La méthodologie et la stratégie de recherche d'information à valeur-ajoutée sur Internet

Carine Dou Goarin, Eric Giraud, Bruno Mannina, Luc Quoniam

La méthodologie et la stratégie de recherche d'information à valeur-ajoutée sur Internet

Dou Carine [1,2], Mannina Bruno [1], Giraud Eric [1], Quoniam Luc[1]

[1] CRRM
Centre de Recherche Rétrospective de Marseille / Université Aix-Marseille III
13397 Marseille Cedex 20
Tel : 04-91-28-87-40, Fax : 04-91-28-87-12

[2] Conseil Régional Provence-Alpes-Côte d'Azur
BP 67, 13441 Marseille Cantini Cedex 06

Résumé :

La croissance exponentielle que connaît Internet impose aux spécialistes de l'information une méthodologie dans leur collecte d'informations.
En effet, outre le World Wide Web que tous les utilisateurs connaissent, Internet recèle d'autres sources, à savoir les listes de diffusion, les groupes de discussion... Ces sources sont également extrêmement riches en informations de toutes sortes (experts du domaine, rapports de recherche, technologie-clé...).

Dans le langage documentaire, ces sources correspondent à de l'information informelle qui a dans la plupart des cas, une valeur-ajoutée indispensable dans la prise de décision.

Tout ceci nécessite donc une bonne stratégie de recherche.

Mots-clés :

Internet, Moteurs de recherche, Index, Liste de diffusion, News, Agent Intelligent, Méthodologie de recherche d'information

La méthodologie et la stratégie de recherche d'information à valeur-ajoutée sur Internet

La croissance exponentielle que connaît Internet impose aux spécialistes de l'information une méthodologie dans leurs collectes d'informations.
En effet, outre le World Wide Web que tous les utilisateurs connaissent, Internet recèle d'autres sources, à savoir les listes de diffusion, les groupes de discussion...

Les informations présentes sur Internet ont des caractéristiques bien spécifiques, qu'il est nécessaire d'intégrer dans sa méthodologie de recherche sur le Web.
En effet, l'information sur Internet est hétérogène, aussi bien en ce qui concerne son contenu, son support, sa langue, que son accessibilité (payant, abonnement ou gratuit).
De plus, elle est dynamique et en continuel renouvellement.

Capter l'information à valeur-ajoutée sur Internet nécessite une certaine maîtrise des techniques de recherche d'information, ainsi que des outils de collecte et d'analyse.

Mais avant toutes choses, il est important de bien comprendre que la démarche de la recherche d'information va fortement dépendre de la question posée. De plus, au vue de la mouvance du réseau et de son contenu, une recherche d'information sur Internet n'est jamais terminée…

Section 1 : Les différents types d'outils pour la recherche d'information sur Internet

1. Les moteurs de recherche

1.1 Les moteurs de recherche (ou robots automatiques d'indexation) liés aux pages HTML

Les moteurs de recherche sont des serveurs spécialisés dans la localisation de ressources sur Internet. Ces serveurs indexent et stockent les informations sur des machines spécifiques dans des bases de données.
Ceux sont les outils les plus utilisés sur Internet.

1.1.1 Le but

Leur but est de rechercher des informations à partir de mots-clés par rapport à une stratégie de recherche. Une requête d'information est saisie par l'utilisateur dans un formulaire HTML. Il peut contenir plusieurs mots-clés combinés avec des opérateurs de recherche.

L'utilisation d'un tel outil permet une première approche de l'information recherchée. Ce survol est nécessaire pour élargir ou affiner sa recherche, voire pour la rediriger.
De plus, cette méthode peut aboutir à la détection de gateways et de sites pertinents.
Un **gateway** est une page HTML contenant une liste de liens hypertextes (pointant sur une multitude de serveurs) se rapportant à une thématique précise.
Les **sites pertinents** serviront à récupérer l'information utile (textes ou fichiers) ainsi que l'identification des experts dans le domaine. Cette localisation d'experts doit être complétée par une recherche plus exhaustive sur des sites spécifiques (cf. moteur de recherche spécifique).

1.1.2 Le fonctionnement

Ils fonctionnent de la manière suivante : un moteur de recherche reçoit une requête de l'utilisateur, puis interroge sa base de données suivant les mots-clés et affiche une liste de liens hypertextes vers les pages Web où ces mots apparaissent le plus souvent, avec, éventuellement, l'affichage de quelques lignes de texte pour chaque page.

Les performances d'un tel système de recherche d'information, résultent de la combinaison de deux types de technologies distinctes. Dans un premier temps, le robot parcourt Internet, explore les serveurs Web. Il navigue au travers des liens hypertexte pour récolter soit de nouvelles pages, soit des pages ayant été mises à jour.
Ensuite, dans un second temps, un moteur d'indexation va intégrer les informations recueillies par le moteur de recherche dans sa base de données. Cette information sera structurée de manière à en faciliter l'accès.
La structure de la base de données est du type : Titre de la page, adresse WWW, ligne de texte, langue, mise à jour…

Il est important de savoir que les méta-données[1] fournis par l'auteur des pages WWW pour décrire le contenu de ses pages ont une pondération grandissante dans l'indexation des pages HTML.

1.1.3 Stratégie d'interrogation

En général, les moteurs de recherche offre à l'utilisateur deux modes de recherche :
- Un mode simple où il n'est pas nécessaire de connaître le langage d'interrogation. L'utilisateur donne simplement une liste de mots, et le moteur lui renvoie une liste de liens hypertextes correspondant aux pages contenant le maximum de ces mots.
- Un mode avancé ou évolué où l'utilisateur aura la possibilité de combiner des mots-clés à l'aide d'opérateurs booléens et de paramètres spécifiques aux différents moteurs de recherche (troncature, parenthèses…).

[1] Les codes Méta permettent de définir des paramètres dans les pages Web. Ces codes indiquent avec précisions aux moteurs de recherche les informations comme la description du contenu des pages, l'auteur du site, les mots-clés…

Les paramétrages sont spécifiques à chaque moteur. Les principales règles de base sont les suivantes :
- Un mot est une chaîne de caractères alphanumérique délimitée par un caractère de ponctuation, un blanc ou caractère spécial. (crrm.univ-mrs.fr : 4 mots, C.E.E. : 3 mots)
- Mot composé est une séquence de mots contigus, séparés par un blanc ou autre délimiteur. Lors de la requête, entourer le mot composé de guillemets
Ex : "veille technologique"
- Les majuscules et minuscules sont différenciées dans l'index. Mettre un mot en minuscule permet de retrouver toutes les casses.
''Veille'' ne retrouvera pas VEILLE ou veille.
''veille'' retrouvera VEILLE ou Veille ou veille.
- L'accentuation suit la même logique.
''économique'' ne retrouvera pas economique
''economique'' retrouvera ''économique'' et ''economique''
- Les opérateurs +,-,* disponibles dans les 2 modes d'interrogation (simple ou avancé)
+ présence obligatoire
- exclusion obligatoire
* opérateur de troncature

Par défaut, le moteur recherche l'information dans tous les champs, mais il est possible d'affiner cette recherche sur des champs liés à la structure HTML: host, link, domain, title, text,...
syntaxe:
nom du champ en minuscule: mot ou expression

host:veille
text:''intelligence economique''
link:univ-mrs.fr
domain:fr
from:dupont

Figure 1 : Exemple de champs pour interroger les moteurs de recherche

1.1.4 Exemple

ALTAVISTA (http://www.altavista.com)
Altavista est un moteur de recherche développé par Digital Equipment, disponible depuis décembre 95. Il indexe le Web et les news quotidiennement. Son taux d'indexation est de 28 % du Web, il fait partie de ceux qui indexent le plus de pages HTML.

