction automatique entre l'anglais et cinq autres langues.

Annexe E : ECILA

URL	http://www.ecila.fr/
Catégorie	Moteur
Versions localisées	Non
Version francophone	Outil francophone (sites en provenance de tous les pays)
Taille de la base de données	?
Possibilité de soumission manuelle d'URL	Oui
Indexation du texte intégral	Limitation à 30 Ko
Prise en compte d'un fichier *robots.txt* ou d'une balise <ROBOTS>	Non
Prise en compte des méta-données	Oui
Indexation des cadres	Non
Présence d'une section de type annuaire	Oui (*Guides d'ECILA*)
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Indifférencié pour les caractères diacritiques Les caractères non alphanumériques (# ! " . : /) sont ignorés.
Mode de recherche par défaut	ET
Options de restriction de recherche	- Sur le titre (recherche avancée) - Sur le nom de domaine (recherche avancée) - Sur le nom de fichier (recherche avancée) - Sur les balises <KEYWORDS> et <DESCRIPTION> (recherche avancée)
Fonctions booléennes	ET, OU (formulaire) Lorsque le OU est sélectionné, il est possible de taper dans la ligne de commande les opérateurs ET, OU (qui est alors implicite), PROCHE, ET NON ainsi que les parenthèses.
Emploi de +/-	Non
Recherche de locutions	Oui (formulaire ou " " quand le OU est sélectionné)
Requête à l'intérieur d'un premier groupe de résultats	Oui (possibilité de restreindre la recherche aux titres, aux URL ou aux balises <DESCRIPTION>)
Classement des résultats	Pertinence présumée
Affichage par défaut	- Titre (sinon URL) - Lien permettant d'accéder à un aperçu du document - Nombre d'étoiles (le nombre d'étoiles permet de savoir combien de mots de la requête ont été trouvés dans le document) - Contenu de la balise <DESCRIPTION> ou, à défaut, les premières lignes du texte - URL - Taille du fichier en KO - Date de dernière mise à jour du fichier, si cette info est fournie sur le serveur
Possibilité de modifier l'affichage par défaut	Non
Choix de la quantité de résultats à afficher	Non. De plus, le nombre total de réponses est limité à 200.

Regroupement des résultats par site (*clustering*)	Non
Affichage d'un taux de pertinence	Non (toutefois, les sites les plus pertinents reçoivent un nombre d'étoiles beaucoup plus important que ce que le nombre de mots de la requête ne laisserait prévoir)
Évaluation de la pertinence	Pour une requête simple, la recherche des mots clés s'effectue d'abord dans la balise <KEYWORDS>, puis dans le titre et, enfin, dans le texte intégral. ECILA tient également compte de la présence de tous les termes de la requête, de même que de la concentration des mots clés dans chaque document : ainsi, un document de quelques lignes contenant tous les mots de la question une seule fois sera préféré à un document très volumineux contenant trois fois les mots de la question.
Particularités	- Les mots clés de la requête sont surlignés dans les résultats. - L'emploi du langage naturel est conseillé. - La syntaxe avancée d'interrogation s'inspire de celle d'ALTAVISTA.

Annexe F : EXCITE

URL	http://www.excite.fr/
Catégorie	Moteur
Versions localisées	Oui (8)
Version francophone	Oui (interface localisée)
Taille de la base de données	Plus de 50 millions d'URL
Possibilité de soumission manuelle d'URL	Oui
Indexation du texte intégral	Oui
Prise en compte d'un fichier *robots.txt* ou d'une balise <ROBOTS>	Oui
Prise en compte des méta-données	Non
Indexation des cadres	Non
Présence d'une section de type annuaire	Oui (*chaînes*)
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Différencié
Mode de recherche par défaut	OU
Options de restriction de recherche	- *Web mondial* vs *Web français* (i.e. France uniquement) - Restriction linguistique (recherche avancée : six langues) - Restriction géographique (recherche avancée : huit options)
Fonctions booléennes	AND, OR, AND NOT, parenthèses L'emploi des opérateurs booléens désactive le mode de recherche par concept (voir plus bas).
Emploi de +/-	Oui
Recherche de locutions	Oui (" ")
Requête à l'intérieur d'un premier groupe de résultats	Non
Classement des résultats	Pertinence présumée
Affichage par défaut	- Indice de pertinence - Titre - Lien pour effectuer une recherche sur des documents similaires - URL - Résumé du contenu
Possibilité de modifier l'affichage par défaut	Oui (recherche avancée)
Choix de la quantité de résultats à afficher	Oui (recherche avancée)
Regroupement des résultats par site (*clustering*)	Oui (option)
Affichage d'un taux de pertinence	Oui
Évaluation de la pertinence	Peu d'informations sont disponibles à ce niveau. L'on sait, du moins, qu'EXCITE tient notamment compte du nombre de liens qui pointent vers une page.
Particularités	- Système ICE (*Intelligent Concept Extraction*) : grâce à cette technologie, EXCITE prétend pouvoir effectuer une recherche par concept, c'est-à-dire déterminer les relations entre mots et concepts pour effectuer ensuite une

