Les principaux problèmes de la recherche d'information sur Internet

NB : En outre ce document a été originellement publié sur support électronique (site web) par le service Information et Documentation Scientifique et Technique du Comissariat à l'Energie Atomique. Il fut retiré de la publication en 2002 en raison de son « ancienneté ». Même si les exemples ou les outils spécifiés appartiennent au passé, le fond est encore largement exploitable et constitue une excellente base de départ pour s'informer sur la recherche d'information électronique.

Bruno Bernard SIMON

AVERTISSEMENT : Ce document n'est pas un guide utilisateur de recherche sur Internet. Vous ne trouverez donc pas la syntaxe d'interrogation détaillée pour chaque outil. En revanche, il vous fournira une évaluation du fonctionnement et des fonctionnalités de différents moteurs de recherche qui débouchera sur une étude critique de ces derniers.

Table des matières

Pour toute information, suggestion, commentaire, contacter motrech@www-dist.cea.fr

CEA / DIST - SITE INTERNET mise à jour : 18/05/98

Introduction

La croissance exponentielle d'Internet (image tirée de [NIC96]) en terme de nombre de documents disponibles, et le besoin stratégique de plus en plus important des entreprises d'aujourd'hui de maîtriser l'information disponible nécessitent la mise en place d'outils de veille informative sur Internet toujours plus puissants et surtout toujours plus efficaces.

La structure d'Internet et l'impressionnant volume des ressources qui y sont disponibles forcent à des **choix quantitatifs et qualitatifs** quant au fonctionnement des systèmes de recherche. Privilégier l'aspect qualitatif oblige à ne prendre en compte qu'un nombre restreint de sources d'information, mais permet d'apporter des services à forte valeur ajoutée. Alors qu'un système mettant l'accent sur l'aspect quantitatif effectue typiquement une indexation automatique sur le plus grand nombre de pages possible, mais avec des méthodes élémentaires pour être rapide.

Un certain nombre d'outils existe aujourd'hui et tente de répondre au problème posé. Ces outils possèdent pour la plupart la puissance de calcul nécessaire à l'ampleur de la tâche à accomplir. Par exemple, les machines utilisées par AltaVista sont d'une impressionnante puissance :

- AlphaStation 500 : Mémoire de 256 Mo, disque de 4 Go
- AlphaServer 8400 5/300 : 10 processeur, Mémoire de 4Go, disque de 210 Go
- DEC 3000/900 Alpha Workstation : Mémoire de 1 Go, disque de 30 Go
- AlphaStation 250 4/266 : Mémoire de 196 Mo, disque de 30 Go
- AlphaStation 400 4/233 : Mémoire de 160 Mo, disque de 24 Go

Mais la puissance n'est pas tout, il faut l'utiliser de manière efficace et intelligente. En effet, avant d'avoir l'ambition d'indexer l'ensemble des documents présents sur le WWW (World Wide Web), et de fournir un système de recherche permettant de retrouver l'ensemble des documents pertinents à une question, il est utile et nécessaire de maîtriser quatre composantes essentielles apparaissant dans ce schéma et détaillées plus bas:

1. Le **parcours** le plus **exhaustif** possible d'Internet : aussi bien en terme géographique que linguistique ou temporel.

2. L'**indexation** des documents collectés permettant d'extraire l'information pertinente de ces derniers. Ceci devrait nécessité pour être réellement efficace une extraction intelligente des structures du document (mots, phrases, sigles, ...), une analyse morphologique, une analyse syntaxique, une analyse grammaticale, ... et la prise en compte d'un maximum de jeux de caractères et de langues.

3. Le **traitement de la question**, la **recherche** et le **tri** des documents pertinents pour cette question, qui encore une fois pour être efficace nécessite de mettre en oeuvre des outils linguistiques élaborés.

4. La **présentation des résultat**s la plus riche et synthétique possible, afin que l'utilisateur puisse facilement juger de l'intérêt des documents, et également avoir une image claire et précise du sujet d'intérêt.

Ces quatre points essentiels soulèvent de nombreux problèmes. Nous allons donc poser l'ensemble de ces problèmes à travers les quatre phases nécessaires à une recherche d'information sur Internet :

- La COLLECTE des Documents
- L'INDEXATION des Documents
- La RECHERCHE des Documents-Réponses
- La PRÉSENTATION des Documents-Réponses

Dans ces quatre parties, nous mettrons également en avant pour chacune d'elles les problèmes que se sont réellement posés les créateurs des moteurs de recherche actuels et leurs solutions éventuelles. Dans cette perspective, diverses comparaisons (basées sur des critères mesurables) des outils existants permettront d'évaluer de manière plus précise leurs fonctionnalités et leurs capacités. Ainsi, nous comparerons les différents modes d'interrogation proposés, les méthodes d'élimination des mots vides utilisées, la liste des

opérateurs booléens, de troncature et de proximité/adjacence disponibles, et enfin les informations fournies à l'utilisateur. Ce dernier point sera développé suivant deux comparaisons : la comparaison des informations générales, et la comparaison des informations relatives à chaque document-réponse.

Ensuite, nous verrons une liste détaillée des fonctionnalités des systèmes actuellement disponibles. Cette liste n'est pas exhaustive, tous les moteurs n'y sont pas encore recensés, mais les plus importants y apparaissent.

Enfin, nous conclurons en montrant clairement quel est réellement l'état de l'art actuel en matière de recherche d'information sur Internet. Nous proposerons alors une synthèse des problèmes à se poser et à résoudre pour parvenir à un résultat vraiment intéressant. Pour terminer, nous évoquerons notre angle de vue vis à vis de ces différents problèmes, et nous élaborerons une ébauche des solutions que nous proposons.

1) La Collecte des documents

Dans le contexte actuel, nous pouvons dégager deux méthodes essentielles de collecte des documents : **manuelle** *ou* **automatique**. *La première apporte une valeur ajoutée de sélection, validation et catégorisation des ressources. Elle limite le bruit par rapport aux méthodes automatiques, mais la mise à jour est beaucoup moins rapide et la couverture beaucoup moins large. Enfin, l'intervention humaine biaise la couverture géographique, linguistique ou thématique de la base. Les modules de collecte automatique, plus généralement appelés robots, spiders, crawlers,... doivent résoudre divers problèmes pour effectuer une collecte suffisamment "intelligente" qui permettra des recherches multilingues en langage naturel :*

1. **Reconnaître les jeux de caractères utilisés dans les documents et agir en conséquence**
   Un grand nombre de moteurs de recherche reconnaît et indexe le jeu ISO-Latin 1 ([ALIS96]). C'est un effort appréciable pour le traitement des langues accentuées. Mais ce n'est pas suffisant, car à l'heure actuelle, à cause de cette lacune, un grand nombre de documents (russes, japonais, arabes, ...) ne sont pas référencés par les moteurs de recherche généraux. Et pour cause, tous ces moteurs, comme nous le verrons plus loin ne traitent qu'une seule langue (en général l'anglais) de manière succincte, voir aucune de manière spécifique, alors pourquoi récupérer des documents n'utilisant pas le jeu de caractère occidental ?

2. **Reconnaître la langue du document**
   Ce problème ne se pose pas pour les systèmes actuels, puisqu'ils ne sont pas multilingues, ni même monolingues d'ailleurs (ils ne traitent aucune langue en particulier, ils traitent des chaînes de caractères). Cependant, un système de recherche multilingue doit être en mesure de déterminer automatiquement la langue des documents récupérés, afin de pouvoir effectuer tous les traitements linguistiques propres à cette dernière. A terme, ceci doit permettre de retrouver des documents de langues différentes, à partir d'une question unique formulée dans une seule langue.

3. **Reconnaître les formats de documents**
   Les documents disponibles sur le WWW sont assez hétérogènes. Si on trouve essentiellement des documents HTML (HyperText Markup Language) et des documents ASCII, il ne faut pas oublier qu'il n'est pas rare de trouver des documents PostScript (.ps - représente environ 1,5% des documents sur Internet), des documents au format Acrobat (.pdf - représente environ 0.2% des documents sur Internet), des documents au format Word (.doc - représente environ 0.2% des documents sur Internet),...Les systèmes actuels ne récupèrent pas et donc n'indexent pas tous ces types de documents. Certes, s'ils sont d'une manière générale relativement marginaux sur le WWW comme le montrent les études Measuring the Web ([BRAY96]) et An Investigation of Documents from the World Wide Web ([WOOD96]), dans un contexte Intranet, ces formats de documents sont largement utilisés. Une entreprise désireuse d'adjoindre un Intranet à ses différents moyens de communication interne ne pourra certainement pas se permettre d'effectuer une conversion de l'ensemble de ses fichiers existants vers le format HTML. La solution retenue sera celle couramment utilisée à l'heure actuelle qui consiste à créer des documents HTML référençant les anciens documents au format Word, Excel, ... offrant ainsi un moyen d'accès uniformisé et relativement simple et clair vers l'ensemble de la mémoire de l'entreprise. Il semble donc évident que dans un contexte Intranet, ou Internet/Intranet, la prise en compte des différents

formats de documents bureautiques les plus courants soit inévitable.

4. **Accéder aux pages dynamiques accessibles à partir de formulaires**
   Aucun des moteurs de recherche existant sur le WWW n'accède aux pages dynamiques créées par formulaire. Or, quand on sait qu'un nombre de plus en plus important des documents est accessible de cette manière, et surtout que les bases de données les plus intéressantes sont accessibles uniquement par le biais de ces formulaires de recherche, il est légitime de croire qu'il y a là encore une grande zone d'ombre non explorée par les moteurs de recherche actuels.

5. **Mettre à jour les documents et tester leur validité**
   La mise à jour régulière de l'index de la base est un critère important pour la recherche d'information. Les documents sur Internet changent fréquemment (image tirée de [LUND96]), de nouveaux documents naissent et d'autres disparaissent. Il faut donc, pour ne pas proposer à l'utilisateur des liens vers des documents inaccessibles (le fameux "File Not Found") parcourir régulièrement l'ensemble des documents déjà indexés et remettre à jour la base de donnée. A ce niveau, les principaux outils actuellement disponibles disposent d'une puissance de calcul leur permettant d'effectuer des mises à jour relativement fréquentes.
   Il semble cependant que ces fréquences de mise à jour ne soient pas encore suffisantes, puisqu'il n'est pas rare d'obtenir de nombreux liens non valides à la suite d'une recherche. Il ne faut cependant pas être trop critique à ce sujet. En effet la cause d'un lien non valide peut être le résultat d'un arrêt momentané du serveur, ce qui est difficilement repérable par le moteur de recherche. Il existe pourtant des solutions à ce problème. En effet, il suffirait d'étendre le protocole HTTP (HyperText Transport Protocol) afin que les serveurs puissent informer l'ensemble des sites de recherche de leur arrêt imminent, des modifications effectuées sur les documents, des documents rajoutés ou supprimés. Notons que certains serveurs HTTP commerciaux commencent à proposer une ébauche de solution dans ce sens, mais ces efforts sont marginaux, et ne seront réellement utilisables que lorsqu'une nouvelle norme d'HTTP sera élaborée. Une autre solution moins élégante, mais méritant tout de même d'être testée consiste à vérifier lors de l'affichage de la liste des documents-réponses la validité des différents liens. Il faut cependant craindre une augmentation significative des temps de réponse, ce qui entraînerait l'utilisateur à se lasser du système, si en plus il ne fournit pas des documents suffisamment pertinents.

