



SOMMAIRE

Avant-propos
par Jean-Claude Le Moal
INRIA, Unité de communication et
information scientifique

1 - Instruments de recherche sur le Web (p.11-70)
par Sylvie Dalbin, ATD – DESYBEL

2 - XML et la documentation structurée: des principes aux techniques (p.71-97)
par François Role. Ministère de la Recherche

3 - Les métadonnées : accès aux ressources électroniques (p.99-135)
par Marie-Élise Fréon. Jouve

4 - Traitement automatique des langues et recherche d'information (p.137-168)
par Pascale Sébillot. IRISA

5 - Des bibliothèques traditionnelles aux « bibliothèques virtuelles» (p.169-201)
par Dominique Lahary. BOP Val d'Oise

6. De la sémantique des contenus à la sémantique des structures (p.203-229)
par Laurent Romary, INRIA/Loria

7. Recherche interactive dans les documents multimédia (p.231-256)
par Nozha Boujemaa, INRIA Rocquencourt

8. Veille stratégique sur les réseaux (p.257-300)
par Armelle Thomas, Inforizon.

<i>Répertoire des sigles utilisés</i>	301
<i>Table des matières</i>	309
<i>Adresses des auteurs</i>	321

Pour citer l'article :

Dalbin, Sylvie. - « Instruments de recherche sur le web ». - La Recherche d'information sur les réseaux. Cours INRIA, 30 septembre - 4 octobre 2002, Le Bono (Morbihan). - Paris : ADBS Éditions, 2002. p.11-70.

Instruments de recherche sur le Web

Sylvie Dalbin

INTRODUCTION : RICHESSE ET COMPLEXITÉ DE L'INTERNET

L'augmentation exponentielle des volumes d'information véhiculée par les réseaux, leur extrême variété tant par leur forme, leur contenu que leur origine, le foisonnement des innovations dans le domaine des technologies de l'information¹, ainsi que les changements opérés dans les pratiques des internautes-utilisateurs d'information, ont fortement complexifié l'environnement de travail des professionnels de l'information et de la documentation (voir annexe I, p. 64).

L'Internet reste un espace éditorial enrichi et complexe, et ses avancées les plus récentes renforcent la dualité de cet espace [10] : tradition / innovation, structuré / non structuré, marchand / non marchand (gratuit / payant), en sous-réseaux professionnels (portails, liens), ouvert / fermé (intranets, extranets, Web invisible), instabilité / stabilité ; volatilité / pérennité (archives du Web), disponibilité / indisponibilité (formats / codages multiples et hétérogènes), traité / non traité, à caractère personnel / public ou professionnel, multilingue...

Les difficultés rencontrées sur le Web (recherche d'information, mais aussi sécurité, etc.) sont à la mesure inverse de la simplicité avec laquelle tout internaute peut éditer et échanger. Beaucoup d'acteurs, économiques, politiques, scientifiques, se sont penchés sur ces questions, chacun apportant sa solution.

En ce qui concerne la recherche d'information, et plus particulièrement les outils d'orientation et d'accès à l'information, les réponses apportées par les différents acteurs aux problèmes rencontrés sur le Web sont de nature variée :

- conceptuelle et organisationnelle : conception de sous-espaces (portails), document structuré et granularité de l'information (voir chapitres 2, 6 et 7), comité d'édition, règles d'identification à la source, normes, meilleure connaissance et prise en compte des utilisateurs de l'information, etc. ;
- technique : perfectionnement d'approches déjà anciennes (calcul de pertinence, etc.), intégration de techniques du langage naturel (voir chapitre 4), traitement des documents de nature autre que textuelle (voir chapitre 7), agents de recherche, exploitation de l'architecture hypertextuelle du Web, agrégateur de contenu, etc. ;
- marketing : co-opération, référencement ou positionnement payant.

De toute cette effervescence, il est possible de faire émerger les grandes tendances à partir de l'étude de l'offre actuelle (ce sera l'objet de la première partie du présent chapitre, ci-après) et d'un catalogue des fonctionnalités mises en œuvre dans les outils de recherche par Internet (deuxième partie, p. 29). Les questions d'évaluation des ressources et des instruments de recherche sont abordées dans une troisième partie (p. 45).

LES INSTRUMENTS DE RECHERCHE POUR LE WEB : L'OFFRE ACTUELLE

Les volumes et les flux de ressources disponibles via Internet ont nécessité, assez rapidement après la création des réseaux mondiaux, le développement d'outils spécialisés pour en permettre le repérage et la localisation. Les annuaires et les moteurs de recherche, surtout ceux à vocation généraliste, sont ainsi devenus les sites les plus utilisés. Devant les difficultés éprouvées par les internautes (spécialistes ou non de la recherche d'information) pour obtenir, avec ces outils généralistes, des réponses adéquates à certains types de requête, d'autres instruments se sont développés : des métamoteurs, des annuaires et des moteurs de recherche spécialisés. Plus récemment sont apparus des outils spécialisés pour le Web invisible : les portails et les anneaux thématiques.

Il semble important aujourd'hui, tandis que ces autres formes d'outils se développent, de préciser la terminologie employée et les caractéristiques qui les distinguent. Le terme « moteur de recherche », par exemple, a souvent été employé dans des contextes et pour des usages différents. Il renvoie en particulier soit à une plate-forme logicielle, soit à un service d'accès à l'information [12]. Nous emploierons ici le terme générique d'« instrument de recherche » pour l'ensemble des outils actuellement exploitables par Internet, qu'ils soient construits à partir de procédures manuelles ou informatisées, et quel que soit le format des fichiers des ressources électroniques manipulés. Le terme « moteur de recherche » sera conservé pour une famille particulière d'instruments de recherche que nous définirons plus avant.