	recherche d'équivalences. - La fonction *Sites similaires* à côté de chaque URL permet d'utiliser le document concerné comme point de départ d'une recherche pour des documents semblables. - Un résumé est établi automatiquement pour chaque URL à partir des phrases dominantes de la page d'accueil.

Annexe G : HotBot

URL	http://www.hotbot.com/
Catégorie	Moteur
Versions localisées	Non
Version francophone	Non
Taille de la base de données	Plus de 110 millions d'URL
Possibilité de soumission manuelle d'URL	Oui
Indexation du texte intégral	Oui
Prise en compte d'un fichier *robots.txt* ou d'une balise <ROBOTS>	Oui
Prise en compte des méta-données	Oui
Présence d'une section de type annuaire	Oui (HOTBOT Directory)
Traitement de la casse	Différencié
Traitement des caractères spéciaux et diacritiques	Différencié
Mode de recherche par défaut	ET
Options de restriction de recherche	- Sur le titre - Sur un nom de personne - Sur les liens vers une URL - Sur un intervalle de temps (huit options) - Sur la langue (dix options) - Sur l'inclusion dans les page d'images, de fichiers vidéo, de fichiers MP3, de fichiers javascript (recherche avancée : huit options supplémentaires sur la présence de types précis de fichier) - Sur l'extension des noms de fichiers (recherche avancée) - Avant/après une date précise (recherche avancée) - Sur le domaine pour l'Amérique du Nord ou le continent (recherche avancée : 16 options) - Sur le nom de domaine (recherche avancée) - Sur les variantes grammaticales (par exemple, avec cette option, une requête sur *thought* retrouvera des occurrences de *think* et de *thinking*) (recherche avancée) - Sur la catégorie de page (pages d'accueil, pages personnelles, pages d'un niveau de profondeur déterminé) (recherche avancée) Les recherches suivantes peuvent également être exploitées à l'aide de directives tapées en toutes lettres dans la ligne de commande : profondeur de la recherche ; titre ; critères temporels ; présence de type précis de fichiers, de formulaires HTML, de cadres HTML, de tableaux HTML ; domaine pour l'Amérique du Nord ou code de pays (une liste des suffixes de domaine peut être consultée à titre d'aide-mémoire) ; nom de domaine.
Fonctions booléennes	Troncature ET, OU (formulaire) Il est également possible de choisir l'option «Boolean phrase» dans le formulaire, ce qui permet d'utiliser les opérateurs booléens usuels (AND, OR, NOT, parenthèses) en les tapant en toutes lettres dans la ligne de

	commande.
Emploi de +/-	Oui
Recherche de locutions	Oui (" " ou formulaire)
Requête à l'intérieur d'un premier groupe de résultats	Oui
Classement des résultats	Pertinence présumée
Affichage par défaut	- Titre - Premières lignes du texte - Indice de pertinence - Date - URL
Possibilité de modifier l'affichage par défaut	Oui
Choix de la quantité de résultats à afficher	Oui
Regroupement des résultats par site (*clustering*)	Oui (optionnel)
Affichage d'un taux de pertinence	Oui
Évaluation de la pertinence	L'évaluation de pertinence se base sur la fréquence d'apparition des termes (dans le document et dans l'ensemble de la base de données) et leur emplacement (le titre et les balises META – en particulier <KEYWORDS> – sont favorisés). HOTBOT prend également en compte la longueur du document, au sens où un document court recevra une meilleure évaluation qu'un document plus long présentant le même nombre d'occurrences d'un terme donné.
Particularités	Suite à une requête, HOTBOT affiche les catégories pertinentes de son répertoire (s'il y en a) avant les pages Web repêchées par le robot.