Nous constatons donc que la collecte des documents, première étape nécessaire à la recherche d'information sur Internet, pose encore de nombreux problèmes. Mais l'un des premiers défis à relever consiste à se poser les questions que nous venons de soulever. En effet, l'étape de collecte des documents est considérée par beaucoup comme triviale, et déjà suffisamment efficace pour être pleinement opérationnelle. Cependant, comme nous venons de le voir, il ne faut pas s'y tromper, de nombreux efforts sont encore à fournir si l'on veut se donner la peine de réellement traiter ce sujet.
Comme le dit Louis MONIER, créateur d'Alta Vista, un document a une réelle existence sur Internet si et seulement si il est référencé par un moteur de recherche. Alors, pourquoi tous ces moteurs de recherche occultent encore un si grand nombre de documents : documents n'utilisant pas l'alphabet latin, documents dynamiques, documents stockés dans des formats propriétaires,...?

2) L'indexation des documents

L'indexation des documents collectés est une étape très importante dans le processus de recherche. En effet, de la qualité de l'indexation dépend la qualité de la recherche. Une "bonne" indexation devrait permettre de constituer un fichier inverse (liste des mots retenus, avec pour chacun d'eux les documents dans lesquels ils apparaissent) regroupant les termes pertinents des documents de la base. Pour cela, il faut tout d'abord parvenir à extraire correctement les mots du document, reconnaître les syntagmes et les expressions idiomatiques, 'tagger' les mots (reconnaître leur genre/nombre, leur fonction grammaticale), lemmatiser les mots (chaque mot est normalisé : belle, belles, beau, beaux seront normalisés en beau; marcher, marchèrent, marcheras seront normalisés en marcher, ...), éliminer les mots vides, extraire les mots "pertinents" d'un document...

Bien sûr, tous ces traitements doivent dans la mesure du possible être disponibles pour un maximum de langues, afin de pouvoir effectuer ensuite des recherches multilingues. Nous décrivons ici des processus et traitements nécessitant une puissance de calcul démesurée pour être appliquée à l'ensemble du

WWW. Les moteurs actuels, comme nous le verrons, sont très loin d'approcher de telles possibilités; ils privilégient une indexation rapide et donc très simple. C'est la raison pour laquelle nous n'avons retenu que quatre critères essentiels:

1. **L'extraction des mots du document.**
   C'est une étape du processus d'indexation qui peut sembler triviale au premier abord, et qui pourtant constituera la base de tout le reste du processus d'indexation. Il faut donc que cette phase soit d'une qualité maximale. L'approche de l'ensemble des développeurs des moteurs de recherche est très simple : ils considèrent un mot comme étant une chaîne de caractères correspondant à l'expression régulière :

   [A-Za-z\-\'$accents]+

   où $accents est une liste de tous les accents possibles dans la table ISO-8859-1 ([ALIS96])

Cette expression régulière signifie qu'un mot est défini comme une chaîne constituée d'au moins un caractère (+) et pouvant contenir :

- n'importe lesquelles des 26 lettres de l'alphabet en majuscule (A-Z),
- n'importe lesquelles des 26 lettres de l'alphabet en minuscule (a-z),
- le tiret (\-),
- l'apostrophe (\'),
- n'importe lesquels des caractères accentués du jeu de caractère ISO-8859-1 ($accents).

Cette définition (simpliste) retenue par AltaVista, InfoSeek Ultra, Excite, AliWeb, OpenText et Carrefour.net entre autres est plus ou moins bien adaptée au traitement de l'anglais (Jerome's Study est découpé en deux mots: Jerome's et Study), mais se révèle catastrophique pour les autres langues. En effet, pour un document contenant les mots L'oeuvre, ces moteurs indexent le terme L'oeuvre comme un seul et même mot. Ce qui signifie que pour retrouver ce document, l'utilisateur devra taper L'oeuvre; s'il tape oeuvre, le document ne sera pas retourné par le système.

2. **La normalisation des mots (lemmatisation)**
   Ce traitement consiste à retrouver pour un mot sa forme normalisée (généralement le masculin pour les noms, l'infinitif pour les verbes, le masculin-singulier pour les adjectifs, ...). Ainsi, dans l'index ne sont conservées que les formes normalisées, ce qui offre un gain de place appréciable, mais surtout, si le même traitement est effectué sur la question, cela permet d'être beaucoup plus "*souple*" et rapide dans la recherche : par exemple, si l'utilisateur effectue une recherche avec un verbe, les documents comportant ce verbe dans toutes ses formes conjuguées seront pris en compte, et pas seulement les documents contenant le verbe dans la forme entrée par l'utilisateur. Peu de moteurs de recherche mettent en place un tel traitement. Ils utilisent parfois la troncature à droite lors de l'interrogation afin de tenter d'approcher ce fonctionnement, mais cette solution est beaucoup plus brutale que la lemmatisation. Voici un tableau regroupant les différentes fonctions de lemmatisation/troncature utilisée par les principaux moteurs de recherche : Les Opérateurs de Troncature (Tableau 4).

3. **L'élimination des mots vides(Tableau 1 : Les Méthodes d'élimination des mots vides).**
   Elle est également importante dans la mesure où c'est un facteur qui a une grande influence dans la précision de la recherche. En effet, le fait de ne pas éliminer les mots vides provoque inévitablement du bruit. L'élimination des mots vides (le, la, et, ...) doit se faire aussi bien à l'indexation qu'à la recherche (élimination des mots vides de la question). Les traitements effectués par les différents moteurs de recherche sont bien sûr monolingues.
   Certains suppriment les mots vides en utilisant la **fréquence d'apparition** des mots dans l'ensemble des documents de la base. Si un mot apparaît trop couramment, il est considéré comme trop fréquent, donc non-informationnel, et il ne sera donc pas pris en compte lors de la recherche. Ce système peut aboutir à des aberrations. Les plus flagrantes sont visibles avec Alta Vista. Par exemple, pour la requête "information retrieval system on the internet", les termes "information", "system", "on", "the" et "internet" apparaissent trop fréquemment dans la base et sont donc ignorés. La recherche ne s'effectue donc que sur le terme "retrieval".

   D'autres (EuroFerret, Galaxy, InfoSeek, Magellan, WebCrawler, WWWWorm) utilisent une **liste des mots vides** (une sorte de dictionnaire) de la langue anglaise. Soulignons le paradoxe d'EuroFerret qui en tant que moteur de recherche européen devrait prendre en considération l'ensemble des langues européennes, mais qui ne possède qu'une liste des mots vides de la langue anglaise (il supprime cependant tous les mots ne dépassant pas un caractère).
   Enfin, des moteurs comme AliWeb, HotBot ou OpenText ne se soucient guère des mots vides...pour eux, tous les mots sont informationnels. D'ailleurs, quelle est la meilleure approche dans un contexte comme le WWW,

où il y a une profusion de documents, dans un multitude de langues différentes, gérer approximativement les mots vides dans une seule langue, ou bien ne pas les gérer du tout ?

4. L'indexation proprement dite consiste à retenir les termes les plus significatifs du document. Ici encore, les moteurs de recherche actuels ne "*font pas dans la finesse*", ils indexent tous les mots!!! A priori, Lycos est l'un des seuls à retenir un certain nombre de mots significatifs (il en retient une vingtaine)...Mais nous n'avons malheureusement aucune information sur les critères lui permettant de retenir tel ou tel mot. Cette aberration de l'indexation va beaucoup plus loin puisqu'il semblerait que certains moteurs qui gèrent les mots vides, indexent tout de même ces derniers, et ne les éliminent en fait que dans la requête de l'utilisateur.

Comme nous venons de le voir, l'indexation réalisée par les systèmes de recherche actuels reste très basique, mais en revanche très rapide. Cette rapidité d'indexation leur permet d'obtenir une fréquence de mise à jour élevée de leur index pour une couverture très large des ressources disponibles sur Internet. Cependant, la qualité de l'indexation est médiocre, ce qui conduit généralement à des systèmes à la fois fortement bruités et fortement silencieux. Si l'exhaustivité des documents indexés sur Internet permet de fournir un certain nombre de documents pertinents, dans un contexte Intranet, de telles méthodes d'indexation ne semblent pas viables.

3) La recherche des documents

La **recherche des documents** *a pour but de faire ressortir les documents de la base les plus pertinents pour la question posée. Nous distinguons deux aspects importants dans cette étape : le mode de formulation de la requête par l'utilisateur (requête par liste de mots clés, requête booléenne, requête en langage naturel), et le méthode interne utilisée par le moteur de recherche (recherche booléenne simple, recherche booléenne raffinée, recherche en langage naturel). Une première partie décrira donc les différents modes d'interrogation existants, et une deuxième les méthodes de fonctionnement interne des moteurs de recherche.*

Les Différents Modes d'Interrogation

Trois types d'interrogation sont envisageables : L'interrogation booléenne, l'interrogation par liste de mots et l'interrogation en langage naturel.

1. L'interrogation Booléenne (Tableau 3 : Les Opérateurs Booléens) qui est en fait le type d'interrogation le plus basique, et surtout le plus simple à implémenter. Il consiste à formuler une question avec une liste de termes séparés par des opérateurs logiques (AND, OR, NOT), et à rechercher les documents correspondant à cette requête. Par exemple, pour la question "(moteur AND recherche) OR (search AND engine)"(les systèmes actuels n'étant pas multilingues, c'est le seul moyen d'effectuer une recherche à la fois en français et en anglais!), le système doit fournir comme réponse tous les documents de la base contenant les deux termes "moteur" et "recherche" ou "search" et "engine". A l'heure actuelle, un grand nombre des systèmes disponibles sur Internet fonctionne suivant ce type de recherche (Alta Vista, Excite, Galaxy, Harvest, HotBot, InfoSeek, Lycos, Magellan, OpenText, WebCrawler, WWWWorm). Notons tout de suite pour les utilisateurs non avertis que la syntaxe "+/-" utilisée par certains moteurs de recherche est tout à fait similaire à celle des opérateurs logiques. En effet, ces systèmes considèrent les coupures de mots comme des opérateurs OU implicites, l'opérateur "+" revient à un ET, puisqu'il permet d'imposer la présence d'un mot, et l'opérateur "-" permet de proscrire la présence d'un mot (équivalent au NON).