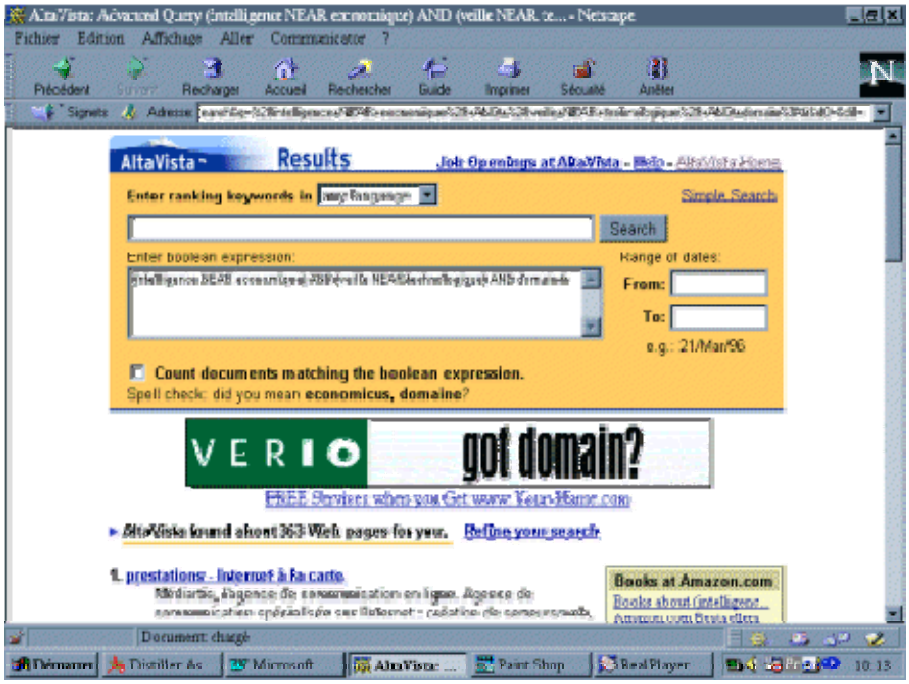

Figure 2 : Exemple de requête sur Altavista

La figure 2 représente une requête sur Altavista correspondant à :
(intelligence NEAR economique) AND (veille NEAR technologique) AND domain:fr

Dans ce cas, Altavista renvoie une liste de liens hypertextes contenant les mots intelligence économique (ou economique) et veille technologique dont les serveurs font partie uniquement du domaine FR (France)
Cette requête n'a pas de contrainte sur les dates ainsi que sur la langue. Il sera donc possible de trouver des pages dans toutes les langues, et de toutes dates. La spécificité domain:fr n'implique en rien une contrainte sur la langue : un serveur du domaine fr peut posséder des pages HTML dans n'importe quelle langue.

1.1.5 Avantages et inconvénients

Les avantages de l'utilisation des moteurs de recherche sont nombreux.
Ces moteurs sont d'abord très simple d'utilisation.
Les moteurs de recherche sont des outils ayant la couverture la plus exhaustive de l'espace public d'Internet. De plus, leur mise à jour est très rapide, puisqu'ils parcourent sans relâche le WWW à la recherche de nouvelles pages ou de pages modifiées.
Ils sont en constante amélioration avec l'intégration de nouvelles technologies.

En ce qui concerne, les inconvénients de ce type d'outil, les informations trouvées sont souvent du domaine du "bruit" c'est-à-dire que ceux sont des liens hypertextes qui ne répondent pas réellement à la requête, mais qui ont été indexés en tant que tel.
Le contenu des pages étant libre, et non vérifié, un individu peut faire indexer ses pages avec des mots-clés (dans les méta balises sur ces pages HTML) qui ne correspondent pas à son contenu.

1.2 Les Index

Les Index ou Sites Répertoires sont des listes de serveurs classés par thèmes. Ces classements peuvent être de n'importe quel ordre : thème, pays…
Cet outil permet de rechercher un ou plusieurs services Internet contenant des informations sur un thème donné.
Les Virtuals Libraries font partie de cette catégorie. Les virtuals libraries sont des sortes de gateways sur des thèmes précis et répertoriés, dans lesquelles l'internaute trouvera tous les liens hypertextes correspondant à son thème de recherche. Ces Librairies sont répertoriées[2] et leurs concepteurs sont tenus de les tenir à jour. C'est un service gratuit sur Internet.

1.2.1 But et fonctionnement

La recherche sur ce type d'outils permet, comme pour les moteurs de recherche de cerner les sites gateways, ainsi que les sites pertinents. Les index sont donc de très bons outils de première approche de l'information sur le Web.

Le principe d'utilisation des index est simple, puisque l'utilisateur n'aura qu'à naviguer dans l'arborescence des thèmes et sous-thèmes pour trouver la catégorie qui correspond à ses attentes.
Dans le cas où l'utilisateur ne trouve pas la catégorie recherchée, il a la possibilité de faire une recherche par mots-clés pour trouver les différents thèmes se rapportant à sa recherche.
Dans ce cas, l'utilisateur récupère les différentes catégories qui comportent ses mots-clés ou un de ses mots-clés suivi de la liste des sites correspondants.
Le problème des index réside dans le fait qu'un webmaster[3] indexe son site dans autant de sous-thèmes qu'ils le désirent. Donc en recherchant l'information par mots-clés, l'utilisateur obtiendra de l'information redondante s'il interroge l'index par mots-clés (non pas en ce qui concerne les rubriques, mais les sites proprement dits).

1.2.2 Exemple

L'index Yahoo (Yet Another Hierarchically Organized Oracle) a été crée par deux étudiants de Standford : David FILO et Jerry Yank. Aujourd'hui c'est une société commerciale.
Yahoo est devenu le plus populaire des moteurs d'indexation. Il permet de faire des recherches dans la plupart des langues par grands thèmes ou directement par sujet.
Yahoo est très convivial et instinctif, son taux de réponse est très satisfaisant.

---

[2] Liste des Virtuals Libraries : http://vlib.org/Overview.html

[3] personne qui conçoit et réalise un site WWW

Yahoo possède son propre moteur de recherche. L'index sur lequel se fait l'interrogation comporte cinq champs : URLs, titres des objets, commentaires des administrateurs, titres des rubriques et news.
S'il ne trouve pas de réponses, Yahoo fait un lien direct vers Altavista.

L'accès de Yahoo est classificatoire : 14 rubriques majeures thématiques avec une profondeur de 4 niveaux maximum pour chaque rubrique. Le nombre de documents associés à une rubrique est indiqué entre parenthèses. Le signe @ indique une rubrique appartenant aussi à une autre branche de la classification. La position dans la classification est toujours indiquée en haut de la page.