Annexe H : INFOSEEK

URL	http://www.infoseek.com/Home?pg=Home.html&sv=FR&svx=INTL_IN_GO_FO_FR
Catégorie	Moteur
Versions localisées	Oui (12)
Version francophone	Oui (interface localisée)
Taille de la base de données	50 millions d'URL selon R. Hock [Hock 1999]
Possibilité de soumission manuelle d'URL	Oui
Indexation du texte intégral	Oui
Prise en compte d'un fichier *robots.txt* ou d'une balise <ROBOTS>	Oui
Prise en compte des méta-données	Oui
Indexation des cadres	Non
Présence d'une section de type annuaire	Oui (NOMADE)
Traitement de la casse	Différencié
Traitement des caractères spéciaux et diacritiques	Différencié
Mode de recherche par défaut	OU
Options de restriction de recherche	- *Tout le Web* vs *France* (i.e. sites en *.fr* seulement) ou choix d'un autre pays (19 options) - Le signe *pipe* (|) permet d'effectuer une recherche avec un mot, pour ensuite réduire les résultats obtenus en utilisant un autre mot, par exemple *fromage* | *chèvre*. - L'emploi des majuscules est suggéré pour les noms de personnes, de lieux, les mots susceptibles d'apparaître entièrement en majuscules, les titres (INFOSEEK considère des mots qui se suivent et qui commencent chacun par une majuscule comme un seul nom ou un titre). Les noms et les titres doivent être séparés entre eux par des virgules. - Sur le titre - Sur le nom de domaine - Sur les liens - Sur l'URL Il convient de noter qu'INFOSEEK FRANCE ne présente pas de page de recherche avancée : sauf pour la première, toutes les options précédentes s'utilisent en tapant les instructions en toutes lettres dans la ligne de commande.
Emploi de +/-	Oui
Recherche de locutions	Oui (" " ou emploi de – entre les mots)
Requête à l'intérieur d'un premier groupe de résultats	Non
Classement des résultats	Pertinence présumée
Affichage par défaut	- Titre - Contenu de la balise <KEYWORDS> ou, à défaut, premières lignes du texte - Indice de pertinence

	- URL - Taille du fichier en KO
Possibilité de modifier l'affichage par défaut	Oui
Choix de la quantité de résultats à afficher	Non
Regroupement des résultats par site (*clustering*)	Non
Affichage d'un taux de pertinence	Oui
Évaluation de la pertinence	Les facteurs suivants sont examinés pour le tri de pertinence : - Présence des termes de la requête dans le titre, dans les balises META ou en début de page - Nombre de termes présents - Présence de termes significatifs (sont considérés comme tels les mots relativement rares dans la base de données de l'outil) - Nombre de liens qui pointent vers une page donnée
Particularités	- L'accès à l'annuaire NOMADE est intégré dans la page d'accueil (affichage des catégories de base).

Annexe I : Lycos

URL	http://www.lycos.fr/
Catégorie	Moteur
Versions localisées	Oui (13)
Version francophone	Oui (interface localisée)
Taille de la base de données	35 millions d'URL selon R. Hock [Hock 1999]
Possibilité de soumission manuelle d'URL	Oui
Indexation du texte intégral	Oui
Présence d'une section de type annuaire	Oui (*Guides du Web*)
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Différencié
Mode de recherche par défaut	ET
Options de restriction de recherche	- *Web français* (i.e. sites en *.fr, .ch* et *.be*) vs *Web mondial* - Recherche d'images (recherche avancée) - Recherche de sons (recherche avancée) - Recherche en langage naturel (recherche avancée) - Sur le titre (recherche avancée) - Sur l'URL (recherche avancée) - Sur un site déterminé (recherche avancée) - Sur la langue (recherche avancée : 15 options)
Fonctions booléennes	AND (ou WITH), OR, NOT, ADJ (mots côte à côte et ordre indifférent), BEFORE (mots côte à côte et ordre respecté), NEAR (fenêtre de 25 mots), FAR (distance d'au moins 25 mots). OADJ, ONEAR et OFAR servent à introduire une notion d'ordre des mots. L'intervalle de 25 mots pour NEAR et FAR peut être modifié par l'usager. De même, on peut faire en sorte d'indiquer un intervalle maximum de séparation des mots pour ADJ (qui est de zéro par défaut), ainsi que pour OADJ, ONEAR et OFAR.
Emploi de +/-	Oui
Recherche de locutions	Oui (" ")
Requête à l'intérieur d'un premier groupe de résultats	Oui
Classement des résultats	Pertinence présumée
Affichage par défaut	- Lien vers le serveur sur lequel est hébergé le document - Titre - Description ou extrait de page - URL - Indice de pertinence - Taille du fichier en KO - Lien pour effectuer une recherche sur des documents similaires
Possibilité de modifier l'affichage par défaut	Non
Choix de la quantité de résultats à afficher	Oui
Regroupement des résultats par site (*clustering*)	En partie seulement : les sites sont regroupés par nom de domaine commun, mais on trouve plus d'une page pour un même site.