   Quelques raffinements sont fournis par certains systèmes, afin de palier aux limitations de la recherche booléenne :

   - La **combinaison des opérateurs** (comme dans notre exemple) va permettre d'effectuer des recherches un peu plus complexes que celles proposées par Open Text, WebCrawler ou WWWWorm qui ne permettent que d'utiliser un seul opérateur entre les différents mots. Ainsi, vous pouvez rechercher "Moteur AND recherche AND search AND engine", ou "Moteur OR recherche OR search OR engine", mais vous n'avez aucun moyen d'effectuer une recherche avec une question du type "(Moteur AND recherche) OR (search AND engine)". C'est une limite très contraignante pour

effectuer une recherche booléenne efficace, et c'est donc un point très important pour un système booléen de disposer de cette fonctionnalité.

- Le **parenthésage** des expressions permet de compliquer un peu plus les requêtes, et permettra ainsi à l'utilisateur averti d'effectuer des recherches complexes. Cette possibilité, qui pourtant semble basique et tout à fait logique dans un contexte booléen n'est cependant disponible que dans ce type de moteurs de recherche (Alta Vista, Excite).

- La **troncature** (automatique et/ou manuelle) permet de rechercher des sous-chaînes. L'opérateur de troncature (généralement "*") remplace un ensemble de caractères afin d'effectuer des recherches plus larges. On distingue trois types de troncature, la troncature droite (la plus commune), la troncature gauche, et enfin la troncature interne.
  La troncature droite permet d'effectuer une recherche en utilisant le début d'un mot, afin d'obtenir les différentes formes dérivées du mot (et d'autres mots n'ayant aucun rapport avec ce que recherche l'utilisateur!). Par exemple, si vous recherchez des documents sur *les chats*, vous pouvez saisir la requête "chat*" afin d'obtenir les documents contenant les termes chat, chats, chatte, chattes, chaton, chatons, ... Le résultat va cependant être surprenant, puisque les documents contenant les termes français chatoiement, chatouille, chatterton vous seront également retournés. Mais ça ne s'arrête pas là, puisque le WWW contient essentiellement des documents de langue anglaise, vous obtiendrez de nombreux documents contenant les mots chat(bavardage), ou chattel(bien mobilier), qui n'ont rien à voir avec votre recherche.
  Les troncatures gauche et interne fonctionnent de la même manière, elles permettent respectivement de rechercher des chaînes sans spécifier le début du mot, ou une partie interne du mot.
  Tous ces opérateurs sont à manipuler avec précaution. En effet, en lançant une recherche avec troncature dans un environnement multilingue comme le WWW, le bruit devient très rapidement trop important pour pouvoir exploiter les résultats obtenus.

  Le Tableau 4 (Les Opérateurs de Troncature) montre les différentes troncatures utilisées par différents moteurs de recherche. Il faut noter que la colonne lemmatisation dans ce tableau n'est pas une réelle lemmatisation, mais un traitement un peu plus évolué que la troncature. Par exemple, pour EuroFerret, la *lemmatisation* consiste à supprimer des mots tous les suffixes s, e, es, ing, ent et à rajouter un opérateur de troncature droite. Nous ne pouvions pas classer un tel traitement dans la troncature droite, mais il faut tout de même mentionner que de l'appellation lemmatisation est un réel abus de langage. En effet, aucun moteur de recherche n'effectue réellement de lemmatisation, et encore moins sur plusieurs langues. Ils utilisent généralement la troncature droite couplée à un ensemble sommaire de règles pemettant de reconnaître quelques terminaisons régulières communes et quelques pluriels irréguliers (pour une seule langue).

- Les opérateurs de **proximité** et d'**adjacence** (Tableau 5 : Les Opérateurs de Proximité/Adjacence). Ce type d'opérateurs a pour but de contraindre le système à rechercher des mots se trouvant proches l'un de l'autre, et donc étant supposés avoir des liens syntaxiques et/ou sémantiques. Ainsi, en posant la question "moteur ADJ recherche", vous allez éliminer un certain nombre de documents concernant la recherche dans le domaine des moteurs automobiles ou aéronautiques. Mais il ne faut pas s'y tromper, cet opérateur ne permet pas de spécifier qu'il doit exister un lien syntaxique entre les mots. Il va considérer qu'un document est pertinent si les mots moteur et recherche ne se trouvent pas séparés dans le texte par plus de 2, 3, 4, 5 ... 10 mots. La limite du nombre de mots est soit déterminée arbitrairement par le système, soit spécifiable par l'utilisateur suivant les moteurs de recherche. La recherche de syntagmes nominaux (phrases) vous permet de rechercher une chaîne de caractère exacte. Par exemple, la question "moteur de recherche" ne vous retournera comme résultat que les documents contenant exactement la chaîne "moteur de recherche".

Il n'est pas besoin d'être un spécialiste des systèmes de recherche d'information pour se rendre compte des limites des systèmes booléens. En fait, leur fonctionnement interne se résume à une simple recherche de chaîne de caractères. Que cette chaîne soit ou non syntaxiquement correcte, peu importe (Essayez d'ailleurs de poser des questions n'ayant aucune signification aux moteurs

de recherche actuels, vous serez étonnés des résultats. Des questions comme "aaaaa", "aieru", ou ":-)" ont de grandes chances de vous retourner des résultats !!!), si elle est présente dans le document, le système la retrouvera. L'utilisation de requêtes booléennes peut rapidement devenir complexe, et finalement, sur Internet, qui utilise ces opérateurs pour effectuer des recherches ? Les documentalistes certainement, les informaticiens peut-être, l'utilisateur lambda sûrement pas!

2. La Recherche par Liste de Mots (ou pseudo Langage Naturel) permet de s'affranchir d'utiliser un langage d'interrogation pour effectuer une recherche. Ces dernières sont donc plus simples à formuler. Plus simples, mais également plus impécises, et donc plus bruitées. En effet, ce mode d'interrogation souvent proposé par les moteurs (AliWeb, AltaVista, EuroFerret, Excite, Galaxy, HotBot, InfoSeek, Lycos, Magellan, OpenText, WebCrawler, WWWWorm) en complément de la recherche booléenne n'est en fait qu'une surcouche logicielle de cette dernière. La requête de l'utilisateur est retranscrite par le système en une expression booléenne suivant un schéma précis pré-établit (ET implicite, OU implicite, troncature droite implicite, etc.).

   En somme, même si la recherche à partir de tels systèmes est à priori plus simple que la recherche à partir d'une requête booléenne, il est néanmoins nécessaire que l'utilisateur connaisse les traitements de reformulation de la question effectués par le moteur. Il pourra ainsi adapter sa requête au fonctionnement interne de ce dernier et il aura une meilleure compréhension de résultats obtenus.

3. La recherche en Langage Naturel, de loin la mieux adaptée au texte va permettre à l'utilisateur de formuler une question totalement libre (en langage naturel = le langage commun de "*tous les jours*") pour effectuer sa recherche. Une telle recherche nécessite une indexation et une recherche *"intelligente"* mettant en oeuvre des modules de traitements linguistiques élaborés. Aucun système de recherche sur Internet ne dispose à l'heure actuelle d'une telle puissance de traitement du langage.

Méthodes de Fonctionnement Interne

Nous envisageons ici trois types de fonctionnement interne des moteurs de recherche : la recherche booléenne, les systèmes vectoriels et enfin les systèmes pondérés autre que vectoriels.

1. La Recherche Booléenne, comme nous l'avons évoqué plus haut est un système basique de recherche. Il repose sur une simple comparaison de chaînes de caractères. Le moteur recherche dans son fichier inversé les mots correspondant à ceux de la requête, tout en respectant les opérateurs séparant les différents termes. L'interrogation est réalisée en exprimant le besoin de l'utilisateur par une fonction booléenne de descripteurs. Ainsi, les documents réponses sont ceux dont l'index fournit la réponse «vrai» à cette fonction. Nous ne reviendrons pas sur les différentes fonctionnalités dont peuvent être agrémentés certains systèmes (combinaison des opérateurs, parenthésage, troncature, proximité/adjacence). En revanche, il faut noter que pour ces systèmes, deux méthodes de fonctionnement interne sont envisageables :

   - La recherche booléenne simple qui se conforme à retrouver les documents contenant les chaînes de caractères correspondant à la requête de l'utilisateur en respectant les restrictions des différents opérateurs. A défaut d'opérateur entre les mots, ce système va utilisé un opérateur implicite qui sera en général le OU logique afin de ne pas trop restreindre la recherche et retourner une liste vide de documents. Il est cependant utile de savoir que quelques systèmes tels que Lycos ou HotBot utilisent l'opérateur ET comme opérateur implicite.

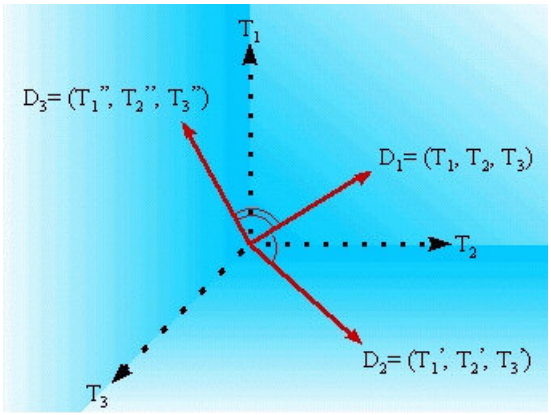
   - La recherche booléenne avec dégradations successives de la question. Cette méthode consiste à reformuler la question de l'utilisateur avec des opérateurs ET entre les termes, et à effectuer plusieurs dégradations de la question de l'utilisateur en insérant des opérateurs permettant d'élargir la recherche (typiquement OU, troncature).

   Nous constatons donc qu'un système de recherche booléen n'est pas en mesure d'évaluer la pertinence d'un document par rapport à une question. Il n'y a que deux états possibles, le

document répond à la question, ou il n'y répond pas. Le moteur de recherche retourne donc une liste de documents réponses pertinents à la question, mais ils ne sont pas triés les uns par rapport aux autres. Comme nous le verrons dans la partie consacrée à l'organisation des documents réponses, plusieurs méthodes sont utilisées par les moteurs booléens actuels pour pallier ce manque et fournir une mesure de pertinence : nombre d'occurences des mots de la question dans le document, présence ou absence des mots de la question dans le titre du document, popularité du site auquel appartient le document, etc. Toutes ces méthodes semblent assez approximatives et leurs fondements sont assez douteux.