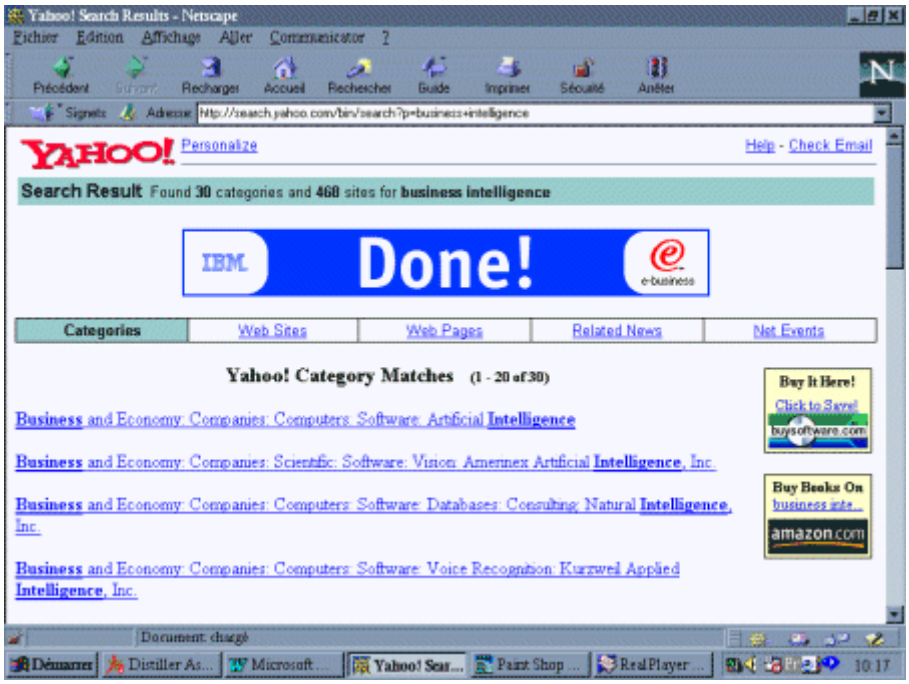

Figure 3 : Exemple d'Index : YAHOO (http://www.yahoo.com)

Dans cet exemple, suite à une interrogation simple avec les mots *business* et *intelligence*, yahoo.com (index anglais) a trouvé 30 catégories et 469 sites.
Maintenant, il est possible de cerner au plus juste sa recherche avec les catégories que Yahoo propose ou simplement de visualiser tous les sites qui ont les mots business et intelligence dans leur présentation.

1.2.3 Avantages et inconvénients des index

Les index apporte une valeur ajoutée grâce à la validation et à la catégorisation des liens hypertextes.
De plus, le bruit est limité.
La recherche se fait sur les pages indexés par le Webmaster (dans la plupart des cas ceux sont les pages d'accueil) donc l'utilisateur ne trouvera pas plusieurs fois le même sites dans ses réponses.

En plus du contenu assez exhaustif des index, les recherches sont exploratoires et intuitives.

En ce qui concerne les inconvénients, la mise à jour des sites est très aléatoire. En effet, la mise à jour régulière et datée n'est pas garantie. Pour ce qui est de l'indexation des nouveaux sites, il suffit de choisir les catégories et l'indexation est assurée dans les 48 heures qui suivent.

Par rapport aux moteurs de recherche, les informations seront moins exhaustives avec les index car ils indexent seulement la description du site.
En général, la description ne doit pas comporter plus de 25 mots et sachant que le contenu des pages n'est pas indexé, la recherche d'information risque de comporter des "silences". Le Silence est un terme utilisé par les documentalistes pour désigner les réponses qui correspondent à la recherche, mais qui ne sont pas retrouvées.

Devant les nombreux avantages qu'offrent les deux méthodes de recherche (moteur de recherche et index), les prestataires de ce genre de services se mettent de plus en plus, à offrir une interface commune à ces deux outils.
Récemment, Altavista a proposé à ses utilisateurs une recherche par catégories.

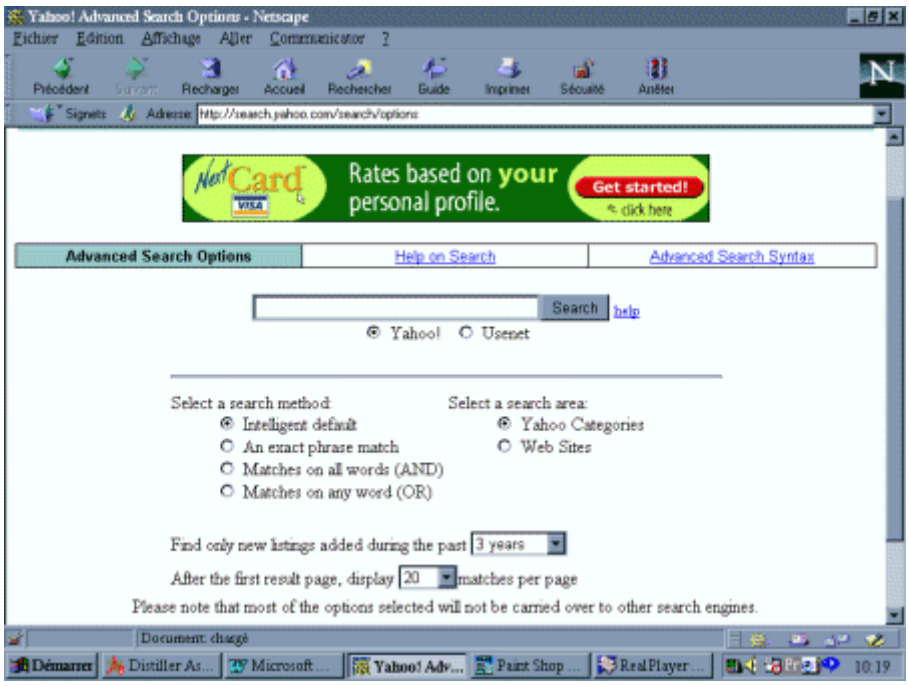

Figure 4 : Le moteur de recherche de Yahoo

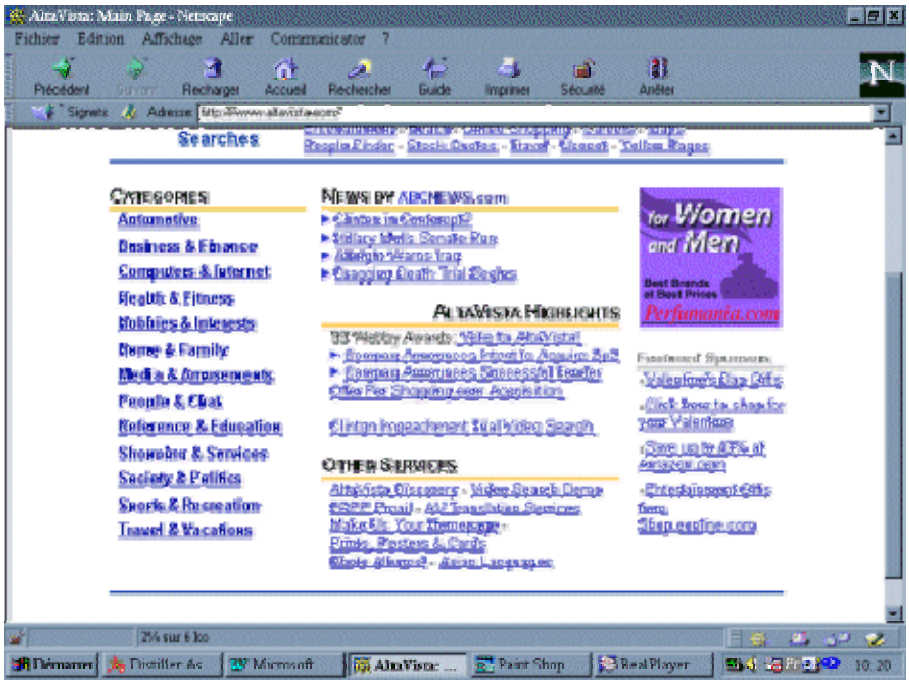

Figure 5 : L'index d'Altavista

1.3 Les moteurs de recherche spécifiques

Les moteurs de recherche spécifiques sont des outils permettant de rechercher de l'information concernant des types de ressources autres que les pages WWW de l'Internet.
Il est donc possible de rechercher de l'information sur les news, sur les ftp (fichiers), ou sur les annuaires.