Affichage d'un taux de pertinence	Oui
Évaluation de la pertinence	Par défaut, LYCOS considère le repérage de tous les mots de la requête, la présence de mots clés dans les titres et en-têtes (mais non les balises META) et le nombre de liens qui pointent vers une page. La recherche avancée offre à l'usager la possibilité de configurer lui-même l'importance basse, moyenne ou haute à accorder, dans l'évaluation de pertinence, aux éléments suivants : *chercher tous les mots*, *occurrences des mots* (ce critère compare le nombre de fois que le mot recherché figure dans un document avec le nombre moyen d'apparitions de ce mot dans tous les documents de la base de données), *mots proches du début de la page*, *mots proches les uns des autres*, *apparition des mots dans le titre (ou les rubriques)*, *mots dans l'ordre*. Comme ce mode de recherche étudie le poids des différents critères de pertinence les uns par rapport aux autres, il est tout aussi utile de mentionner les critères jugés d'importance «basse» que de mentionner ceux évalués d'importance «haute».
Particularités	- Sous chaque entrée, la fonction *Pour plus de réponses comme celle-ci* permet d'utiliser le document concerné comme point de départ d'une recherche pour des documents semblables. - Un service de traduction d'unitermes (mots simples) en 32 langues est disponible.

Annexe J : VOILA

URL	http://www.voila.fr
Catégorie	Moteur
Versions localisées	Oui (7)
Version francophone	Oui (interface localisée)
Taille de la base de données	100 millions d'URL, dont plus de six millions d'URL francophones
Possibilité de soumission manuelle d'URL	Oui
Indexation du texte intégral	Oui
Prise en compte d'un fichier *robots.txt* ou d'une balise <ROBOTS>	Oui
Prise en compte des méta-données	Oui
Présence d'une section de type annuaire	Oui (*chaînes*)
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Indifférencié pour les caractères diacritiques Les caractères non alphanumériques (# ! " . : /) sont ignorés.
Mode de recherche par défaut	OU
Options de restriction de recherche	- *Web francophone* vs *Web mondial* vs *Newsgroups* vs *Dépêches AFP* - *Sons* vs *images vs vidéos* - Pays francophones ou type d'organisation (10 options) - Type de fichier (recherche avancée : 11 options) (pour le *Web* uniquement) - Mots proches (recherche avancée) - Pays mondiaux ou type d'organisation (recherche avancée : 89 options) - Sur le nom de domaine (recherche avancée) - Recherche thématique à l'aide de moteurs spécialisés[51] (recherche avancée)
Fonctions booléennes	ET, OU, SAUF (formulaire, recherche avancée)
Emploi de +/-	Oui (recherche avancée)
Recherche de locutions	Oui (recherche avancée)
Requête à l'intérieur d'un premier groupe de résultats	Non
Classement des résultats	Pertinence présumée (classement chronologique inverse disponible en option)
Affichage par défaut	- Symbole (boussole ou loupe) permettant d'identifier l'origine des réponses (robot ou chaînes) - Titre - Résumé

[51] VOILA offre une possibilité intéressante et novatrice : celle de pratiquer des recherches thématiques sur près de 30 sujets différents grâce à une technologie algorithmique qui permet, lors de la constitution de la base de données suite aux investigations du robot, de classer automatiquement les pages recueillies à l'intérieur d'une arborescence de thèmes (la base de données du moteur évoque donc un peu la structure d'un annuaire). La restriction thématique d'une recherche peut ainsi s'effectuer en amont (en optant pour un thème spécifique dans le formulaire de recherche avancée) ou en aval (suite à une requête, le moteur propose à l'usager un liste de thèmes susceptibles de correspondre à la thématique de recherche, ce qui permet de filtrer les réponses obtenues dans un premier temps). Cette pratique limite le problème des «false drops» que nous avons évoqué plus haut.

	- Date de dernière modification du document - Taille du fichier en KO - URL - Lien pour atteindre d'autres pages sur le même site
Possibilité de modifier l'affichage par défaut	Oui
Choix de la quantité de résultats à afficher	Oui
Regroupement des résultats par site (*clustering*)	Oui
Affichage d'un taux de pertinence	Non
Évaluation de la pertinence	Un poids est assigné à chacun des mots d'une page Web. Cette pondération tient compte de la fréquence des mots et de leur position dans la page (titre, balises META, gros caractères, gras, italique, etc.).
Particularités	- Les mots clés de la requête sont surlignés dans les résultats. - Les résultats de recherche combinent les pages indexées par le robot et les réponses qui proviennent des chaînes. - Un résumé est constitué automatiquement pour chaque entrée à partir des balises META et/ou du contenu textuel des pages. - Les sites comportant un tilde (~) dans leur URL sont indexés comme pages personnelles[52].