Sont exposés ici les changements majeurs intervenus ces dernières années dans ce domaine² (voir aussi le chapitre 5 et [22] [23] [24] [25]).

Territoire informationnel couvert par les instruments de recherche

L'information disponible par Internet est d'une grande hétérogénéité, avec des contenus dynamiques et en renouvellement continu, des ressources non directement visibles (Web invisible), une très forte instabilité des localisations (de plus en plus d'erreurs de type « 404 »), une grande diversité linguistique et une couverture géographique mondiale.

La diversité des types de formats pris en compte par les instruments de recherche – autres formats de nature textuelle, comme les formats PDF ou Word, mais également ressources « images » ou sonores – et le développement des pratiques et des usages des réseaux sont des facteurs qui concourent à l'explosion des volumes et des flux disponibles.

Dans ce contexte, les outils de repérage et d'orientation déploient des efforts importants pour améliorer l'étendue (profondeur, largeur) des ressources électroniques prises en compte et les délais de rafraîchissement. Mais la situation technologique et financière réclamait des solutions autres que strictement technologiques – comme le déploiement de machines, la segmentation des fonds ou encore le repérage des doublons. Des solutions de type commercial sont apparues en premier lieu, dont certaines, comme la soumission payante, furent assez controversées. Parallèlement se dessinent une segmentation forte du marché et un déploiement d'instruments de recherche spécialisés, comme les portails ou les moteurs spécifiques, par exemple pour l'information d'actualité (Google, Fast avec des recherches possibles sur la base News ou l'onglet News & Ressources). Enfin, on assiste à un renouveau des pratiques collaboratives pour le repérage et le référencement des ressources – pratiques qui sont à l'origine d'Internet, avec la constitution de ressources coproduites, comme l'annuaire ODP.

Web visible et Web invisible³

Toutes les ressources disponibles ne sont pas exploitées par les moteurs de recherche ou les annuaires, et ce pour différentes raisons.

– Les moteurs de recherche sur le Web ne prennent en compte qu'un nombre limité de formats de fichiers : le format HTML ou XML, puis celui d'Adobe Acrobat (PDF) ou Powerpoint de Microsoft.

– De nombreuses ressources sont accessibles sous des protocoles différents de celui du Web (comme FTP). Des évolutions récentes de plusieurs moteurs de recherche autorisent l'exploitation simultanée de plusieurs protocoles, dont HTTP (pour le format HTML), FTP et Gopher.

- Des pages possèdent des caractéristiques techniques rendant difficile, sinon impossible, l’indexation par les moteurs : *frames* (cadres), scripts modifiant le contenu des pages, technologies propriétaires (par exemple Flash, Active X, Java).
- D’autres pages sont produites dynamiquement à partir de bases de données ou d’applications ; leurs URL comportent des paramètres non exploitables par la plupart des moteurs ; on parle de « pages dynamiques » (.asp, .php, .pl, etc.). On peut noter que le robot exploité par Google suit des liens dynamiques, indexant des pages ayant des liens du type `index.php?var1=valeur1&var2=valeur`. Lors d’une recherche, la visualisation de la page trouvée reste toutefois un problème, qui nécessite d’utiliser la fonction « cache » proposée par le moteur.
- Certaines pages dites « orphelines », se trouvant sur des serveurs non identifiés, n’ont fait l’objet ni d’un référencement direct ni d’aucun lien à partir d’une autre page. Ces pages orphelines peuvent correspondre à des liens morts, à des liens erronés, dupliqués, à des pages protégées, à des pages sur un intranet, à des pages sur lesquelles le robot ne passe jamais. Il arrive cependant que certaines parties de ces pages – les mots des adresses (URL) et ceux des ancres ou autres pages liées – soient indexées par des moteurs comme Google.
- De nombreuses pages nécessitent une identification préalable de la part de l’internaute pour pouvoir être affichées.
- Des pages d’un site peuvent être « interdites de référencement », information indiquée dans le fichier Robots.txt. [24].
- Ces pages sont produites à partir de données saisies par l’internaute via un formulaire. C’est en particulier le cas des formulaires d’interrogation des banques de données. À ce sujet, on peut citer le produit Quigo⁴, qui traite les pages avec un formulaire de requête. Des algorithmes décodent ces formulaires pour déterminer quels types de requêtes ils exigent (mot de passe, nom, chaînes de caractère, variable numérique), puis Quigo crée un agent qui pourra spécifiquement interagir avec le formulaire.
- Lorsque le moteur n’indexe que partiellement la page, le(s) mot(s) recherché(s) peu(ven)t ne pas apparaître dans les parties indexées.
- Le dernier motif est l’incapacité technique des instruments de recherche (même spécialisés, et donc brassant un territoire plus restreint) à assurer une prise en compte exhaustive des informations disponibles, face aux volumes et aux flux des ressources publiées ou mises à jour.

En résumé, le Web invisible correspond à l'ensemble des documents (textes, images fixes, son, vidéo, etc.) non indexés par les instruments de recherche automatiques ou manuels (moteurs, annuaires...). On parle également du Web caché, par opposition au Web visible manipulé par les instruments de recherche.