2. Les Systèmes Vectoriels sont basés sur le modèle de Salton [SALT74] qui peut brièvement être décrit comme suit :

   Considérons un espace à n dimensions constitué de D_i documents identifiés par un ou plusieurs termes d'indexation T_j. Les termes peuvent êtres pondérés suivant deux méthodes : discrète (0 : le document ne contient pas le terme, 1 : le document contient le terme) ou continue (selon l'importance du terme d'indexation dans le document, T_j prendra une valeur plus ou moins importante entre 0 et 1). Par exemple, un espace d'indexation à trois dimensions a l'aspect suivant :



Nous pouvons alors étendre ce modèle tri-dimensionnel à t dimensions, où t est le nombre de termes d'indexations utilisés. Nous représentons alors chaque document D_i par un vecteur à t dimensions :

$D_i = (d_{i1}, d_{i2},..., d_{it})$,

où d_{ij} représente le poids du $j^{ème}$ terme. Ainsi, à partir des vecteurs d'indexation de deux documents D_i et D_j, nous pouvons calculer un coefficient de proximité $s(D_i, D_j)$ entre D_i et D_j. Le coefficient de proximité obtenu $s(D_i, D_j)$ prend en compte la similarité des termes utilisés dans les deux documents D_i et D_j tout en tenant compte du poids des termes dans chacun d'eux. Une telle fonction de similarité peut être définie comme une fonction inverse de l'angle entre les deux vecteurs (typiquement le cosinus). Ainsi, quand les termes d'indexation sont identiques dans les deux documents, l'angle est nul et donc la mesure de proximité est maximale.
Pour effectuer une recherche dans un espace à t dimensions composé de n documents , il suffit donc de modéliser la requête de l'utilisateur par un vecteur $Q = (q_1, q_2,..., q_t)$ où q_i représente le poids du $i^{ème}$ terme et de calculer chaque degré de proximité $s_j(Q, D_j)$, où j représente le $j^{ème}$ document ($j=\{1;2;...;n\}$).

Ce modèle, contrairement au modèle booléen, permet donc de pondérer les résultats. Il est possible de calculer un degré de similarité entre chaque document de la base, mais aussi entre une question et l'ensemble des documents. Or, comme le mentionne Salton [SALT74], un problème majeur de ce modèle réside dans le fait que ses performances sont fortement liées à la topographie de l'espace créé lors de l'indexation.

La quasi totalité des outils de recherche sur Internet utilisent un système de recherche booléen, et il semblerait que l'un des seuls à utiliser le modèle vectoriel soit WebCrawler.

3. Les Systèmes Pondérés autre que Vectoriels. Nous regroupons ici tous les autres systèmes de recherche pondérés. Cependant, nous n'allons ni les énumérer ni les détailler. En effet, il y a deux raisons à cela. Tout d'abord, les méthodes utilisées peuvent enormément varier d'un système à l'autre, et ensuite, il n'y a pas de moteurs de recherche sur Internet utilisant de tels méthodes. Nous allons tout de même illustrer cette partie par un exemple de système de recherche pondéré non vectoriel afin de pouvoir comprendre de quoi il s'agit. Nous avons choisit le système SPIRIT dont nous résumons ici de manière très sommaire la méthode de calcul d'un degré de similarité entre les documents. En fait, SPIRIT calcule un coefficient de proximité sémantique entre les documents. Pour cela, il utilise à la fois des méthodes statistiques (calcul du poids des termes par rapport à leur apparition dans la collection de documents), et linguistiques (normalisation des mots, extraction des mots composés, analyse des liens syntaxiques, etc.).

4) La présentation des résultats

C'est par le bais de la présentation des résultats que l'utilisateur fait une première évaluation des documents qui lui semblent intéressants. Les informations affichées doivent permettre deux choses essentielles : comprendre globalement comment le moteur de recherche fonctionne (quels ont été les mots considérés comme vides, quels ont été les mots retenus, les mots inférés par le système, quels mots ont été retrouvés par le système dans les documents-réponses), et évaluer facilement la pertinence des documents-réponses sans avoir besoin de tous les consulter pour vérifier.

Une présentation des résultats présentant de telles qualités, couplée à un système de recherche efficace doit permettre de créer un climat de confiance entre l'utilisateur et le système qu'il utilise. C'est en créant un telle relation que la recherche devient réellement efficace. Les critères que nous avons retenus concernent les informations délivrées par le système sur son fonctionnement et sur les documents, l'organisation des documents-réponses, et enfin les moyens d'apporter une valeur ajoutée à la liste des documents.

1. **Quelles sont les informations que doit afficher le système?**
   Ce problème doit être vu suivant deux angles différents : Les informations générales concernant la recherche, et les informations portant sur chaque document-réponse.

   - **Au niveau des informations générales (Tableau 6 : Les Informations Générales du Résultat) concernant la recherche, nous en avons dégagé trois qui nous semblent inévitables :**

     1. Le nombre de documents retrouvé par le système pour la recherche. Cette information va permettre à l'utilisateur d'évaluer la pertinence de sa question. Si ce nombre est trop élevé, il va essayer de préciser sa question et de relancer une recherche. En revanche, s'il est relativement faible, il va certainement reformuler sa question de manière un peu plus générale afin de ne pas "*passer à côté*" de documents intéressants.
        Cette information est généralement fournie par l'ensemble des outils disponibles sur Internet. Seuls les moteurs AliWeb et WWWWorm ne jugent pas utile de fournir cette information à l'utilisateur.

     2. Le rappel de la question posée par l'utilisateur. Cette information ne semble a priori pas extrêmement importante, puisque l'utilisateur sait quelle question il a posée. Certes, mais c'est une information importante et même essentielle du contexte de l'interrogation, et elle se doit d'être présente sur la page de résultat. Prenons un exemple concret : L'utilisateur imprime la page des résultats, comment pourra-t-il se souvenir un mois plus tard à quelle question correspond cette liste de documents si la question n'apparaît pas?
        S'il y a bien un point sur lequel les différents moteurs de recherche sont unanimes, c'est celui du rappel de la question. En effet, tous les moteurs fournissent cette information de manière

plus ou moins évidente.

3. Les listes des termes reconnus, non-reconnus et ignorés par le système, qui doivent permettre à l'utilisateur de mieux comprendre comment sa question a été traitée et donc de mesurer l'adéquation entre sa question telle qu'il se l'était formulée et telle qu'elle a été interprétée par le système. Pour pouvoir distinguer des termes reconnus et des termes non-reconnus dans le lexique de la langue ou du domaine, le système doit fonctionner avec un dictionnaire (ou thésaurus), ce qui n'est pas le cas des systèmes actuels qui considèrent toutes les chaînes de caractères comme étant valides. Ils ne sont donc qu'en mesure de fournir une liste des termes présents ou absents de la base. Or, comme les moteurs indexent pour la plupart tous les mots, il est très rare que des mots ne soient pas reconnus, et l'information *"termes reconnus/non-reconnus"* perd son intérêt pour évaluer l'adéquation entre le vocabulaire de l'utilisateur et celui utilisé dans la base. Il y a donc très peu de moteurs de recherche qui fournissent de telles listes.
   Alta Vista donne la liste des termes non-ignorés, ainsi que celle des termes ignorés (car trop fréquents dans les documents). Il ne possède pas la notion de termes reconnus ou non reconnus.
   Quand à Lycos, il fournit à l'utilisateur une liste partielle des termes qui ont servis à la recherche. Dans cette liste ne figurent pas les mots vides, mais par contre un ensemble de mots dérivés de ceux de la question à partir d'une troncature ou pseudo-lemmatisation.
   Enfin, WWWWorm affiche la liste des mots retenus par le système. C'est à dire la liste des mots de la question moins la liste des mots vides.
   Les autres moteurs de recherche disponibles sur Internet ne fournissent aucune information de ce type à l'utilisateur. Ce dernier ne dispose donc d'aucun moyen d'avoir une vision globale des résultats de sa recherche. Il devra consulter les informations relatives à chaque document, si elles existent...

- **Au niveau des informations propres à chaque document (Tableau 7 : Les Informations sur les Documents), nous avons retenu 9 critères importants pour une bonne évaluation du contenu des documents et de leur pertinence :**

1. L'URL (Uniform Resource Locator) du document. Ceci dans le but de connaître la localisation du document, le type de protocole qu'il utilise, et de quelle organisation il provient. De plus, ceci permet de disposer de l'adresse du document après une impression de la page de résultat. Quasiment tous les moteurs de recherche fournissent l'URL du document-réponse, mis à part AliWeb, EuroFerret, WWWWorm et WebCrawler qui l'offre en option.

2. Un lien hypertexte vers le document d'origine (ça semble être la moindre des choses!) afin de pouvoir le consulter. L'ensemble des outils disponibles affichent un lien vers le document trouvé.

3. Tout comme le lien vers le document d'origine, le titre de ce dernier paraît être une information vitale pour un minimum de souplesse d'utilisation, et là encore, tous les moteurs de recherche nous fournissent le titre des documents-réponses.

4. Un bonne majorité des moteurs affiche un extrait du document. Il s'agit bien généralement des deux ou trois premières lignes de ce dernier...ce qui nous donne un texte souvent très peu informationnel. Excite, quant à lui nous propose un *"résumé automatique"*, qui se révèle être à défaut de résumé une collecte de trois ou quatre phrases prises ça et là dans le document. Les seuls outils fournissant un résumé intéressant sont bien entendu ceux pour lesquels ce dernier est rédigé manuellement (AliWeb, Magellan), mais en contre partie, très peu de documents sont disponibles avec un résumé.
   Comme nous le voyons, certains choisissent de privilégier la quantité plutôt que la qualité (extrait des documents), d'autres font l'inverse (résumé manuel). Il serait donc intéressant de se poser le problème de réussir à combiner quantité et qualité, c'est à dire parvenir à générer pour chaque document un résumé automatique. Divers travaux existent dans ce domaine, et les résultats semblent réellement exploitables. Il serait donc intéressant d'étudier s'il serait possible d'adapter de tels systèmes aux contraintes imposées par Internet.

5. La taille du fichier qui n'est pas une donnée vitale est fournie par la majorité des moteurs de

recherche. Elle permet tout de même de se rendre compte si le document a plutôt la taille d'un résumé, ou d'un livre entier (bien que cette taille soit toujours exprimée en octets ou kilo-octets, ce qui n'est pas forcément très significatif pour un non-informaticien).

6. La date de dernière visite du document par le robot de collecte donne à l'utilisateur une bonne idée de la "*fraîcheur*" de l'index et donc de la pertinence de la recherche. Les systèmes proposant ce type d'information sont peu nombreux, on trouve Alta Vista, EuroFerret, et HotBot.

7. La date de dernière mise à jour du document. Cette information qui est liée à la précédente permet de connaître la date de dernière modification du document. C'est une information clé dans un soucis de veille informationnelle. En effet, on peut très bien imaginer de créer une interface permettant d'effectuer des recherches mais ne fournissant à l'utilisateur que les documents nouveaux, ou ceux ayant changés depuis la dernière interrogation. Il n'y a pourtant que Galaxy qui considère que cette information ait autant d'importance.