1.3.1 Les News (ou Forums de discussion)

Il existe des moteurs de recherche permettant de cibler les news dans lesquelles l'internaute pourra trouver les différents messages qui sont déjà parvenues aux forums.
Il pourra ainsi avoir des renseignements sur les personnes travaillant dans des domaines précis, ainsi que les réponses à ses questions et le titre de la news groupe de son domaine de recherche.

Pour ce type d'outil, il est possible de rechercher de l'information soit sur des mots-clés, soit sur des titres de news, soit sur des E-mails, soit sur le sujet.
Les réponses donneront la date, le sujet de la news, ainsi que le nom de son auteur.
Dans le cas de l'exemple suivant, l'information rapatriée concerne les internautes intéressés par l'"intelligence business" dans le forum "misc.industry.quality". Il y a donc 7 personnes qui ont discuté sur ce sujet dans ce forum.

Cette recherche permet de connaître l'auteur, son mail, la date du mail, le sujet, le contenu, l'organisation qui gère le forum, ainsi que tout ce qu'il a écrit et dans quel autre forum.
Dans les réponses figure 6, il y a deux messages postées à un autre forum, alors qu'il est bien spécifié dans la gestion. C'est pour la simple raison que la personne qui a posté son message, l'a fait dans plusieurs forums à la fois, et le forum sur lequel la recherche est effectué est en "Carbon Copy", c'est-à-dire qu'il y a plusieurs forums destinataires.

Cette stratégie de recherche permet donc de cerner les experts, leur préoccupations, ainsi que l'émergence de certaines technologies (dans le cas où les messages postés sont exempts de toutes méfiances).

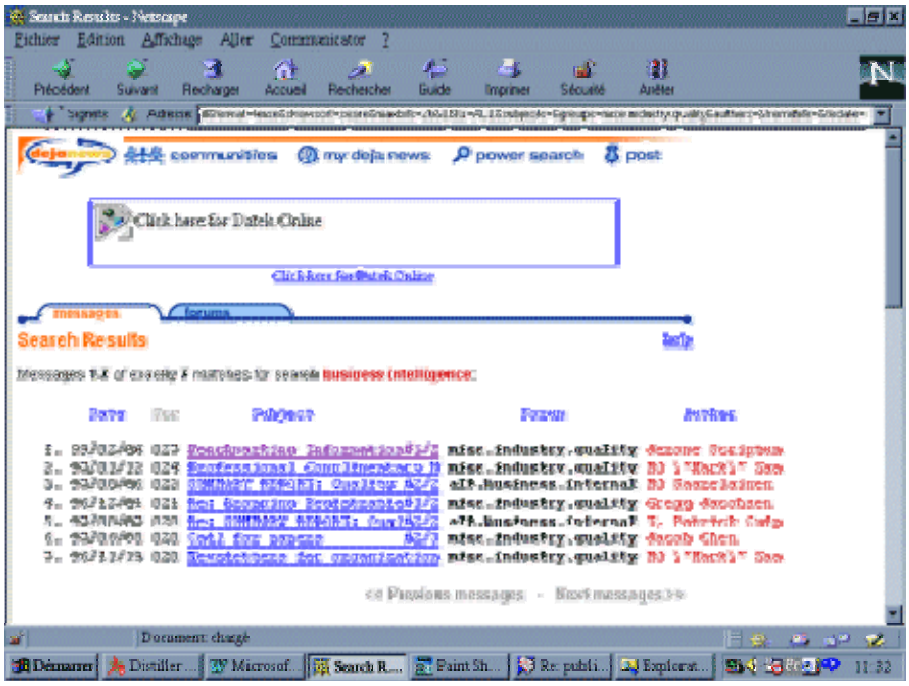

Figure 6 : Exemple de recherche sur dejanews (http://www.dejanews.com)

Il est aussi possible de connaître les préoccupations des employés d'une société via les forums. Pour cela, il faut simplement inscrire, par exemple, "*@ibm.com" dans le champ auteur (Author), et les messages postés par les employés d'IBM apparaîtront.
Cette méthode permet aussi de savoir dans quelles news, la personne a communiqué, donc quels sont ses centres d'intérêt.

Les avantages de ce type d'outils sont assez explicites, au vue de ces quelques exemples. Il est possible de connaître les acteurs et les experts des domaines, ainsi que d'essayer de cerner l'émergence de nouveaux concepts.
De plus, si l'utilisateur considère que toutes les questions ont dores et déjà été posées, il sera aisé de trouver les réponses à ses propres questions.

L'inconvénient majeur de ce type d'outil réside non pas dans la collecte de l'information sur un concurrent par exemple, mais sur l'utilisation que peuvent faire les employés d'une société dans ces news. Par exemple, il est possible de surveiller dans les news d'offre d'emplois, les offres des concurrents et donc de percevoir quels vont être leur stratégie de développement et les nouvelles technologies qu'ils vont utiliser.
Il est donc important d'avoir une stratégie bien définie dans la manière d'utiliser les e-mails. Le repérage des mails, et donc les préoccupations des internautes est facilité grâce à ce type d'outil.

1.3.2 Les Fichiers

Il existe des moteurs de recherche permettant de trouver des fichiers. Ces outils fonctionnent sur le même principe que les moteurs de recherche classique, à la différence qu'au lieu d'indexer des pages WWW, ils indexent les fichiers des serveurs FTP (File Transfert Protocol). Ces serveurs sont principalement dédiés au télédéchargement. Il est donc possible via ces serveurs de récupérer des fichiers quelque soit leur type (.doc, .ppt, .zip, .exe …)
Les moteurs de recherche sur les FTP indexent seulement le nom du fichier ainsi que les répertoires d'accès aux fichiers.

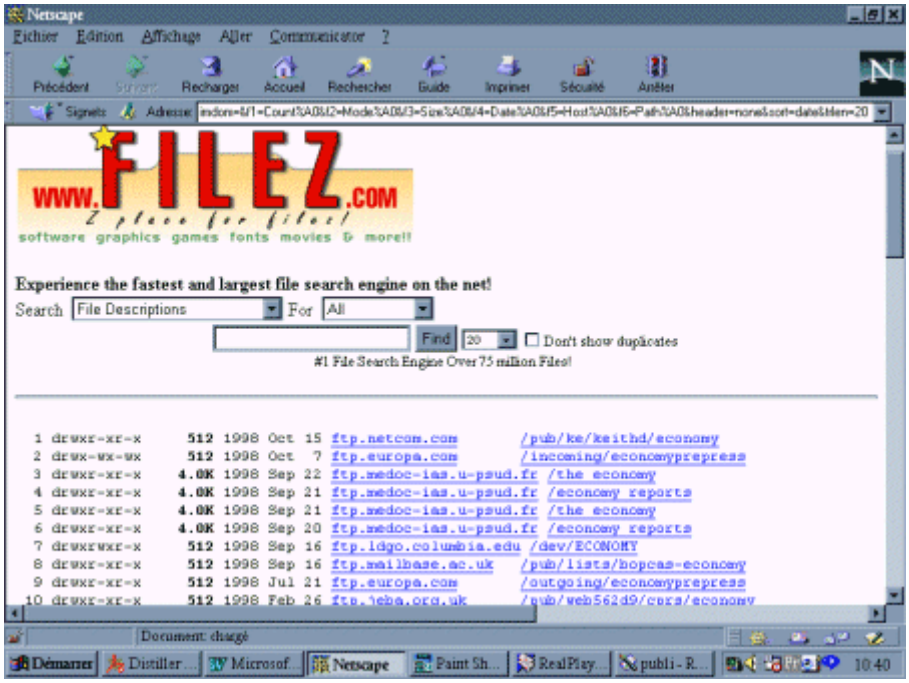

Figure 7 : Exemple de recherche sur Filez (http://www.filez.com)