Le volume estimé du Web invisible ne cesse de croître même si de nombreux outils, souvent très spécialisés par thème ou type de données, permettent de traiter ces ressources cachées [8]. On peut citer l'annuaire InvisibleWeb⁵, qui scrute parmi plus de 10 000 bases de données d'archives, de catalogues, et organise le fonds repéré de manière thématique.

Moteurs et annuaires généralistes

Nous traiterons, dans cette partie, des deux familles d'instruments de recherche les plus anciens et les plus usités :

- les moteurs de recherche (*search engines*), qui indexent automatiquement les pages des sites pour y permettre un accès par mots clés ;
- les annuaires (*directories*), qui classifient les sites par des méthodes manuelles, offrant un accès par catégories.

Les moteurs de recherche

L'objectif de ce type d'instrument de recherche est de répertorier des sites et de mémoriser tout ou partie de leur contenu sous forme d'une base d'index (*catalogue, index database*) afin d'en faciliter l'accès par des méthodes de recherche en texte intégral. Ces outils se décomposent en trois modules.

- **Un robot logiciel explorateur (*spider, crawler*).** Ce module explore les réseaux de liens, et collecte le contenu des ressources accessibles dans une base de données. Ces robots sont peu nombreux, et certains – comme celui d'Inktomi, nommé Slurp – sont exploités par plusieurs moteurs de recherche. Le paramétrage de ces robots, essentiel pour la qualité de la collecte, permet de caractériser les formats de fichiers des ressources pris en compte, le traitement du fichier robot.txt, la profondeur et / ou la largeur des sites pris en compte, la nature des traitements opérés sur les liens hypertextuels contenus dans les pages et éventuellement ceux des pages liées, le rythme de la surveillance. Pour améliorer la performance des moteurs, le rythme de passage du robot, qui s'étend sur une période comprise entre quelques heures et plus d'un mois, peut être programmé en fonction, par exemple, d'un type de sites : les plus évolutifs sont contrôlés plus fréquemment. La date de fraîcheur entre la page la plus récente (a) et la page la plus ancienne (b) au sein d'une base d'index donne la mesure des différences entre les moteurs. Le site SearchengineShowdown⁶ fournit, à la date du 4 avril 2002, les chiffres suivants :

- Google (a) 1 jour (b) 68 jours
- MSN (Inktomi) (a) 1 jour (b) 80 jours
- HotBot (Inktomi) (a) 1 jour (b) 136 jours
- AltaVista (a) 12 jours (b) 51 jours

• **Un système d'indexation.** Un index est alors construit à partir des données fournies par le robot. L'indexation suit des règles différentes suivant les moteurs : tout ou partie de la ressource, indexation différenciée en fonction de la structure de la ressource, traitement linguistique, etc.

• **Un logiciel de recherche (*searcher*).** Le module de recherche est la partie visible qui permet d'interroger la base d'index en formulant une requête. Ce module effectue un appariement entre la requête posée et les éléments contenus dans la base d'index. Les fonctionnalités proposées par les moteurs de recherche sont très diverses (voir annexe II, p. 65) ; les évolutions récentes, qui prennent en compte les liens et affinent les traitements, se détachent fortement de ces anciens produits. En général il est possible, en recherche dite « avancée », de formuler une requête complexe, en particulier :

- par date ou période, avec une particularité sur le Web où la date indexée est celle de la dernière mise à jour des documents dans la base d'index, et non la date intégrée au document par l'auteur ;
- par métadonnées : mots clés, titre, mots de l'URL, etc. ;
- en exploitant les opérateurs booléens et / ou de proximité.

Certains moteurs assurent des traitements de la requête (reformulation, correction orthographique). Les moteurs de recherche ordonnent les réponses à la requête en s'appuyant sur différents critères dits « de pertinence » (voir p. 33).

De nombreux outils de recherche exploitent avec une interface différente (module *searcher*) un même index. Ainsi, par exemple, la base d'index d'Inktomi est exploitée par MSN et Hotbot.

Les annuaires

Les annuaires sont des catalogues de ressources classées par catégories, et organisées hiérarchiquement [51]. La sélection des ressources identifiées et leur classement s'effectuent manuellement. Le repérage des ressources fait l'objet d'un travail de veille par les documentalistes des annuaires, complété par des actions d'autoréférencement par les éditeurs des ressources.

Les descriptions des ressources sélectionnées y sont plus ou moins riches. Des méthodes et des règles, parfois décrites dans des cahiers des charges (« chartes éditoriales ») proposés à l'internaute, assurent une relative homogénéité de traitement.

De nombreux critères de sélection peuvent être mis en œuvre à priori : sites illégaux, sites en construction ou sans contenu réel, sites jugés trop personnels, sites et catégories sélectionnés strictement en fonction des centres d'intérêt des internautes étudiés à travers l'étude de leurs requêtes (AOL), etc. D'autres annuaires ont un fonctionnement plus contributif et font appel, pour évaluer la qualité des sites, soit à des experts rémunérés (comme About⁷, par exemple), soit à des internautes bénévoles dont la compétence est reconnue pour réaliser cette qualification des sites (ODP), soit enfin à des centres spécialisés répartis, telle que la Virtual Library⁸. D'autres répertoires, souvent plus spécialisés et très sélectifs, mettent en place des critères de qualité, voire des notations (voir l'exemple de la grille d'évaluation de Netscoring, p. 51).