8. Une mesure de pertinence permettant d'évaluer approximativement la pertinence d'un document par rapport à un autre. Il faut néanmoins se méfier de ces valeurs souvent très subjectives mesurées de manière parfois très originales et différentes selon les moteurs. Nous verrons ce point plus en détail par la suite (Comment organiser les documents ?).

9. La mise en évidence des mots de la question dans les documents-réponses est le moyen le plus commode de se rendre compte de la pertinence d'un document. En effet, si le système vous indique que tel document contient quatre des mots de la question, alors qu'un autre n'en contient qu'un, le premier semble beaucoup plus proche de ce que vous recherchez. Aide indéniable pour l'utilisateur dans sa recherche documentaire, ce service n'est malheureusement disponible qu'avec Lycos et EuroFerret.

2. **Comment organiser les document ?** Le classement des documents-réponses fournis par les systèmes de recherche est un point important pour l'utilisateur. En effet, si ce classement est performant, l'utilisateur trouvera dans les premiers documents retournés par le système un nombre important de documents pertinents. Il sera alors amené à faire confiance au système, et il gagnera énormément de temps en ne consultant que les premières réponses du système. Bien sur, il n'existe pas de critères absolus pour définir la pertinence d'un document, et deux utilisateurs différents ne jugeront pas les mêmes documents comme pertinents pour une même question, même si le nombre de documents communs sera relativement important. Les moteurs de recherche disponibles sur Internet utilisent tous des heuristiques plus ou moins simplistes, voir "*exotiques*". En effet, les critères retenus pour donner du poids à un mot de la question sont :

   - Le nombre d'occurrences du mot dans le document
   - La présence du mot dans le titre du document
   - La présence du mot dans l'URL du document
   - La présence du mot dans les premières lignes du document
   - L'adjacence des mots de la question dans le document
   - Le nombre faible d'occurrences du mot dans la base. Ce critère provient du postulat que plus un mot est rare plus il est informationnel.
   - La popularité du site dans lequel les mots de la question sont trouvés. Un site est populaire s'il est référencé par de nombreux documents "*extérieurs*".

   Comme nous le voyons, les méthodes de classement varient énormément d'un moteur à l'autre. Aucune ne semble donner de bons résultats. En effet, avec de telles méthodes de tri des documents-réponses, il n'est pas rare d'obtenir en premiers résultats d'une recherche au sujet des moteurs de recherche des documents portant sur des travaux de recherche sur les moteurs automobiles ou aéronautiques.

3. Quelles sont les **informations à valeur ajoutée** que l'on peut envisager de fournir à l'utilisateur?

- Un certain nombre de recherches porte actuellement sur les systèmes graphiques de navigation sur Internet. De telles représentations graphiques devraient permettre de synthétiser un grand nombre d'informations sur les documents et surtout sur leur contexte, permettant ainsi à l'utilisateur d'avoir une vision claire des résultats obtenus. Le Xerox PARC (Xerox Palo Alto Research Center) est dans ce domaine l'un des acteurs les plus actifs. L'ensemble de leurs recherches et développements menés au niveau de la représentation graphique et de la visualisation de l'information est très prolifique. Celà va de la visualisation en arbre conique permettant de montrer de vastes hierarchies en répartissant les noeuds dans un espace 3D, jusqu'au "Fouille Web" (Web Forager) utilisant également un espace 3D pour permettre à l'utilisateur de rassembler et ranger des pages WWW de manière souple et informelle, en passant par le Table Lens ou le Document Lens qui offrent à la fois une vue synthétique et détaillée de l'information.

- La possibilité d'effectuer de nouvelles recherches à partir des documents-réponses sélectionnés est une fonctionnalité que l'on trouve dans certains moteurs de recherche (EuroFerret, Excite, WebCrawler). Son but est de pouvoir, à partir d'un ou de plusieurs des documents obtenus en résultat, de lancer une nouvelle recherche afin de trouver les documents similaires. Il s'avère que les résultats sont actuellement très peu convaincants. La recherche à partir de documents entiers provoque des résultats énormément bruités et donc inexploitables.

- Plusieurs moteurs de recherche tentent de mêler l'aspect moteur de recherche et taxinomie (Galaxy, InfoSeek, Magellan). C'est un point très intéressant qui mériterait d'être mieux géré et exploité. En effet, il semble tout à fait intéressant de pouvoir effectuer en même temps des recherche à deux niveaux. A la fois en effectuant des recherches par questions (structure de recherche souple), et en naviguant dans la structure thématique (structure de recherche rigide). Il est certain que la taxinomie ne peut pas être exhaustive, mais elle doit référencer les plus grands cites clés pour un domaine (les grandes organisations : NASA, CERN, ...) afin d'accéder à des documents de référence. Couplée à la taxinomie, les moteurs en langage naturel permettrait d'affiner la recherche d'information et d'accéder plus souplement et plus efficacement à l'ensemble des documents pertinents concernant le sujet d'intérêt de l'utilisateur.

Conclusion

La recherche d'information

Les systèmes de recherche d'information ont connu plusieurs évolutions majeures :

- Les systèmes booléens (années 60-70)
  - Indexation sur les chaînes de caractères.
  - Utilisation d'une liste de mots vides.
  - Troncature des termes.
  - Opérateurs booléens et de proximité entre les chaînes de caractères.

- Puis les modèles statistiques (années 70-80)
  - Calcul d'une mesure de pertinence des documents et tri (*"relevance ranking"*).
  - Modélisation vectorielle des documents et des questions.
  - Calculs de probabilité pour l'indexation et la recherche.
  - Algorithmes de clusterisation des documents.

- Enfin, le traitement du langage naturel et les systèmes à base de connaissances (depuis les années 80).
  - Traitements morphologiques, syntaxiques et sémantiques.
  - Questions en langage naturel.
  - Reconnaissance des racines, lemmatisation.
  - Repérage des syntagmes (mots composés).
  - Reformulation automatique de question.
  - Recherche par concept.

Actuellement, en 1997, comme nous l'avons souligné tout au long de ce document, les robots généralistes de recherche d'information sur Internet utilisent:

- Le traitement de chaîne.
- La logique booléenne, les opérateurs de proximité, d'adjacence et de troncature.
- Le *"relevance ranking"*.

C'est à dire des techniques développées pour les premiers systèmes de recherche datant des années 60-70. Cela montre clairement qu'aujourd'hui, le spécialiste du WWW redécouvre le domaine de la recherche documentaire.

Internet et l'information

Comme le mentionne à juste titre Ghislaine CHARTRON [CHAR96], poser le problème de la recherche d'information sur Internet nécessite avant tout de connaître les caractéristiques majeures de cet espace d'information. C'est une condition nécessaire pour parvenir à se poser les questions pertinentes et à apprécier avec critique les systèmes actuellement disponibles. Nous pouvons ainsi dégager 7 caractéristiques majeures de l'information sur Internet :

1. une grande hétérogénéité dans le contenu et dans le codage numérique,
2. une instabilité des localisations,
3. une fragmentation plus ou moins importante,
4. un multilinguisme,
5. un renouvellement continuel,
6. un caractère public et commercial,
7. une information en grande majorité non structurée.

L'ensemble de ces caractéristiques nous montre clairement que la recherche d'information sur Internet est une problématique plus complexe que la recherche d'information sur un espace plus circonscrit tel que le Minitel ou les banques de données sur un serveur.

Reprenons une par une les spécificités de l'information disponible sur Internet, et étudions à travers ce que nous avons vu dans les chapitres précédents comment les systèmes de recherche actuels s'adaptent à ces caractéristiques:

La grande hétérogénéité dans le contenu et dans le codage numérique
Nous l'avons évoqué lors de la collecte des documents, et nous avons vu qu'aucun moteur de recherche d'information actuel ne s'adaptait aux codages non latins tels que le russe, le japonais ou l'arabe. Dans ce même chapitre, nous avons également indiqué que ces systèmes n'étaient pas en mesure d'indexer les documents dans un autre format que le format HTML. Les documents de type .pdf, .doc, .ps, etc ne sont donc pris en compte d'aucune manière.

L'instabilité des localisations
Les moteurs de recherche traitent ce problème par leur fréquence de visite des pages Web. Ainsi, si un document est déplacé, il va être une nouvelle fois indexé, mais aucun système ne sera en mesure de spécifier à l'utilisateur que le document est le même que celui qui se trouvait anciennement à une autre localisation.

La fragmentation
C'est certainement l'un des plus grands problèmes que pose la structure de l'information sur Internet. En effet, beaucoup de documents sont fragmentés en plusieurs pages Web, sur un ou plusieurs sites. Comment prendre en compte cette fragmentation ? Comment reconstruire une pseudo structure linéaire permettant à l'utilisateur d'accéder à un document aussi bien dans son intégralité, que par fragment ? Aucun des systèmes actuels ne s'est posé ce problème, et il n'est pas rare qu'une recherche fournisse comme résultats les différentes pages d'un document, sans que l'utilisateur ne puisse avoir une vue synthétique lui permettant de savoir qu'il s'agit de la même source d'information.

Le multilinguisme
Composante importante d'Internet, le multilinguisme n'est pris en compte par aucun outil. C'est pourtant un point essentiel qui permettrait à toutes les langues de trouver leur place sur Internet, et la quasi omniprésence de la langue anglaise ne serait plus un problème pour la formulation de requêtes. Le problème des systèmes actuels provient tout simplement du fait qu'ils n'effectuent aucun traitement linguistique, ils ne traitent donc aucune langue en particulier, alors comment pourraient-ils traiter de manière spécifique plusieurs langues?

Le renouvellement continuel
C'est certainement l'une des caractéristiques les mieux intégrées par les moteurs de recherche. En effet, ils mettent à jour leur index très régulièrement pour la plupart et les renouvellements et mises à jour sont généralement pris en compte dans un délai se situant entre une semaine et un mois. Ce n'est peut-être pas encore tout à fait suffisant pour une activité de veille intense, mais le plus gros problème réside dans la présentation des résultats qui devrait permettre de disposer d'un aperçu des documents ayant subi des modifications. A l'heure actuelle, on peut imaginer une activité de veille qui consisterait à relancer régulièrement la même recherche. Pour connaître la liste des nouveaux documents, il faut alors comparer la liste des documents retournés par le système avec celle obtenue la fois précédente, et pour connaître la liste des documents mis à jour, il est nécessaire de consulter tous les documents afin de rechercher les modifications éventuelles.

Pourtant, un outil de veille utilisant des profils et des contextes utilisateurs, permettrait à la suite d'une recherche, de présenter de manière synthétique (en utilisant des codes de couleurs ou des pictogrammes) l'évolution des documents (nouveaux documents, documents modifiés, documents ayant expirés, ...).