L'inconvénient majeur réside dans le fait que l'utilisateur ne peut pas directement connaître le contenu du fichier, et qu'un seul nom n'est souvent pas assez explicite.
La recherche est donc réduite, car elle ne se fait pas sur le contenu du fichier.
Dans l'exemple de la figure 7, la recherche sur Filez.com s'est effectuée sur le nom : economy. Les réponses obtenues sont donc des fichiers mis à disposition du public comportant economy dans leur nom ou leur répertoires d'accès, mais l'utilisateur

n'aura aucun autres renseignements sur le contenu à proprement dit des fichiers. Les renseignements accessibles seront la date, l'adresse FTP du site, la taille du fichier, son extension (si c'est le nom du fichier qui comporte le mot-clé) et les droits du fichier.
Les droits de fichiers sont en unix (ou linux). Il faut décomposer en 4 parties :
d : directory
rwx : accès en lecture écriture exécution pour le propriétaire
r-x : accès en lecture exécution pour son équipe
r-x : accès en lecture exécution pour le reste du monde

1.3.1 Les Annuaires

Pour retrouver des personnes qui n'ont jamais discuté via des news, il est possible d'utiliser un autre type d'outil : Les Annuaires.
C'est un service qui permet de retrouver une personne ainsi que son mail, via certains critères comme une partie de l'adresse, le nom ou le prénom. Il est aussi possible de trouver le numéro de téléphone de celle-ci.
Pour cela, il faut que préalablement cette personne est remplie un formulaire d'inscription pour se faire recenser auprès des annuaires.

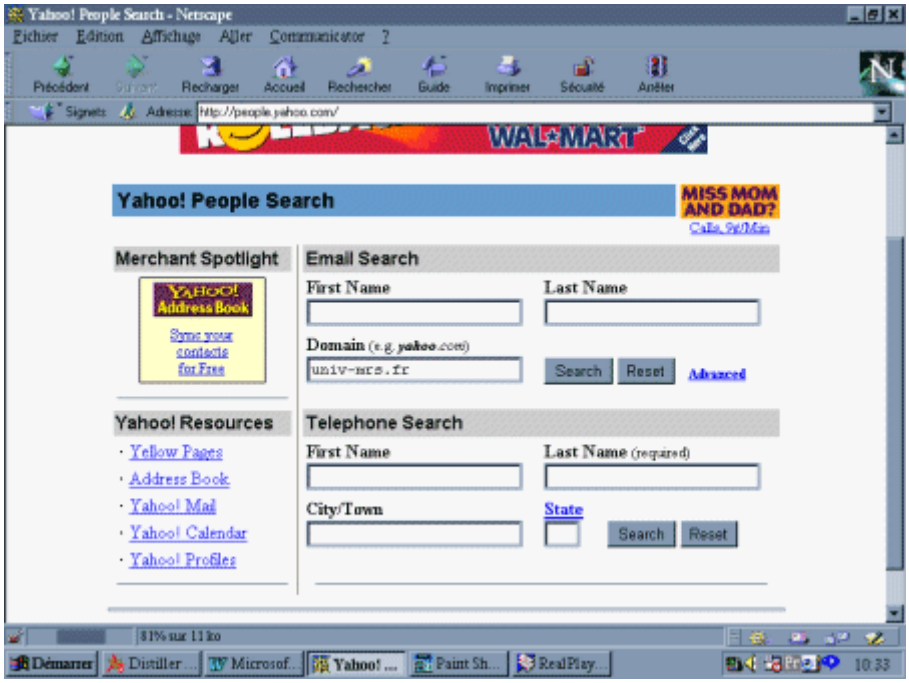

Figure 8 : Exemple d'annuaire : L'annuaire de Yahoo (http://people.yahoo.com)

2 Les Méta-outils

Il existe deux types de Méta-outils :
Les méta-index : ils interrogent simultanément différents index et moteurs de recherche.
Les guichets uniques : ils offrent sur une même page de leur site, les formulaires de multiples outils de recherche.

2.1 Les méta-index

Ces outils ont une interface de recherche unifiée, c'est-à-dire qu'ils proposent à l'utilisateur un formulaire unique donnant accès à plusieurs outils de recherche de façon simultanée. C'est la même requête (avec la même syntaxe) qui sera envoyée à tous les outils.
Les meilleurs Méta-outils fournissent les caractéristiques suivantes :

- élimination des doublons,
- vérification de la validité des liens,
- calcul d'un nouvel indice de pertinence
- présentation une liste unifiée de résultats

L'exemple suivant permet de collecter l'information sur la veille technologique en interrogeant simultanément les principaux moteurs de recherche du Web. 37 réponses sont obtenus et les résultats sont du même type qu'un simple moteur. Pour information, Webcrawler fournit le nom du ou des moteurs ayant trouvés la réponse.

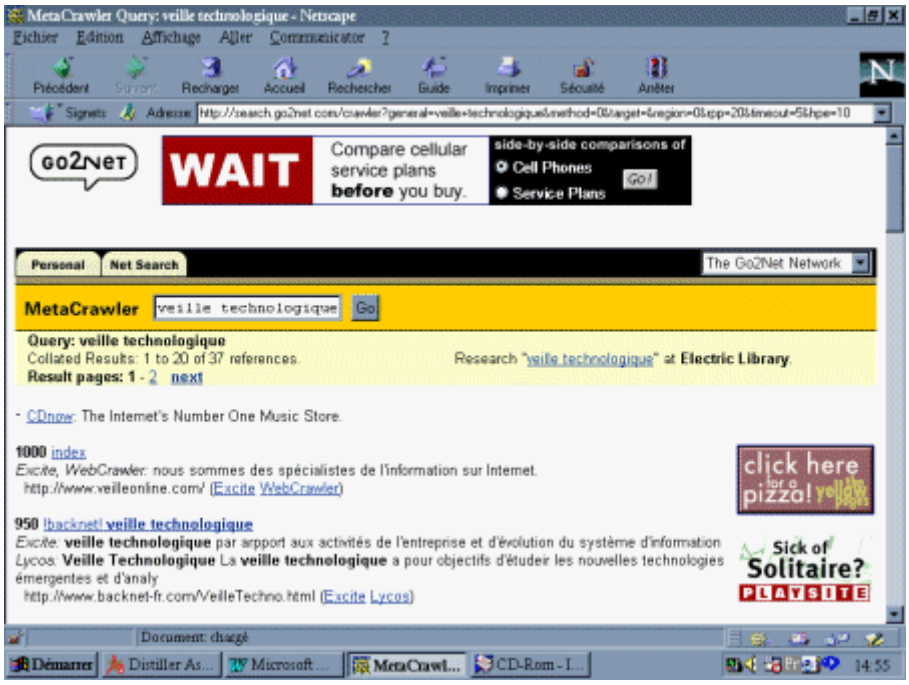

Figure 9 : Exemple de résultats sur Webcrawler (http://www.webcrawler.com)

L'inconvénient de ce type d'outil réside dans l'appauvrissement des fonctions d'interrogation. En effet, la syntaxe étant unique à tous les moteurs, les spécificités de chacun ne sont pas pris en compte, ce qui implique un résultat moins ciblé.

De plus, le fonctionnement de ce type d'outil est limitatif. Ne pouvant pas traiter pour chaque moteur l'intégralité des réponses, le méta-moteur va restreindre son choix de réponses aux premières pages de chaque moteurs interrogés.
Pour preuve, Webcrawler renvoie 37 réponses pour veille technologique, alors qu'Altavista à lui seul, trouve 6230 réponses.