Les annuaires donnent une vue d'ensemble d'un domaine à l'utilisateur ; celui-ci peut ensuite naviguer de sous-catégories en sous-catégories, ou d'un site à un autre à l'intérieur d'une catégorie. Certains permettent une recherche par mots clés, mais uniquement sur les catégories et sous-catégories et non sur les sites. Ils ne permettent pas de réaliser des requêtes complexes, celles-ci s'effectuant sur les sites, et non sur les pages des sites. Les mises à jour, des sites mais également des catégories, ainsi que le désherbage, restent des problèmes importants inhérents à ce type d'instrument de recherche.

Développements actuels

Les annuaires et moteurs de recherche généralistes qui visent la représentation du Web plutôt que son exhaustivité poursuivent leur déploiement sur plusieurs axes.

COMPLÉMENTARITÉ ANNUAIRES ET MOTEURS DE RECHERCHE

Chaque type d'outil a ses avantages et ses inconvénients, et il est vite apparu nécessaire de proposer à l'internaute les deux formes d'accès. Des alliances se sont donc construites entre moteurs et annuaires. Par exemple, l'annuaire Yahoo! exploite le moteur Google ; l'annuaire ODP est utilisé par plusieurs fournisseurs d'accès Internet ou moteurs de recherche (Google, par exemple) pour une recherche complémentaire. Les partenariats évoluent : ainsi Tiscali (anciennement Nomade) a changé de moteur de recherche, préférant Google à Inktomi.

AMÉLIORATIONS TECHNIQUES

Les moteurs de recherche poursuivent les améliorations techniques (voir p. 29-45) sur l'ensemble de la chaîne : la prise en compte d'objets de formats divers et un travail sur l'architecture de la base d'index pour

améliorer les performances, l'ajout de traitements linguistiques pour enrichir l'indexation, des aides à la formulation de la requête ou à la présentation et à l'exploration des résultats (classification des résultats, cartographies). Quant aux annuaires, ils proposent des formulaires pour des référencement directs par les webmestres.

PARTICULARISMES RÉGIONAUX ET LINGUISTIQUES

En favorisant les échanges entre communautés au niveau mondial, Internet a engendré des besoins forts en termes de traduction ou de recherche sur des ressources multilingues. La qualité des services gratuits et en ligne de traduction (Systran sur le moteur de recherche d'Altavista, ou la solution de Google – en version bêta vers l'anglais uniquement) suffit soit au lancement d'une recherche élargie sur d'autres versions linguistiques par une traduction de la requête, soit au repérage des meilleurs documents dans le lot de résultats grâce à une traduction de type « lecture à vue » dans d'autres versions linguistiques. Dans ce deuxième cas, la traduction – même médiocre – doit permettre à l'internaute de décider de la suite à donner : abandonner le document ou tenter de réaliser une traduction plus affinée du document ainsi choisi. De plus, les grands annuaires et moteurs généralistes ont déployé des efforts importants pour étendre la prise en compte d'autres régions linguistiques, soit par référencement manuel, soit par des traitements appropriés des particularités linguistiques⁹.

La recherche inter-lingue, c'est-à-dire la recherche exprimée dans une langue pour manipuler un fonds dans une autre langue, ne semble pas à ce jour opérationnelle sur des outils généralistes sur le Web.

SERVICES COMPLÉMENTAIRES

Les moteurs de recherche et les annuaires constituent des points majeurs d'entrée à Internet. À ce titre, ces outils ont développé, pendant quelques années, divers services complémentaires à la recherche d'information, se transformant ainsi en portails généralistes. Après une période d'enrichissement qui a parfois pu nuire à la lisibilité des sites, les moteurs et annuaires cherchent depuis plusieurs mois à épurer leur interface (voir Altavista, par exemple).

Annuaire et moteurs spécialisés

Caractéristiques

Devant les difficultés rencontrées avec les instruments généralistes, de nombreux répertoires et moteurs spécialisés se sont développés. La notion

de spécialisation est prise ici au sens large, et les critères pour la définir sont variables. On trouve ainsi des instruments spécialisés :

- sur un domaine ou un secteur particuliers : le vin, le tourisme, le sport, l'industrie... Les répertoires spécialisés thématiques sont souvent le cœur d'un portail thématique, appelé alors « vortail » (portail vertical) ;
- sur une zone géographique ou linguistique précise ;
- sur la nature du document : foires aux questions, forums électroniques, messages de forums ou listes de discussion, bases de données, thèses, périodiques, actualités, articles de quotidiens, dépêches d'actualité, encyclopédies, bibliothèques électroniques, cartes, monographies, etc. ;
- sur les instruments de recherche : spécialisés dans le signalement des répertoires ou moteurs, généralistes ou spécialisés, métamoteurs, voire portails. Ces outils proposent parfois un signalement géographique (comme Indicateur¹⁰, par exemple). Certains (Beaucoup¹¹, SearchEngine¹², etc.) jouent parallèlement un rôle de métamoteur ;
- pour le Web invisible : Searchability¹³, liste des bases de données gratuites [15] ;
- suivant le type de fichiers : sur le site d'Adobe PDF Search¹⁴, plus d'un million de documents sont disponibles à cette adresse ;
- suivant la nature du média : son, image animée ou fixe.

Ces critères peuvent se croiser : par exemple, un annuaire thématique restreint à une zone géographique donnée, multipliant les offres possibles (voir Scirus, à titre d'exemple, en page suivante).