Le caractère public et commercial
Lors d'une recherche sur Internet, seules les informations publiques sont disponibles. Ce problème n'est pas simple, et il revêt plus un caractère politique que technique. En effet, techniquement il ne serait pas des plus compliqués d'interfacer un système de recherche avec les grandes bases commerciales, mais encore faut-il trouver un accord sur les méthodes d'accès à ces dernières.

L'information non structurée
L'édition des documents sur Internet est totalement libre, et ne répond à aucune norme ou standard. De ce fait, il n'est pas aisé de retrouver certaines informations à priori importantes : Le nom de l'auteur du document, de l'organisation dont il dépend, de la date d'édition, de la langue du document, etc. Il va donc falloir (contrairement aux systèmes actuels qui ne se posent pas ce genre de questions) trouver des heuristiques permettant d'extraire de telles informations.

Nous constatons que les systèmes de recherche d'information sur Internet actuellement en service ne

répondent que très partiellement aux caractéristiques particulières de l'information disponible sur Internet. Il y a donc deux axes importants à retenir pour la mise en place de systèmes efficaces: Le traitement multilingue du langage naturel, et la prise en considération des spécificités de l'information sur Internet.

La situation paradoxale à laquelle nous sommes tous aujourd'hui confrontés et qui consiste à disposer d'une profusion et d'une abondance de matières premières sans posséder les outils permettant de l'exploiter et d'en tirer une valeur ajoutée ne doit plus exister. Et puisque comme le mentionne très justement Alvin TOFFLER, *"nous assistons à une fusion entre l'argent et l'information"*, un système performant permettrait à la fois de donner une valeur à l'information disponible sur Internet et de rendre ce dernier économiquement exploitable (temps de recherche plus courts, information plus pertinente, recherches plus précises, évaluation de l'évolution de l'information, etc.).

Description des critères d'évaluation des moteurs de recherche

DESCRIPTION des moteurs

Localisation
Localisation du moteur : Pays dans lequel est situé le serveur original (dans le cas où plusieurs serveurs existent) du moteur de recherche.
URL de la page principale
URL de la page d'accueil du moteur de recherche.
Remarque : Cette dernière correspond souvent à la page d'interrogation.
URL de la page d'interrogation
URL de la page disposant d'un formulaire permettant d'effectuer une recherche.
Editeur(s)
La liste des organisations/sociétés qui sont intervenues dans la mise en place du moteur de recherche. Cette liste contient le nom de l'éditeur, un lien vers cet éditeur (s'il possède un site Web), ainsi que son URL et le pays de son site.
Type de Service
Le type de service fourni par le moteur : Index Automatique, Index Manuel, Annuaire, Index Automatique et Annuaire, Index Manuel et Annuaire.
Type d'accès
L'accès à ce service est-il gratuit (Public) ou payant (Commercial), ou bien les deux types d'accès sont-ils proposés (Public et Commercial).
Fréquence de mise à jour
Evaluation moyenne de la fréquence de mise à jour de l'ensemble de l'index de la base.
Fréquentation moyenne
Evaluation moyenne du nombre de requêtes adressées au service pour une période donnée.
Sites Miroirs
Liste des sites miroirs du moteur de recherche. Les sites miroirs sont des répliques du site original à des localisations différentes, afin de répartir la charge des machines et de réduire les temps de communication. Cette liste comprend, s'il y a lieu, le nom du site miroir, un lien vers ce site, son URL ainsi que sa localisation.

COLLECTE des documents

Méthode de collecte
Décrit la manière dont les documents qui seront plus tard indexés sont déjà dans un premier temps collectés. Quatre cas sont possibles :

- Manuelle, des *net-surfer* passent leurs journées à parcourir le Web et à noter les adresses des sites intéressants.
- Automatique, un robot (petit programme) se promène sur le WWW et rapatrie les documents qu'il trouve en se déplaçant de lien en lien.
- Soumission d'URL, dans ce cas, ce sont les auteurs de pages et/ou sites Web qui envoient l'adresse de leurs créations aux moteurs afin que ces derniers indexent leurs pages.
- Suppression d'URL. Ce n'est pas un moyen de collecte de document, mais au contraire un moyen de suppression de document qui devrait logiquement être présent dans tout système permettant de soumettre des URLs (ce qui n'est d'ailleurs bien souvent pas le cas!). Dans le cas d'un moteur acceptant cette fonctionnalité, l'auteur ayant fourni l'URL de ses pages pour indexation a la possibilité de retirer ses dernières de l'index du moteur. Cette décision peut avoir plusieurs causes : Les pages ont été déplacées, ou bien elles n'existent plus, ...

Robot de collecte
Le nom (s'il en possède un) du logiciel effectuant la collecte automatique des documents.

Méthode de parcours
Méthode utilisée par le robot de collecte pour parcourir le Web. Deux possibilités : Largeur d'abord pour les programmes qui, à partir d'une page, parcourent d'un seul niveau tous les liens présents sur celle-ci, profondeur d'abord pour ceux qui à partir d'une page, explorent le premier lien, puis sur la page résultante parcourent à nouveau le premier lien, etc...

Protocole Standard d'Exclusion
Deux réponses possibles, oui ou non. Dans le cas de l'affirmative, cela signifie que le robot de collecte respecte le protocole standard d'exclusion permettant à tout WebMaster de spécifier des pages Web ne devant pas être collectées par le robot.

Serveurs Collectés
Ce critère décrit l'ensemble des types de serveurs collectés par le moteur de recherche. Nous avons restreint leur nombre aux plus essentiels : WWW, Usenet, F.T.P., Gopher, et une rubrique 'Autre', pour les outils collectant des documents à partir d'autres sources.

Couverture Géographique
Décrit la couverture géographique du robot de collecte des documents. En effet, de plus en plus d'outils sont spécifiques à un domaine géographique particulier (Europe, France, Pays Francophones, Suisse,...).

Type de Contenu
On trouve ici le sujet des documents collectés par le système. Si la plupart s'intéressent à tous les documents (dans ce cas, le type de contenu est étiqueté général), certains vont restreindre leur processus de collecte à certains sujets bien précis (médecine, brevets, informatique, ...)

Fréquence de Visite des Documents
Ce critère donne une évaluation de la fréquence moyenne de visite des documents par le robot de collecte. En effet, ce dernier doit parcourir le plus fréquemment possible les pages qu'il a déjà récupérées afin de tenir compte de toute modification du document. Ainsi, plus cette fréquence est élevée, et plus les résultats d'une recherche seront à jour par rapport à la réalité (si l'indexation est aussi fréquente).

INDEXATION des documents

Méthode d'Indexation
Nous distinguons ici deux méthodes d'indexation, l'indexation automatique et l'indexation manuelle.

Nom du Moteur
Le nom du moteur d'indexation dans le cas d'une indexation automatique.

Données Indexées
Les critères d'indexation peuvent être multiples et variés. Nous avons retenu :

- Le titre du document,
- ses différents sous-titres (balises <H1>...<Hn>),
- son en-tête (le <META> tag),
- sa date de création et/ou modification,
- sa taille,
- les URLs qu'il cite,
- le texte des URLs qu'il cite,
- d'autres balises éventuelles,
- un résumé du document,
- un extrait du document,
- et enfin le texte intégral du document.

Traitements Manuels
Nous regroupons ici quelques traitements manuels complémentaires à l'indexation et pouvant apporter une valeur ajoutée pour la recherche. Ainsi, nous prenons en compte le cataloguage des documents qui consiste à regrouper ces derniers par thèmes ou sujets, et la création de résumés manuels, permettant de mieux se rendre compte de la pertinence d'un documents lors de l'affichage des résultats qu'un simple extrait.

RECHERCHE des documents

Moteur de Recherche
Le nom du moteur de recherche utilisé.

Type de Recherche

Type de Question
Les systèmes de recherche peuvent proposer différents modes d'interrogation. Cela va de la requête en langage booléen (mots séparés par des opérateurs ET, OU et NON) à la question en langage naturel (formulation d'une question en langage libre) en passant par la requête comportant une liste de mots (qui revient souvent à une question booléenne dont tous les mots sont séparés par un opérateur implicite - généralement le OU) ou l'interrogation par expression régulière (expression définissant des ensembles de chaînes de caractères à rechercher).

Opérateurs Booléens
Nous distinguons ici,

- tout d'abord les opérateurs booléens classiques (ET, OU, NON),
  Remarque : Nous ne distinguons pas ici l'usage des opérateurs ET, OU NON, et des opérateurs "+" et "-" souvent utilisés par les moteurs de recherche. En effet, les deux sont plus ou moins interchangeables comme nous le soulignons dans la partie consacrée aux problèmes de la recherche.
- la combinaison des opérateurs, qui doit permettre dans une même requête booléenne de mêler les différents opérateurs disponibles afin d'effectuer des recherches plus élaborées que celle ne pouvant contenir qu'un type d'opérateur,
- le parenthésage des expressions donnant alors la possibilité de créer des requêtes très complexes comprenant plusieurs niveaux de parenthèses et plusieurs types d'opérateurs.
  Remarque : Il va de soit que cette option n'est logiquement présente que si le système permet de combiner les opérateurs.
- la possibilité de spécifier une certaine distance en terme de mots entre certains termes de la question (opérateur de proximité),
- enfin, le support de l'opérateur d'adjacence qui est en fait un cas particulier de l'opérateur de proximité. Il permet de spécifier que certains des mots de la recherche ne doivent être séparés par aucun autre terme. Nous pouvons trouver dans les systèmes de recherche actuels deux formes d'opérateurs d'adjacence. Tout d'abord sous la forme d'un opérateur du type <mot_1> ADJ <mot_2> qui signifie que l'on recherche des documents où <mot_1> et <mot_2> sont adjacents. Mais on trouve plus couramment ce que les moteurs de recherche appellent la recherche de phrase qui revient au même (mais qui est cependant moins souple dans le cas d'expressions parenthésées complexes). La syntaxe généralement utilisée est "<mot_1> <mot_2>" (phrase entre guillemets) pour recherche les documents dans lesquels les mots <mot_1> et <mot_2> sont adjacents.

La Troncature Automatique
Nous évaluons ici les possibilités de troncature automatique du système. La troncature automatique consiste à ne fournir qu'une partie du mot et à ce que le système recherche un ensemble de chaînes de caractères dérivées de cette sous chaîne. Ainsi, nous testons si le système supporte

- La troncature automatique gauche : on spécifie une chaîne de caractères, et le système recherche les documents comportant le mot spécifié ou les mots terminant par cette chaîne. Par exemple, pour la chaîne matique, le moteur de recherche va récupérer les documents contenant les termes informatique, télématique, mathématique, ...