2.2 Les guichets uniques

Les guichets uniques sont des pages Web donnant accès à différents outils de recherche d'informations (robots, index, annuaires…). Ceux sont en général de bons aiguilleurs, car les outils sont classifiés.
Une requête est envoyée à un seul outil à la fois, et l'usager n'a pas besoin de connaître les adresses de ces outils.
Il faut quand même utiliser la syntaxe d'origine de l'outil choisi.

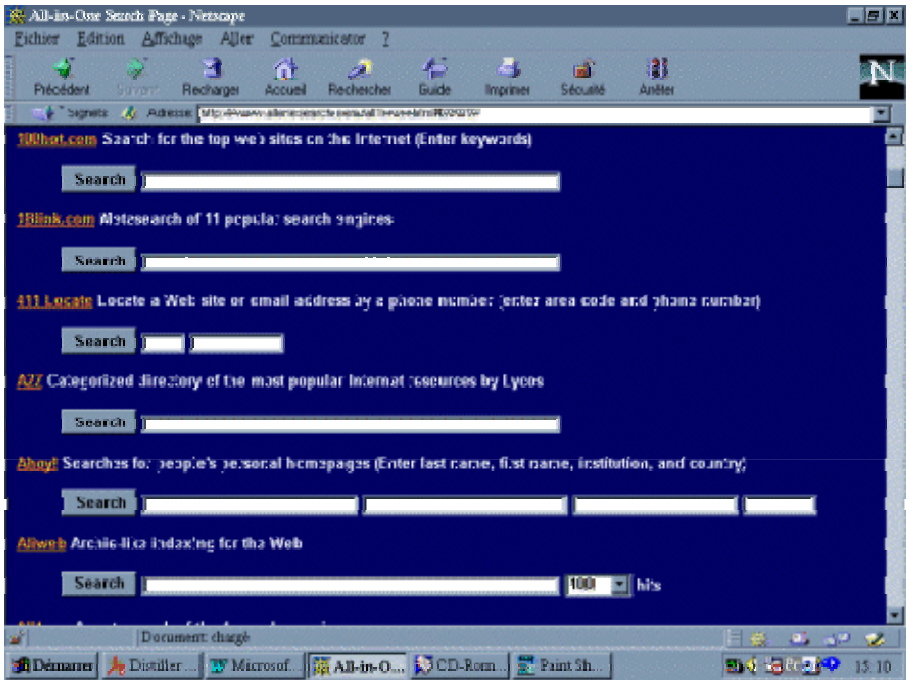

Figure 10 : Exemple de recherche sur All-in-one
http://www.allonesearch.com/all1www.html#WWW

3. Les News

Les News sont des moyens de communication qui permettent d'échanger de l'information sur des thèmes précis via la messagerie. Dans ce cas, l'utilisateur doit aller chercher l'information qu'il désire dans le groupe de news approprié.
Ces outils fournissent un certain nombre de forums, les newsgroups qui sont classés par thème de façon hiérarchique. Par exemple, tous les groupes qui commencent par fr. sont des groupes français. Les groupes commençant par fr.bio. sont les groupes français dédiés à la biologie, etc... Un article (une news) posté dans un groupe sera donc lu par toute personne lisant ce groupe.

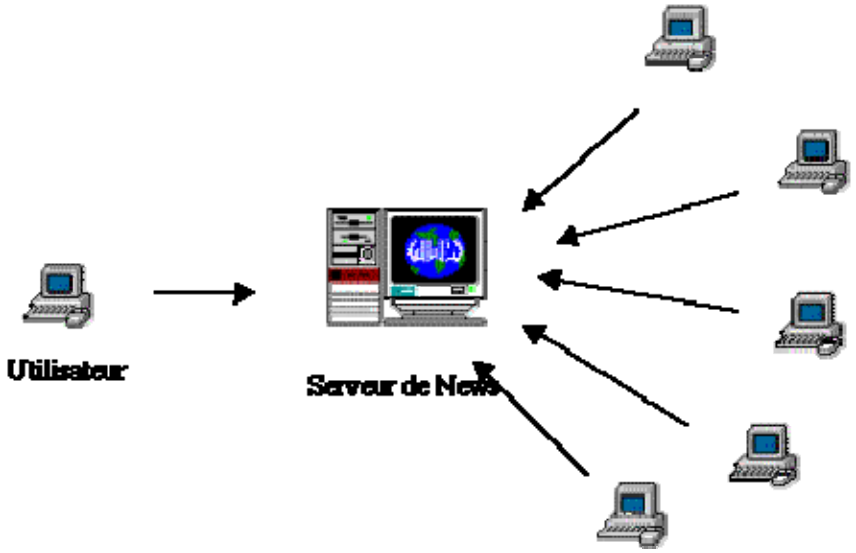

Figure 11 : Fonctionnement des serveurs de News

L'avantage principal de ce type de ressources est d'être aiguillé et même parfois renseigné sur le problème que l'utilisateur se pose. De plus, une bonne utilisation des news permet à un utilisateur d'avoir rapidement une ou plusieurs réponses.
Les réponses seront d'autant plus pertinentes que les experts du domaine sont souvent présents sur leurs forums.

Le problème majeur des forums réside dans leur accessibilité. En effet, chaque serveur de news interdit l'accès en écriture (impossible de poster des messages) aux utilisateurs ne faisant pas partie de son réseau physique. Par exemple, le personnel de l'université de Marseille (dont le domaine est u-3mrs.fr) ne peut accéder en écriture qu'au forum news.u-3mrs.fr. Il ne pourra que consulter tous les autres serveurs de news (news.univ-tln.fr : serveur de news de l'université de Toulon).
Cette restriction est d'autant plus gênante que les serveurs de news ne donnent pas accès à tous les forums. C'est à l'administrateur des news de choisir les groupes auxquels ils désirent s'abonner.
De plus, il n'existe aucun outil spécifique ou organisme qui référence la liste des serveurs de news. Seuls des outils comme dejanews peuvent seulement donner une indication sur le nom des groupes, mais l'utilisateur ne connaîtra pas le serveur auquel il faut être connecté.

4. Les listes de diffusion

Les listes de diffusion sont des discussions via la messagerie autour de divers thèmes. L'utilisateur reçoit l'information directement dans sa boite aux lettres.

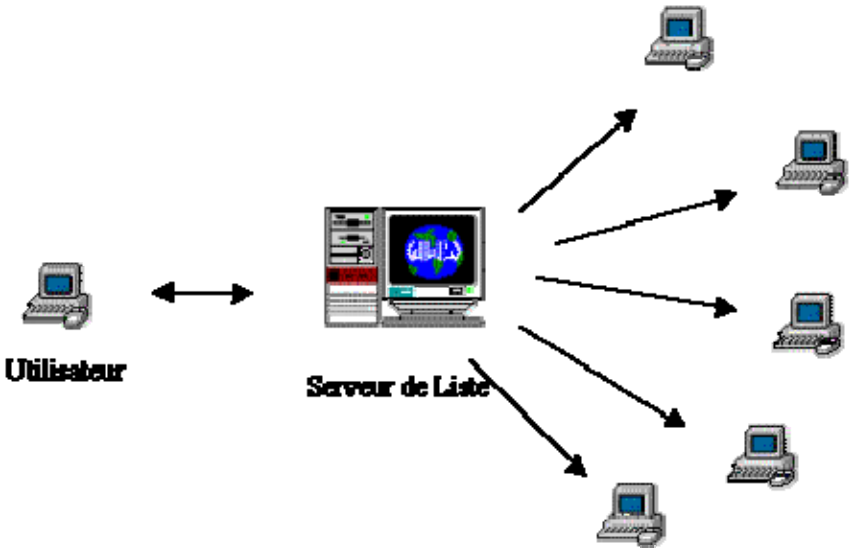

Figure 12 : Fonctionnement des listes de diffusion

Les listes de diffusion rassemblent des personnes qui partagent un même centre d'intérêt.
L'utilisateur a la possibilité de s'inscrire dans une liste par un certain nombre de commandes placés dans le corps du message du courrier électronique qui est adressé au robot de la liste. Par exemple :

"subscribe votre nom", permet de s'inscrire dans la liste de diffusion.
"unsubscribe" permet de se retirer de la liste.