Tendance : des instruments de recherche de plus en plus spécialisés

Poussés par les contraintes inhérentes aux réseaux et aux besoins des internautes – sauvegarde des ressources du Web, niveau de qualité exigé – et pour répondre à certaines faiblesses des outils généralistes, les instruments de recherche se spécialisent de plus en plus. C'est dans ce contexte que les portails et les anneaux sont apparus.

Métamoteurs et moteurs humains

Plutôt que de développer de nouveaux instruments, certains se sont lancés dans une exploitation pour optimiser des outils existants : soit automatiquement, avec les métamoteurs ; soit par l'intermédiaire de spécialistes des réseaux : les moteurs humains.

Métamoteurs

Les métamoteurs¹⁶ interrogent simultanément, par le biais d'une interface unique, plusieurs moteurs de recherche et/ou répertoires. Selon les

Scirus est un moteur de recherche via Internet, spécialisé dans les domaines scientifiques. Il est l'aboutissement d'un partenariat entre le moteur de recherche Fast (AllTheWeb) et l'éditeur de revues scientifiques *Elsevier*. Son contenu est constitué de sites web à vocation scientifique, quel que soit le type d'éditeur (pages personnelles, sites universitaires...), de ressources libres ou issues de périodiques payants, de bases de données spécialisées (Medline, Ideal, Biomed...), ainsi que d'articles en *peer-review* extraits de ScienceDirect. Ces ressources sont issues pour une partie non négligeable du Web invisible, et prennent en compte également des formats de fichiers autres qu'HTML (PDF ou Postscript, par exemple). La recherche s'effectue, en fonction du type d'abonnement, sur environ 90 millions de ressources web et 17 millions d'enregistrements de type base de données grâce à un accès unifié.

Les niveaux d'interrogation, au nombre de trois, permettent d'étendre ou de réorienter la requête par présentation de termes extraits des documents entourant les termes de la requête (fonction affiner : *refine*). D'autres fonctionnalités, dont celles de surveillance sur des revues ou des domaines, sont proposées. L'éventail des fonctionnalités, maintenant classiques, est proposé avec Fast : opérateur + pour forcer la présence d'un mot clé, opérateurs booléens, recherche sur zones structurées (titre, auteur, affiliation, revue, mot clé, adresse, nom de domaine), choix de présence de tous les termes et de leur ordre, filtrage en fonction des types de documents, de la source, du domaine. Les résultats peuvent être triés par pertinence et sont présentés suivant les deux grandes ressources : celles du Web et celles des périodiques. Les résultats de la recherche peuvent être affinés par rétroaction. Ils peuvent être sauvegardés, tout comme les paramétrages de la requête.

Un débat ¹⁵ est en cours sur la question de la présentation des articles en *peer-review* sur le Web, et non périodiques.

produits, il peut y avoir une adaptation de la requête pour chaque moteur, puis une phase de fusion des résultats obtenus pour chaque moteur, et enfin le calcul d'un nouveau tri global de pertinence. Ils n'assurent pas eux-mêmes la collecte et l'indexation de ressources via Internet. Ils informent sur le moteur ou sur l'annuaire source. Mais, n'exploitant pas les fonctionnalités élaborées de chacun des moteurs ou annuaires, ils génèrent souvent beaucoup de bruit. Leur vocation est d'apporter une vision panoramique de résultats trouvés. Ils sont intéressants pour effectuer des recherches larges sur le Web, mais sont moins efficaces pour des requêtes précises.

Le développement des liens payants (voir p. 38) remet en cause la pertinence même des métamoteurs, d'autant que ce type de liens est moins facilement identifiable dans les résultats fournis par les métamoteurs que dans ceux fournis par l'outil d'origine ; ceci est d'autant plus vrai que les métamoteurs n'exploitent souvent que les premiers résultats de chacun des moteurs, c'est-à-dire majoritairement les liens payés ¹⁷.

Moteurs humains

En généralisant la notion d'instrument de recherche, on peut y inclure les moteurs « humains », reprenant ainsi le postulat que la recherche d'information nécessite des compétences particulières.

Les moteurs humains offrent un service de réponses, par des professionnels, aux questions des internautes. Ces services peuvent être gratuits ou payants, en mode direct ou différé¹⁸.

On peut citer le service WebHeld¹⁹, version française de Web Wizards. Google vient d'introduire en version bêta un service payant de même type : Google Answers. Google Answers²⁰ propose à l'utilisateur un formulaire pour préciser la question, incluant le choix d'une catégorie associée et l'indication d'un montant a priori en fonction de la difficulté et de l'urgence de la question. Le développement de ce service sur Google amène à envisager ce type d'offre en complément des modèles proposés par les autres instruments de recherche (moteurs, annuaires, portails, anneaux).

Portails

Qu'y a-t-il de commun entre les « portails » de Yahoo!, de MSN, d'AOL, de Medisite (l'Internet de la santé) ou des professionnels de l'éducation (EduClic), ou encore le portail gouvernemental français ? Il ne s'agit pas ici de figer dans une définition précise un objet relevant du monde de l'Internet, mais, devant tant de diversité, de tenter d'en préciser les caractéristiques principales.

Le portail est souvent considéré comme une page web regroupant un ensemble d'informations qui abordent le même thème. La définition donnée sur le site de Portail 2²¹ ne nous semble pas suffisante pour distinguer un portail à fort contenu d'un site spécialisé ou d'un répertoire.