- La troncature automatique droite : dans ce cas, on fournit au système la partie gauche d'un mot, et il recherche toutes le chaînes de caractères commençant par cette partie gauche. Par exemple, pour la chaîne inform, le moteur de recherche doit retourner les documents contenant les mots informatique, informaticien, informaticiens, informaticienne, informaticiennes, information, informations, informationnel, informationnels, informationnelle, informationnelles, ... informe, informes, informel, informels, informelle, informelles, informulé, informulés, informulée, informulées. Comme nous le voyons, c'est donc un bon moyen d'étendre une recherche au pluriel, féminin, ou mots de la même famille qu'un terme. Mais comme nous le constatons également, le bruit de la recherche croit énormément, et il est facile d'être rapidement submergé par des mots n'ayant aucun rapport avec la signification recherchée.

- La lemmatisation, qui consiste à rechercher pour un mot toutes ses formes possibles dans la langue (tous les genres, nombres, conjugaisons). Etant donné l'état de l'art dans ce domaine au niveau des moteurs de recherche sur Internet, nous avons étendu le terme de lemmatisation à tout traitement un peu plus évolué qu'une simple troncature droite.

- La possibilité pour l'utilisateur de désactiver les traitements de troncature automatique effectués par le système.

La Troncature Manuelle
Tout comme pour la troncature automatique, nous distinguons ici la troncature manuelle gauche et droite. Elles ont les mêmes effets que la troncature automatique, et généralement elle est mentionnée en plaçant le caractère '*' à l'endroit où doit se faire la troncature. Pour reprendre les exemples précédents de la troncature automatique, il faudra donc saisir *matique et inform*. Nous trouvons également la troncature interne, qui permet de spécifier le début et la fin d'un mot, en laissant une partie *libre*. Par exemple, poi*on donnera poison, poisson, poivron, ...

Champs de Recherche
Cette rubrique énumère les divers champs dans lesquels le système effectue la recherche. Nous avons dégagé cinq champs : l'URL du document, son titre, son résumé (généralement dans le cas d'une indexation manuelle. Ce critère regroupe aussi les moteurs recherchant dans une liste de mots-clés créée manuellement)., son texte intégral (l'ensemble du document moins son titre puisque ce critère est déjà pris en compte).

Champs Spécifiables
Ce critère énumère la possibilité offerte à l'utilisateur de spécifier dans sa requête dans quel(s) champs la recherche doit s'effectuer. Aux vues de ce que proposent certains moteurs de recherche, nous avons été très exhaustifs pour ce critère, et les champs spécifiables que nous avons retenus sont les suivants :

- L'URL du document,
- son titre,
- ses mots clés (META tag),
- son résumé,
- le texte intégral (pas l'URL ni les mots clés ou le résumé,...)
- les URL citées dans le document.

L'Elimination des Mots Vides
L'élimination des mots vides permet de ne pas prendre en compte certains mots (les mots vides!) trop communs et surtout n'apportant que peu de sens dans un texte tels que les articles. Nous distinguons deux moyens utilisés par les moteurs de recherche pour éliminer les mots vides. La méthode consistant à utiliser une liste de mots vides (Liste), et une autre à éliminer les mots dépassant un certain nombre d'occurrences dans la base. Bien sur, chaque moteur de recherche effectue des variations sur chacune de ces deux méthodes, et c'est pourquoi pour certains d'entre eux un commentaire complémentaire est disponible.

Prise en Compte de la Casse
Exprime comment le moteur de recherche réagit aux caractères majuscules et minuscules. Le tableau présente les chaînes recherchées pour une question en majuscule et une question en minuscule. Dans chacun des deux cas, nous envisageons que le système peut ne rechercher que les chaînes majuscules, ou majuscules, ou bien les deux à la fois.

Prise en Compte de l'Accentuation
Exprime comment le moteur de recherche réagit aux caractères accentués et non-accentués. Le tableau présente les chaînes recherchées pour une question accentuée et une question non-accentuée. Dans chacun des deux cas, nous envisageons que le système peut ne rechercher que les chaînes accentuées, ou non-accentuées, ou bien les deux à la fois.

Améliorations de la Recherche
Nous énumérons dans cette rubrique quelques moyens plus ou moins courants permettant d'effectuer une meilleure recherche. Nous distinguons l'utilisation d'un thésaurus qui va permettre de contrôler le vocabulaire, la lemmatisation des mots permettant de factoriser les mots identiques mais exprimés sous différentes formes (singulier/pluriel, masculin/féminin, verbe conjugué,...), la dérivation qui pour un mot lemmatiser va fournir toutes ses formes possibles, la reformulation qui doit permettre de gérer la synonymie et d'étendre les mots de la question à un ensemble de termes "proches" sémantiquement (la reformulation peut dans certains cas être multilingue), et enfin la recherche des documents similaires qui

à partir d'un ou plusieurs documents doit retrouver l'ensemble des documents traitant du ou des sujets identiques.

Présentation des RÉSULTATS

Informations Générales
Ce sont des informations de nature diverses relatives à la recherche : le nombre de documents-réponses trouvés par le moteur de recherche, la liste des termes de la question reconnus, de ceux non reconnus, et enfin la liste des mots vides de la question.

Organisation de Documents-Réponses
Elle doit permettre d'avoir une vue synthétique et efficace des documents-réponses. Pour cela, plusieurs critères essentiels :

- La méthode de tri des documents-réponses qui est propre à chaque système est décrite en quelques mots.
- La caractérisation des documents qui correspond à un regroupement des documents en sous-ensembles permettant d'avoir une bonne vue synthétique des résultats.
- L'élimination des liens dupliqués qui évite à l'utilisateur de voir apparaître plusieurs fois le même document dans la liste des documents-réponses.

Informations Concernant les Documents (par défaut)
La liste des informations affichées par le système concernant chaque document-réponse. Nous avons retenu :

- L'URL du document,
- le lien vers le document permettant d'aller le consulter directement (normalement toujours présent),
- le titre du document,
- la liste des mots-clés du document,
- un résumé du document,
- un extrait du document (différent du résumé dans le sens que l'extrait correspond souvent aux premières lignes du document),
- les URL citées dans le document,
- les URL citant le document,
- la taille du document,
- sa date de dernière mise à jour connue (par l'auteur),
- sa date de dernière visite (par le système),
- une mesure de pertinence (score),
- et enfin la mise en évidence des mots de la question présents dans le document.

Informations Concernant les Documents (option)
Nous retrouvons ici les mêmes critères que dans la rubrique précédente. La différence est que nous mettons ici en évidence les informations qui peuvent être affichées par le biais d'une option proposée par le système de recherche.

Bibliographie

[ALIS96] - Internet Society et Alis Technologies. (1996).
'La série de normes ISO 8859', [En ligne].
URL: http://babel.alis.com:8080/codage/iso8859/jeuxiso.htm
[AYMO96] - AYMONIN David. (1996).
'Liste des fiches techniques concernant les outils de recherche d'information sur Internet', [En ligne]. URFIST Alsace Lorraine Franche-Comté.
URL: http://www-scd-ulp.u-strasbg.fr/urfist/Fiches_Techniques/fiches.htm
Mise à jour permanente En une ou deux pages un résumé complet des astuces et techniques d'utilisation des moteurs, des répertoires thématiques, et de bien d'autres outils et services permettant de mener des recherches d'information sur le Web.
[BARL96] - Barlow Linda. (Novembre 1996).
'The Spider's Apprentice: how to use Web search engines' , [En ligne].
URL: http://www.monash.com/spidap.html
[BARR96] - Tony Barry et Joanna Richardson. (Juillet 1996).
'Indexing the Net - A Review of Indexing Tools' , [En ligne].
URL: http://bond.edu.au/Bond/Library/People/jpr/ausweb96/
Ce document évoque les outils d'indexation traditionnels d'Internet (WAIS, archie, veronica, ...), et nous fournit quelques tableaux synthétisant les résultats de différentes évaluations effectuées à travers le monde.
[BIRM96] - Birmingham Judy. (Novembre 1996).
'Selected Internet Search Engines', [En ligne].
URL: http://www.stark.k12.oh.us/Docs/search/
Différents tableaux récapitulatifs des différentes fonctionnalités supportées par les principaux outils de recherche.
[BOCH96] - Bocher Bob. (Octobre 1996).
'A Higher Signal - To - Noise Ratio: Effective Use Of Web Search Engines', [En ligne]. WETC: Wisconsin Educational Technology Conference, Green Bay, Wisconsin, USA.
URL: http://www.state.wi.us/agencies/dpi/www/search.html
[BORT96] - Bortzmeyer Stéphane. (Avril 1996).
'Systèmes de recherche par mots-clés' , [En ligne].
URL: http://www.freenix.fr/Web/recherche.html
Un bref guide utilisateur des fonctions de recherche généralement utilisées, ainsi qu'un classement par type de service des différents outils de recherche.
[BRAY96] - Bray Tim. (1996). *'Measuring the Web'*, [En ligne].
Computer Networks and ISDN Systems, Volume 28, issues 7–11, p. 993.
URL: http://www-di2.cea.fr/www5/www134/overview.htm
De très nombreuses statistiques sur le WWW: Les types de documents disponibles, la taille des documents, l'utilisation des balises, ...
[CARD96] - Card Stuart K.. (Mars 1996).
'Visualizing Retrieved Information: A Survey' , [En ligne]. Special Report - Computer Graphics and Visualization in the Global Information Infrastructure, CG&A, Vol. 16, No. 2.
URL: http://www.computer.org/pubs/cg&a/report/g20063.htm
[CHAR96] - Chartron Ghislaine. (Octobre 1996).
'Recherche d'Information sur Internet' . Cours INRIA « La recherche d'Information sur les réseaux », ADBS Editions.
Une très bonne synthèse sur la problèmatique de la recherche d'information sur Internet, et sur les différents types d'outils de recherche.
[CONT96] - Conte Ron Jr.. (1996).
'Guiding Lights', [En ligne]. Internet World, Vol. 7 No. 5.
URL: http://www.iw.com/1996/05/guiding.html
[GRAY96] - Gray Terry A.. (Juillet 1996).
'How to Search the Web: A Guide To Search Tools' , [En ligne].
URL: http://issfw.palomar.edu/Library/TGSEARCH.HTM
Une description des fonctionnalités des moteurs de recherche les plus courants.
[KOCH96] - Koch T.. (Septembre 1996).
'A review of robot based internet search services' , [En ligne].
Lund University Library, NetLab.
URL: http://www.ub2.lu.se/desire/radar/reports/D3.11/
Etude préparatoire au projet DESIRE, ce document très détaillé sur la problématique de la recherche d'information sur Internet dresse un état de l'art relativement complet.
[KOST96] - Koster Martijn. (1996).
'The Web Robots Pages', [En ligne].
URL: http://info.webcrawler.com/mak/projects/robots/robots.html
Un document complet consacré aux robots de collecte. Une liste des robots existants, leur description, le protocole d'exclusion, ...
[LAGE96] - Lager Mark. (1996).
'Spinning a Web Search : Trends in Information Retrieval' , [En ligne].
California Lutheran University.
URL: http://www.library.ucsb.edu/untangle/lager.html
Quelques explications sur les principes de fonctionnement des systèmes de recherche d'information, une description sommaire des principaux outils de recherche sur Internet et enfin une bibliographie abondante.
[LANT96] - Lanteigne Diane. (Avril 1996).
'Recherche de ressources dans Internet' , [En ligne].
Cemagref de Grenoble.