Il existe des serveurs qui répertorient toutes les listes de diffusion[4]. Lorsqu'un utilisateur s'abonne à une liste de diffusion, il faut qu'il garde à l'esprit le problème du décalage horaire. En effet, une liste de diffusion américaine comporte l'inconvénient que les internautes abonnés à cette liste ne pourront répondre seulement le lendemain. Pour des réponses instantanées, il est préférable de s'abonner à des listes nationales ou dans le même fuseau horaire (ou adjacents).

Ce type de ressource permet de se faire rapidement un réseau virtuel d'experts.
Les avantages qu'offrent les listes de diffusion sont très proches des news, voire identiques. Seule la façon de communiquer est différente puisque le principe de la liste de diffusion se rapproche du "Dataware House". Le dataware house est un concept de rapatriement de l'information sur le poste de l'utilisateur. C'est par le biais des mails et de sa boite aux lettres électronique qu'il aura l'information directement sur son poste. Il n'aura pas à aller la chercher sur les serveurs de news pour les forums.

---

[4] Listes de diffusion internationales : http://catalog.com/vivian/interest-group-search.html
Listes de diffusion francophones : http://www.cru.fr/listes/

A la différence des news, toutes les listes de diffusion sont accessibles et répertoriées.

Ce principe de communication présente très peu d'inconvénients. Un détail important à ne pas négliger concerne la quantité du nombre de messages renvoyés par la liste quotidiennement. En effet, selon les listes de diffusion, un abonné peut recevoir plusieurs dizaines de messages par jour pour les listes les plus actives. Le cumul de plusieurs abonnements peut vite saturer le serveur de messagerie de l'utilisateur.
Dans ce cas, l'utilisateur sera vite submergé de messages et risquera de ne plus porter attention aux informations susceptibles de l'intéresser. "Trop d'informations noie l'information". Pour éviter cela, il est conseiller de ne pas s'abonner à trop de listes ou d'utiliser un outil de filtrage automatique des mails. Ces outils sont des robots passifs qui ne laissent passer dans la boite aux lettres que les messages correspondant à certains critères prédéfinis (filtre sur une partie de l'adresse électronique, mots-clés dans le corps du message…).
Dans les dernières versions des navigateurs, cet outil est directement intégré à celui-ci.

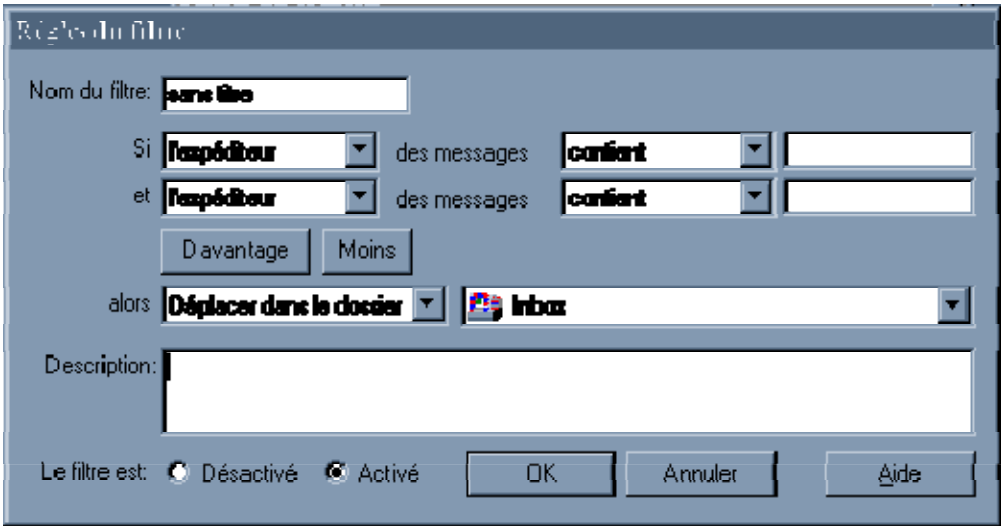

Figure 13 : Filtre des messages de Netscape

Netscape Communicator 4.03 offre la possibilité de paramétrer sa messagerie de façon à filtrer ses mails.
L'utilisateur a la possibilité de choisir comme champ de filtre : l'expéditeur, l'objet, le corps du message ou la priorité. Ces champs peuvent "contenir", "ne pas contenir", "être", ou "commencer" par ce que l'utilisateur va paramétrer.
Cette restriction peut être effectuée deux fois.
Il s'ensuit une action de déplacement du message, de sa suppression, de changement de priorité,…

5. Les Agents Intelligents

Les Agents Intelligents sont des logiciels «permettant d'automatiser, périodiquement ou à la demande, des tâches de façon transparente pour l'utilisateur qui bénéficie des résultats». Les agents intelligents sont capables d'aller chercher une information pertinente pour leurs utilisateurs dans la jungle de l'information que représente le monde du Web. Ils permettent de gagner du temps sur les connexions et sur le ciblage de l'information. Ils permettent de plus de surveiller, de détecter des sites selon des critères préétablis (même s'ils ne sont pas indexés par les moteurs de recherche).

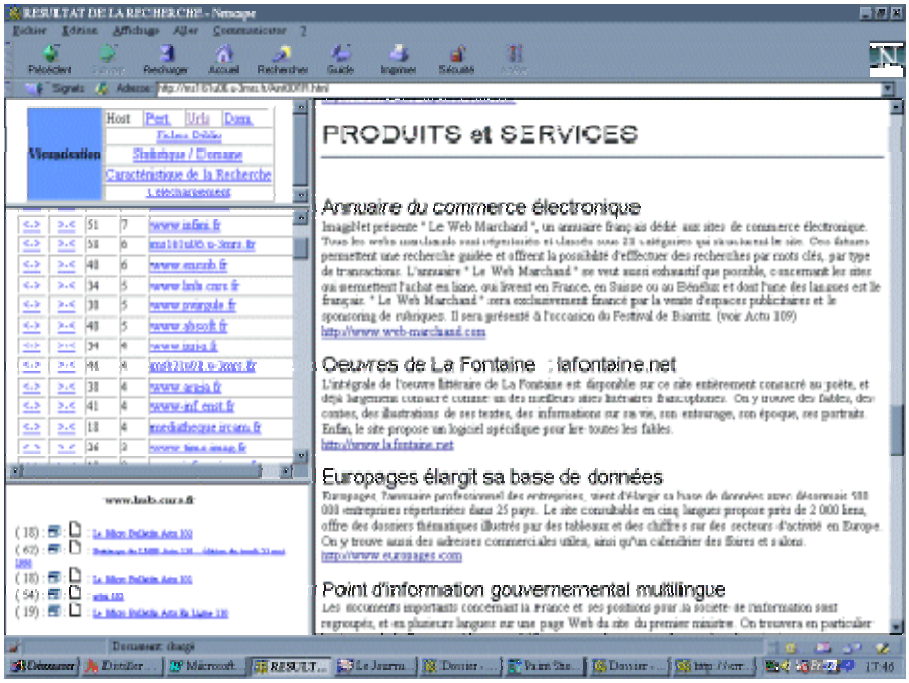

Figure 14 : Exemple de recherche avec un agent intelligent

Cet exemple montre les résultats d'une interrogation avec la stratégie suivante :

((agent* near intelligent*) and logiciel*) and domain : fr

Les réponses sont structurées de manière à améliorer la navigation de l'utilisateur dans les différentes réponses rapatriées par l'agent intelligent. Ces informations sont stockées sur le poste utilisateur et donc il n'y a pas de temps de connexion supplémentaire pour la navigation. L'utilisateur bénéficie ainsi d'un gain de temps très important.