Un point d'histoire

L'idée du portail, en tant que point d'accès privilégié à des ressources pour des utilisateurs ciblés, n'est pas nouvelle. Il s'agit, comme son nom l'indique, d'une porte qui offre, à partir d'un point unique, l'accès à un regroupement d'informations ou de services qui, séparément, auraient représenté pour l'utilisateur un investissement important (temps, ressources).

Mais jusqu'à ce que cette idée se développe et prenne la forme enrichie que nous lui connaissons actuellement, en partie grâce à l'avènement du Web depuis 1994, les techniques limitaient grandement les fonctionnalités et donc la portée de ces outils.

Ainsi, entre 1986 et 1994, les BBS comme AOL, CompuServe, Calvacom... offraient un point unique d'entrée à un ensemble de services et produits sélectionnés en fonction des besoins supposés des clients, mais ceux-ci comportaient certaines limites :

- l'accès à cette offre n'était possible qu'aux membres ou clients dûment référencés, ce qui permettait par contre des services marchands – mis en œuvre dès le début des années 1990 (porte-monnaie électronique) – et des services spécifiques (accès à telle encyclopédie ou tel magazine pour lesquels un partenariat avait été effectué) ;
- l'offre était moins importante qu'aujourd'hui ;
- la personnalisation était très limitée ;
- l'ergonomie des interfaces était celle de cette époque ;
- l'accès à Internet (à l'époque, aux ressources Gopher ou Wais) à partir de cette porte était limité, voire impossible.

Les BBS ont évolué, tout d'abord en donnant un accès à Internet – le terme de « fournisseur d'accès à Internet » a alors été utilisé, y compris pour les BBS – puis, aujourd'hui, pour les plus importants d'entre eux, en proposant des « portails » accessibles sur Internet, comportant un accès particulier pour leurs membres.

Les fonctions (accès unique, fonction d'achat, d'échange, télédownload, actualités...) ne sont donc pas réellement nouvelles ; mais les techniques de l'époque, ainsi que les publics et leurs pratiques des réseaux, en limitaient la portée et la diversité.

Exemples de portails

Caractéristiques d'un portail spécialisé en éducation : EduClic ²²

Ce site propose un ensemble bien défini de ressources web (sites ou documents) français, sélectionnés pour leur contenu lié au domaine éducatif. Les sites référencés sont diversifiés tant par leur contenu que par leurs éditeurs : sites institutionnels, académiques, culturels, scientifiques, associations d'enseignants... Ces ressources sont organisées suivant une classification reflétant les divers centres d'intérêt du monde éducatif. Des documentalistes identifient les différentes rubriques des sites sélectionnés et élaborent une présentation synthétique adaptée. Le portail propose en complément des liens vers des sites particuliers en fonction de l'actualité (sujets d'examen, par exemple).

Le portail exploite les technologies de Verity Inc. pour l'indexation en texte intégral des ressources, et Arisem pour leur classification automatique (voir p. 31-32). Des liens organisés vers d'autres ressources extérieures offrent une ouverture vers des gisements d'information complémentaires à ceux proposés. Le portail donne une vision homogène et structurée d'un ensemble ciblé de ressources spécialisées.

Caractéristiques d'un portail spécialisé dans le secteur de la santé : Medisite ²³

Ce site propose :

- des informations sous la forme de notes ou fiches sur les maladies, médicaments ou examens, rédigées par des spécialistes (médecins et / ou documentalistes) ; des références bibliographiques d'articles ou d'ouvrages ; des liens complémentaires ;
- des informations (communication) : messages d'actualité tirés de l'AFP-Santé, des informations sur les problèmes de santé à l'approche des vacances estivales, des quizz pour tester sa santé, les coordonnées d'associations, de maisons de retraite, etc. ;
- des informations à caractère pédagogique (sauvetage, prévention...).

La fonction de recherche en texte intégral s'effectue sur l'ensemble du portail (ici appelé « site ») et sur 6 000 ressources référencées. Ce portail offre aux internautes des fonctions interactives à différents niveaux : participation à des forums, commentaires, paramétrage d'un dossier personnel de santé (avec toutes les réserves d'usage pour ce type de service), paramétrage de requêtes personnalisées qui seront mises en œuvre automatiquement... L'utilisation du portail peut être personnalisée en fonction de son profil : grand public (enfant, adolescent, femme, homme, senior) ou professionnels de la santé.

Medisite fournit, bien sûr, des informations sur son fonctionnement (partenariats) et sa déontologie (charte, fonctionnement vis-à-vis de la publicité).

Définition

À partir de ces deux exemples, nous pouvons proposer les éléments pour une définition de la notion de portail d'information et de communication.

Le portail est une ressource accessible via Internet²⁴, constituant un point d'accès unique, simplifié, facile d'emploi et unifié, pour un public cible, à des ressources (services, produits) électroniques distantes, variées et hétérogènes :

- d'une part, en terme de nature (et de format) : services de recherche d'information (annuaire de sites, moteurs de recherche interne ou externe ; services transactionnels (de type achat / vente, calcul d'un tarifs), impliquant des contraintes de sécurité ; services d'information généralistes (actualités, météo, cours de la bourse...) ou plus spécialisés (exploitation de catalogues de sites, de bases de données interactives ou transactionnelles à des fins de tarification ou de réservation de déplacements...) ; services de communication comme le mail gratuit, les forums, la gestion de fichiers, l'hébergement de sites ou pages, ainsi que des agents intelligents ou des fonctions de personnalisation, d'alerte (*push*), etc. ; services d'accès à des ressources logicielles dédiées ;
- d'autre part, en terme d'origine (émetteurs de la ressource), internes et / ou externes au portail ou à l'entreprise qui le met en œuvre.