URL: http://www.grenoble.cemagref.fr/navigator/recherche.html
Une revue très complète, classée par types d'outils des différents
systèmes de recherche d'information sur Internet.
[LEON?] - Andrew J. Leonard. (sans date).
'reviews - where to find anything on the Net', [En ligne]. CNET.
URL: http://www.cnet.com/Content/Reviews/Compare/Search/index.html
Une description des différents outils de recherche disponibles sur Internet,
quelques comparaisons sommaires.
[LIU96] - Liu Jian. (Septembre 1996).
'Understanding WWW Search Tools ', [En ligne]. Reference Department, IUB Libraries.
URL: http://www.indiana.edu/~librcsd/search/
Une brève description des principaux outils de recherche, méta-moteurs,
et moteurs spécialisés, ainsi que quelques liens vers des pages recensant
les différents outils de recherche disponibles sur Internet.
[LUND96] - S. Lundberg et al.. (Août 1996).
'The European Web Index: An Internet Search Service for the European Higher Education, Research and Development Communities', [En ligne].
NetLab, Lund University Library.
URL: http://www.ub.lu.se/desire/radar/reports/D3.12/
[MELA96] - Mélard Anne-Sophie. (Octobre 1996).
'La recherche documentaire sur l'Internet', [En ligne]. URFIST.
URL: http://www-scd-ulp.u-strasbg.fr/urfist/Anne_Sophie/rechdoc.htm
Une synthèse aussi complète que possible qui présente un répertoire commenté des cours, guides et tutoriaux existant sur l'Internet et ailleurs pour apprendre à chercher de l'information ou simplement à utiliser Internet.
[NIC96] - NIC France (Network Information Center). (1996).
'Comptage du nombre de machines, domaines et réseaux dans le monde', [En ligne]. NIC France.
URL: http://www.nic.fr/Statistiques/world/
[NORT96] - Northern Webs. (1996).
'Search Engine Tutorial for Web Designers', [En ligne].
URL: http://www.digital-cafe.com/~webmaster/set01.html
[NORT97] - Northwestern University Library. (Janvier 1997).
'Search Tools', [En ligne].
URL: http://www.library.nwu.edu/resources/internet/search/
Une liste d'outils de recherche classée par type, ainsi qu'une description des *"meilleurs systèmes"*.
[NOTE96] - Notess Greg R.. (Octobre 1996).
'Comparing Internet Search Engines and Finding Aids', [En ligne].
URL: http://www.imt.net/~notess/compeng.html
Un document intéressant qui fournit une revue détaillée des principaux
moteurs de recherche, ainsi que des tableaux comparatifs.
[OVER96] - Overton Richard. (Septembre 1996).
'Search Engines Get Faster and Faster, But Not Always Better',
[En ligne]. PC World.
URL: http://www.pcworld.com/workstyles/online/articles/sep96/1409_
engine.html
[PAGE96] - Page Adam. (Octobre 1996).
'The Search Is Over', [En ligne]. PC Computing.
URL: http://www.zdnet.com/pccomp/features/fea1096/sub2.html
[PLOU96] - Plourde Jean-Noël. (1996).
'Définition et application de critères d'évaluation d'outils de recherche dans Internet', [En ligne]. Cursus (périodique électronique étudiant de l'École de bibliothéconomie et des sciences de l'information). EBSI de l'Université de Montréal.
URL: http://mistral.ere.umontreal.ca/~beaudryg/cursus/vol1no2/plourde.html
Un article exhaustif sur le fonctionnement des différents moyens de recherche d'information sur Internet, suivi d'une description des principaux moteurs de recherche.
[PROS96-1] - Prosise Jeff. (Juin 1996).
'Researching with the Web', [En ligne]. PC Magazine, 15(11), p235-236.
URL: http://www.pcmag.com/issues/1511/pcmg0081.htm
[PROS96-2] - Prosise Jeff. (Juillet 1996).
'Crawling the Web', [En ligne]. PC Magazine, 15(13), p277-278.
URL: http://www.pcmag.com/issues/1513/pcmg0045.htm
[RAND95] - Randall Neil. (1995).
The Search Engine That Could', [En ligne]. Ziff-Davis Publishing Company.
URL: http://www.zdnet.com/pccomp/features/internet/search/
[SALT74] - Gerard Salton, A. Wong et C. S. Yang. (Juillet 1974).
'A Vector Space Model for Automatic Indexing', [En ligne]. Department of Computer Science, Cornell University, Ithaca, N.Y.
URL: http://cs-tr.cs.cornell.edu/Dienst/UI/2.0/Describe/ncstrl.cornell/TR74-218
[SLOT96] - Slot Matt. (1996).
'The Matrix of Internet Catalogs and Search Engines', [En ligne].
URL: http://www.ambrosiasw.com/~fprefect/matrix/
Un site exhaustif et bien structuré évaluant de nombreux outils de recherche d'information sur Internet.
[SMIT96] - Smith Richard J.. (1996).
'Web Search Cheat Sheet', [En ligne].
URL: http://www.colosys.net/search/
SULL96] - Sullivan Danny. (Décembre 1996).
'A Webmaster's Guide to Search Engines and Directories', [En ligne].
Calafia Consulting.
URL: http://calafia.com/webmasters/
Un guide relativement complet, allant des conseils pour que votre page apparaisse dans les premières réponses d'une recherche, jusqu'à l'étude détaillée de différents critères tels que la fréquence de mise à jour de l'index des moteurs, en passant par une description des différentes fonctionnalités disponibles sur chaque système. De plus, ces informations sont remises à jour très régulièrement.

[SUNS96] - SunSITE Manager. (Septembre 1996).
'Internet Search Tool Details', [En ligne]. Berkeley Digital Library SunSITE.
URL: http://sunsite.berkeley.edu/Help/searchdetails.html
Description des fonctionnalités essentielles des principaux outils de recherche : Alta Vista, Excite, HotBot, Infoseek Ultra, Lycos, Open Text Web Index.
[TILL96] - Tillman Hope N.. (Février 1996).
'Evaluating Quality on the Net', [En ligne]. Computers in Libraries Preconference, Crystal City, Arlington, Virginia, USA.
URL: http://www.tiac.net/users/hope/findqual.html
[TRAU97] - Traugott Koch. (1997).
'Searching the Web - Systematic overview over indexes' , [En ligne]. Lund University Electronic Library .
URL: http://www.ub2.lu.se/tk/websearch_systemat.html
Une revue exhaustive très complète d'un grand nombre de système de recherche.
[TYNE96] - Tyner Ross. (Mai 1996).
'Sink or Swim: Internet Search Tools & Techniques' , [En ligne]. Online workshop held at Connections '96 Conference, Vancouver, B.C., Canada.
URL: http://www.sci.ouc.bc.ca/libr/connect96/search.htm
Un guide sur les méthodes de recherche (catalogues, moteurs de recherche), et une description de plusieurs outils de recherche.
[URFI96] URFIST Alsace Lorraine Franche-Comté. (Novembre 1996).
'Chercher l'information sur Internet : Le défi' , [En ligne].
URL: http://www-scd-ulp.u-strasbg.fr/urfist/Recherche_information/
sommair.htm
Un bref document qui rappelle le principe d'utilisation des principaux types d'outils de recherche. Avec des liens et des exemples de recherches types.
[VEND96] - Venditto Gus. (1996).
'Search Engine Showdown: IW Labs tests seven Internet search tools.' ,
[En ligne].
Internet World, Vol. 7 No. 5.
URL: http://www.iworld.com/plweb-cgi/idoc.pl?575+unix+_free_user_+
pubs.iworld.com..80+Publications+iWORLD+Internet_World+Internet_World+
+search%26engine%26showdown
Des informations générales sur le fonctionnement des moteurs de recherche, ainsi qu'une évaluation (peu formelle) de ceux-ci.
[WEBS96] - Kathleen Webster et Kathryn Paul. (Janvier 1996).
'Beyond Surfing: Tools and Techniques for Searching the Web' , [En ligne].
Feliciter 42.1, 48-54.
URL: http://magi.com/~mmelick/it96jan.htm
Des explications sur le fonctionnement des différents types d'outils de recherche, ainsi qu'une description de ces derniers.
[WEST96] - Westera Gillian. (1996).
'Robot-Driven Search Engine Evaluation Overview' , [En ligne].
URL: http://www.curtin.edu.au/curtin/library/staffpages/gwpersonal/
senginestudy/index.htm
Un guide très complet avec de nombreuses explications, et comparaison des outils de recherche.
[WINS95] - Winship Ian R.. (1995).
'World Wide Web searching tools - an evaluation' , [En ligne]. Information Services Department, University of Northumbria at Newcastle, UK.
URL: http://www.bubl.bath.ac.uk/BUBL/IWinship.html
Cette évaluation relativement vieille porte sur six sites de recherche (WWWorm, WebCrawler, Lycos, Harvest, Galaxy, Yahoo). Les critères sont clairement définis et des tableaux comparatifs permettent de bien identifier les similarités et les différences entre les sites.
[WOOD96] - Woodruff Allison et al.. (1996).
'An Investigation of Documents from the World Wide Web' , [En ligne]. Computer Networks and ISDN Systems, Volume 28, issues 7–11, p. 963.
URL: http://www-di2.cea.fr/www5/www273/overview.htm
De très nombreuses statistiques sur le WWW: Les types de documents disponibles, la taille des documents, l'utilisation des balises, ...
Erreu de 'Web Search Engine Comparison', [En ligne].
URL: http://sawfish.lib.utexas.edu/Staff/ICO/training/topics96/
webcomparison.html

Vous pouvez trouver d'autres bibliographies intéressantes aux adresses:

- http://www.ub2.lu.se/desire/radar/lit-about-search-services.html
- http://www.hamline.edu/library/bush/handouts/comparisons.html
- http://www3.sympatico.ca/jn.plourde/moteurs.htm
- http://www-scd-ulp.u-strasbg.fr/urfist/Anne_Sophie/classe1.htm
- **http://www-scd-ulp.u-strasbg.fr/urfist/Anne_Sophie/classe5.htm**