Les agents intelligents sont en grande majorité des logiciels payants. Ceux qui sont utilisés pour faire de la Veille Technologique sont plus onéreux car ils permettent, outre la récolte et le stockage, une analyse des documents recueillies. Ces analyses sont représentées sous forme graphique (cartographie) afin d'avoir un aspect plus visuel de l'information recueillie.

Les agents intelligents se positionnent en aval des moteurs de recherche, car leur principe de fonctionnement se compose de deux phases.

Dans un premier temps, l'agent va se comporter comme un méta-outil, c'est-à-dire qu'il va se connecter à plusieurs moteurs de recherche; puis dans un second temps, il va parcourir les réponses renvoyées par les moteurs.
Les pages répondant à la requête seront ensuite stockées de manière structurée, voire même indexées sur le poste de l'utilisateur.

Section 2 : La Méthodologie de recherche

La section précédente donne un très large aperçu des ressources et outils de recherche sur Internet. Le problème est donc de savoir quel outil doit être utilisé pour récupérer tel type de données.

Le tableau 1 répertorie le type d'information recherchée, quels outils ou ressources sont à utiliser et quels résultats sont obtenus.

QUOI	OUTILS	RESULTATS
Informations générales	➢ Moteurs de recherche standard ➢ Virtuals Libraries	➢ Gateways de liens hypertextes ➢ Entreprises ➢ Sites personnels se rapportant au sujet
les experts dans le domaine	➢ Moteurs de Recherche standards ➢ Moteurs spécifiques (news) ➢ Liste de diffusion	➢ Nom des personnes ➢ Fonctions ➢ Mails ➢ Domaines d'intérêt
Revue de presse sur le sujet	➢ Sites de presses spécifiques et générales	➢ Articles gratuits ➢ Articles payants ➢ Archives ➢ Résumés ➢ Publications scientifiques
Rapports, analyses	➢ Moteurs de recherche standards ➢ Index ➢ Sites de prestataires publics et gouvernementaux ➢ Bases de données accessibles et gratuites	➢ Rapports ➢ Thèses ➢ Mémoires ➢ Publications scientifiques ➢ Exposé ➢ Présentations
Echange d'information	➢ Listes de diffusion sur le sujet ➢ News	➢ Réponses aux questions ➢ Noms des experts ➢ Informations informelles à valeur ajoutée
Propriété industrielle	➢ Bases de données accessibles et gratuites ➢ Sites de prestataires publics et gouvernementaux	➢ Brevets ➢ Marques déposées ➢ Informations générales sur le dépôt de brevets

Information de contrainte	➢ Sites gouvernementaux ➢ Sites spécifiques	➢ Réglementation, ➢ Juridique, ➢ Environnement, ➢ Sécurité

Tableau 1 : Comment rechercher l'information en fonction de son type

Dans une recherche d'information, la première étape indispensable, souvent occultée, est de bien définir les mots-clés. En effet, afin d'éviter les bruits et les silences documentaires, il est nécessaire de bien cerner son sujet : il ne faut pas être trop large pour ne pas être amené à analyser trop de réponses, mais à l'extrême, il ne faut pas être trop restreint pour passer "à côté" de l'information utile.
Dans un premier temps, il faut donc affiner ou élargir la stratégie d'interrogation sur un ou deux moteurs de recherche généraliste afin de trouver la meilleure stratégie.
Une fois celle-ci au point, toutes les ressources du web doivent être explorées. Puisque celles-ci sont hétérogènes, l'éventail des réponses et leurs contenus en seront d'autant plus larges, allant de l'information générale à l'information spécifique.
De plus, il est important de passer outre la timidité pour certains ou la gène pour d'autres pour poser des questions aux personnes spécialistes du sujet. Dans la plupart des cas, celles-ci répondront volontiers aux questions et/ou aiguilleront l'utilisateur vers des sites pertinents. Cette approche très communicante est fort courante sur Internet. C'est d'ailleurs par le biais de la construction d'un réseau virtuel d'experts que l'utilisateur trouvera le maximum de réponses pertinentes.

Ce type d'information qualifiée d'informelle, est souvent à valeur ajoutée. En effet, les ressources Web sont accessibles à tous. Il faut pour trouver de l'information pertinente (ou que les autres n'auront pas) dans une première phase, avoir la possibilité de faire des recherches rapides et efficaces et de trouver des sites non indexés avec les agents intelligents.
Dans une seconde phase, la communication avec les experts, sur des listes de diffusion ou dans des forums permet de capter de l'information supplémentaire, soit sur des sites qui étaient passés au travers de la recherche, soit de l'information brute ou élaborée directement sur le sujet. Ce type d'information ne peut être récupéré que si l'utilisateur entre en contact direct avec les experts. Dans ce cas, cette information sera personnelle, car les concurrents n'auront pas accès à cette même information (s'ils ne font pas la même démarche).
Il est évident qu'il est primordial de valider ce type d'information auprès d'autres sources, car l'infoguerre et la désinformation sont souvent pratiquées sur Internet.
Que ce soit sur les sites Web, ou dans des forums, listes ou par contact direct, la validation de l'information est obligatoire.

De plus, lorsque la recherche d'information engendre un contact direct, soit par le biais des listes, forums ou mails, il est important d'avoir une stratégie de communication pour ne pas se dévoiler, surtout si le sujet est critique pour l'entreprise.

En plus de la méthodologie de recherche d'information sur Internet, il est important d'avoir une stratégie de mémorisation et d'agencement de l'information collectée. Que ce soit des pages HTML, des images, des fichiers ou des mails, il est important de conserver des traces de sa source. De plus l'agencement de cette information est primordiale. Qu'elle soit stockée dans un répertoire, une base de données ou directement dans un bookmark, il faut penser à la structuration de ces données, par types, thèmes ou par importance.
Avant toutes choses, il faut garder à l'esprit à qui sont destinés ces informations, et par quel biais elles vont être communiquées. Ceux sont deux aspects indispensables pour élaborer son dossier d'information, qu'il soit électronique, sur Intranet, en base de données ou simplement sous format papier.

Bibliographie

"Recherche d'information sur Internet", Ghislaine Chartron, CNAM, septembre 1998

"GIRI - Guide d'initiation à la recherche dans Internet"
http://www.bibl.ulaval.ca/vitrine/giri/
CREPUQ, 1996

"AURESYS 2.0 : Un agent Intelligent au service de l'information stratégique"
Quoniam Luc, Bruno Mannina, Dou Henri, CRRM, SFBA'97, Ile Rousse

"Nouvelles Technologies. Pas si bêtes, ces agents…" [en ligne]
M. Baccar, Génération Internet : http://www.entreprises-virtuelles.ch/art_12.htm
(consulté le 28-09-98)

"L'information Scientifique et Technique", F. Jakobiak, 1996, Que sais-je?

"Veille Technologique et Compétitivité", H. Dou, 1995, Dunod

"La recherche intelligente sur Internet" H. Samier, V. Sandoval, Ed. HERMES

"L'Internet et l'Intelligence Economique", P. Oberson, Les éditions d'organisation

"Beaucoup Serach Engine" http://www.beaucoup.com/