Les ressources sont sélectionnées parmi l'offre disponible sur Internet / intranets / extranets ; elles sont organisées et structurées en fonction d'objectifs liés à des usages (et non à des offres), c'est-à-dire intégrées et agrégées (suppression des doublons). Le portail propose des fonctions de personnalisation plus ou moins poussées.

Deux aspects particuliers différencient un « portail » d'une simple page d'accueil de site web (Internet / extranets) :

- d'une part, une approche orientée usage / utilisateur, et non offreur : on s'adresse à une communauté d'utilisateurs avec des besoins spécifiques ;
- d'autre part, un portail suppose une intégration d'applicatifs informatiques : les ressources, issues ou produites par différents outils logiciels (base de données, pages HTML, messagerie...), doivent être utilisées de la façon la plus transparente au niveau de l'interface homme-machine [12] [14] (Voir aussi le chapitre 8).

Cette approche a un triple impact pour les professionnels de l'information et de la documentation :

- une offre de services et produits qui dépasse la mise à disposition stricte d'information documentaire ;
- une offre à valeur ajoutée qui organise et agrège l'information utile. Il s'agit ici de dépasser la juxtaposition d'informations plus ou moins redondantes ;
- une prise en compte de l'utilisateur comme acteur du dispositif avec un accès personnalisé par et pour lui (ou par et pour un groupe d'utilisateurs), voire une participation interactive des utilisateurs – avec une seule identification (*single sign one*).

Cette définition permet d'envisager deux types extrêmes de portail : un portail-seuil assurant une sélection raisonnée des ressources, celles-ci étant entièrement extérieures au portail lui-même, ou au contraire un portail ne fournissant que des services ou ressources produits en interne. Toutes les solutions sont possibles, mais un portail ne se résume pas à une liste de signets ou à un moteur spécialisé ; un portail apporte une valeur ajoutée (sélection, validation, description...) par rapport à un répertoire de sites sur plusieurs points.

– Un portail est un produit éditorial exigeant qui assure une certaine responsabilité vis-à-vis des publics visés et s'impose, par exemple, le contrôle des ressources en cas de changement d'adresse d'un site, ou la fourniture d'informations sur toutes les sources signalées. À l'inverse, un moteur ou un annuaire n'est pas maître des ressources qu'il signale ; il ne peut que signaler leur disparition et rechercher leur nouvelle localisation.

– Des fonctionnalités complémentaires permettent une exploitation plus riche des ressources sélectionnées que leur consultation indépendante et

successive, par exemple en proposant un index thématique commun aux ressources, un glossaire multi-sources, des fonctions de *push*...

– Des services additionnels peuvent être proposés, comme la gestion d'abonnements à des revues spécialisées dans le thème traité par le portail avec un lien avec l'éditeur ; ou, dans le cas d'un portail de revues, la vente à l'article en s'appuyant sur le principe du kiosque ; ou encore un méta-moteur spécialisé sur des ressources externes.

Certaines fonctionnalités, de même que certains services additionnels, peuvent être payants [1].

Portails généralistes et portails spécialisés

Les portails généralistes sont généralement construits autour d'un annuaire ou d'un moteur de recherche généraliste ; ils offrent sur la page d'accueil un certain nombre de fonctions et d'outils complémentaires à cet instrument de recherche, comme des services d'annuaire, de réservation... Cette famille de portails évolue vers une simplification de l'interface qui, malgré des possibilités de paramétrage, restait jusqu'alors très dense, rendant difficile sa manipulation.

Les portails spécialisés ou professionnels, appelés également « vortails », sont des sites fédérateurs à forte valeur ajoutée. Ils visent à créer une dynamique au sein d'un secteur, d'un public particulier, d'une région, etc., en assurant une identification de qualité des ressources propres à un environnement professionnel (une région, une thématique, etc.). Les portails spécialisés sont d'une extrême diversité en fonction des objectifs assignés : fédération d'un secteur d'activité, d'un métier, d'une profession ; fédération des acteurs du marché d'une entreprise, des employés. Globalement, ces critères renvoient à une typologie des usages (vendre, communiquer, former, informer) en rapport avec des groupes d'utilisateurs. Ces critères se retrouvent avec plus ou moins de force dans les portails actuels, en fonction de la culture et des missions des différents acteurs.

Pour faciliter le repérage des portails, tout comme pour les annuaires et les moteurs, des catalogues de portails se développent. Nous pouvons citer Portail 2²¹ pour les ressources francophones.

Outils collaboratifs

Pour tenter d'apporter une réponse aux difficultés rencontrées par les moteurs ou annuaires généralistes face aux volumes et flux des ressources à traiter et à la diversité des usages, et pour fournir une alternative aux méthodes de nature marchande souvent proposées pour y faire face, des

projets et / ou des technologies de nature collaborative se développent. Ces actions s'appuient sur le principe d'une répartition des ressources, humaines ou technologiques, nécessaires au fonctionnement du dispositif. Nous retrouvons là les principes qui furent à l'origine de l'Internet.

Open Directory Project (ODP)

L'Open Directory Project²⁵ est né début 1998 sous la dénomination de Gnuhoo. C'est un annuaire de ressources Internet réparties en catégories et alimenté par les internautes eux-mêmes, la plate-forme technique proposant des outils pour gérer les catégories de façon autonome. Aussi la qualité du contenu de ces catégories est-elle très inégale. Un internaute est donc responsable de sa catégorie, la met à jour, la développe, l'enrichit. Aujourd'hui, l'ODP est l'annuaire par défaut de nombreux moteurs de recherche, dont HotBot ou Altavista.