[52] On entend par *page personnelle* tout site édité par une personne physique et qui présente cette personne et ses activités. Ce genre de site peut parfois renfermer, néanmoins, des informations détaillées sur un sujet d'intérêt général (art, sport, région touristique, etc.).

Annexe K : COPERNIC 99

URL	http://www.copernic.com/fr/ (pour téléchargement)
Catégorie	Métamoteur
Versions localisées	Oui (3)
Version francophone	Oui
Outils interrogés	Quelque 40 sources[53] : principaux moteurs et annuaires francophones et anglophones, groupes de discussion, répertoires d'adresses de courriel et sites de vente en ligne de livres.
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Différencié
Mode de recherche par défaut	ET
Options de restriction de recherche	*Le Web* vs *Le Web francophone* vs *Groupes de discussion* vs *Adresses email* vs *Achat de livres*
Fonctions booléennes	ET, OU (formulaire) ET, OU, SAUF (écrits en toutes lettres) sont disponibles pour la fonction *raffiner*, soit lorsqu'on effectue une recherche plus approfondie dans les documents téléchargés.
Emploi de +/-	Non
Recherche de locutions	Oui (" " ou formulaire) Les moteurs de recherche ne supportant pas les guillemets ne sont pas utilisés si ces signes apparaissent dans la requête.
Requête à l'intérieur d'un premier groupe de résultats	Oui
Classement des résultats	Pertinence présumée (tri par titre, date de la recherche ou URL en option) Possibilité de tri ascendant ou descendant
Affichage par défaut	- Titre - Indice de pertinence - Sommaire - Outil(s) ayant repéré le document - URL
Possibilité de modifier l'affichage par défaut	Oui
Choix de la quantité de résultats à afficher	Oui
Regroupement des résultats par site (*clustering*)	Non
Affichage d'un taux de pertinence	Oui
Particularités	- Les doublons sont supprimés, de même que (sur commande) les liens inaccessibles ou invalides. - Les mots clés de la requête sont surlignés dans les résultats. - Un historique détaillé des recherches est accessible.

[53] La version commerciale permet de consulter plus de 140 sources.

	- Les résultats de recherche peuvent être consultés hors ligne, mis à jour et raffinés à l'aide d'opérateurs logiques. - Le logiciel peut télécharger sur le disque dur de l'usager les documents repérés suite à une requête. - Le programme est perpétuellement mis à jour, le téléchargement de la version la plus récente étant automatique au moment de l'utilisation.

Annexe L : DIGOUT4U (version 1.5)

URL	http://www.arisem.com (pour téléchargement)
Catégorie	Agent intelligent
Versions localisées	Oui (2)
Version francophone	Oui
Outils interrogés	Par défaut, DIGOUT4U consulte les outils de recherche les plus usuels (ALTAVISTA, HOTBOT, YAHOO!, etc.) ou les *newsgroups* classiques. On peut également lancer les agents à partir d'une URL spécifique ou d'un fichier au format HTML qui propose des liens vers d'autres pages.
Traitement de la casse	Indifférencié
Traitement des caractères spéciaux et diacritiques	Différencié
Options de restriction de recherche	- Langue des documents - Limitation de la recherche à un seul site - Recherche sur un poste local (fichiers *.htm, .html, .txt*)
Classement des résultats	Pertinence présumée (tri par titre disponible)
Affichage par défaut	- Titre - URL - Indice de pertinence - Taille du fichier en KO - Nombre de citations (dédoublonnage) - Temps mis pour repérer le document L'exportation des résultats inclut les titres, les URL et les notes de pertinence des documents. Dans le cas d'une exportation avec résumé, le fichier généré inclut également, pour chaque document, des extraits pertinents par rapport à la requête initiale.
Possibilité de modifier l'affichage par défaut	Oui
Choix de la quantité de résultats à afficher	Oui
Regroupement des résultats par site (*clustering*)	Oui (optionnel)
Affichage d'un taux de pertinence	Oui
Évaluation de la pertinence	Les documents sont évalués en fonction des concepts présents, indépendamment des termes employés, de la syntaxe et de la langue. Les icônes à côté des documents sourient ou non en fonction du taux de pertinence. Un document avec un taux de pertinence de 30 contient un des éléments du thème de la recherche ; un document noté 90 ou plus contient tous les concepts recherchés.
Particularités	- Il faut formuler les requêtes en langage naturel. - Le système gère indifféremment le français et l'anglais. - Il est possible de filtrer les résultats obtenus selon différents critères (par exemple, pour ne garder que ceux dont la pertinence est supérieure à 50, ou encore les «50 meilleurs»).