Anneaux

De nombreux sites sur Internet traitent de sujets proches, voire du même sujet, sous un angle légèrement différent. Ces services se considèrent comme complémentaires, et les liens créés entre eux par les administrateurs en font une communauté de fait, liée par une thématique et répartie sur le réseau.

Un anneau est le nom donné à un type d'association entre sites qui vise à promouvoir et à faire connaître ses membres. Gérée manuellement, la sélection effectuée garantit un certain niveau de qualité et de fiabilité de l'information. Mais, comme pour les forums, il existe une grande disparité entre les anneaux.

Au départ, des sites qui veulent former un groupe échangent leurs adresses. Mais très vite les contraintes de mise à jour de ces adresses et les demandes d'adhésion incitent à la prise en main par un site pivot qui recense et établit la liste des membres. En général, un signe distinctif sur l'une des pages des sites membres permet de repérer l'adhésion à un anneau particulier. L'anneau se distingue d'une page thématique de liens par l'aspect volontaire de l'adhésion.

Ce type d'instrument de recherche exploite la structure hypertextuelle du Web puisque l'on peut naviguer d'un site à un autre grâce à ces liens.

Pour trouver un anneau sur un thème donné, on peut utiliser l'instrument de recherche spécialisé WebRing qui assure une relative visibilité de ces ressources.

Le système WebRing donne un nom (*ring ID* pour identification) suite à une demande d'un site ; ce numéro permet l'identification du site sur le réseau. En parallèle, un logo est mis au point, ainsi qu'un outil (sous forme de tableau) qui permet la navigation entre les sites adhérents. D'autres sites peuvent, à partir de cette première identification, demander leur adhésion ; celle-ci doit être validée par le « ringmestre ». La plate-forme offre également des fonctions pour administrer les anneaux (statistiques de visites, formulaires d'adhésion...).

EXEMPLE

Tout comme les portails, les anneaux peuvent exister sur n'importe quel thème. Une recherche sur la formation à distance sur le site WebRing (*distance learning*) a fourni 26 réponses, dont les 2 suivantes :

FR : Sites de documents JTMATH (4 Sites, 9 Hits ²⁶)

[<http://M.webring.com/hub?ring=frsitesdedocumen>]

Ce groupe répertorie les sites web proposant des ressources pédagogiques réalisées avec le logiciel JTMATH. L'intérêt de ces technologies novatrices est d'apporter une solution originale pour l'enseignement à distance, notamment avec l'aide contextuelle, l'affichage progressif de la solution qui permet une véritable réflexion individuelle pour une meilleure assimilation des notions abordées en cours, etc.

04/04/2002

Distance Learning WebRing (16 Sites, 915 Hits)

[<http://Q.webring.com/hub?ring=dlring>]

Linking sites that offer Distance Learning education course, programs and/or materials for all the world to enjoy.

05/21/1999

Auto-archivage par l'auteur

Dans cette mouvance, ici tout à la fois collaborative et patrimoniale (voir les archives du Web, ci-après), la notion « d'auto-archivage par l'auteur » fait référence à la mise en dépôt public sur Internet (d'où le terme d'« archives ouvertes ») de publications scientifiques par leurs auteurs eux-mêmes. Ces archives peuvent être consultées gratuitement et par tous.

L'auto-archivage peut concerner des pré-publications, c'est-à-dire des pré-tirages non encore soumis au contrôle des pairs, et des tirés-à-part accompagnés de la référence complète de la revue²⁷. Il répond à un fort besoin de

communication, comme l'indiquent les auteurs du projet de création d'une archive ouverte en InfoCom [52] pour favoriser le développement scientifique de cette communauté. Les scientifiques des sciences physiques, mathématiques ou biomédicales ont été les précurseurs de ce type de lieux ressources.

Cette modalité de constitution de collections organisées sur le Web est de même nature que les anneaux, en ce sens que la démarche est volontaire et s'appuie, à priori, sur des aspects qualitatifs.

Archives du Web

Tous les jours, des millions de nouvelles pages sont créées sur le Web, tandis que des millions d'autres sont modifiées ou disparaissent. La question de la pérennité et de la sauvegarde de ces archives est aujourd'hui posée de façon cruciale. De nombreux sites archivent certaines de leurs rubriques comme les entreprises sauvegardent certaines archives. Des moteurs de recherche comme Google proposent, par exemple, la consultation des archives des forums. Mais la question classique est maintenant posée : archiver quoi ? comment ? pour quels usages ? où ?

Parmi les projets en cours, le plus avancé est sans doute celui du fonds américain d'archivage du Web²⁸ ouvert en 2001²⁹. Ce projet se propose de fournir un accès public à un fonds d'archives du Web stocké depuis 1996. Pour cela, ses promoteurs, organisés en association à but non lucratif, Internet Archive, ont constitué un premier fonds à partir de dons : l'histoire du réseau Arpanet, une vaste sélection de sites créés à l'occasion des campagnes présidentielles de 1996 et surtout de 2000 aux États-Unis, les archives des forums de discussion du réseau Usenet, plus de 50 000 bases de données permettant de télécharger des documents de toutes natures, ainsi que 956 courts métrages légués par un collectionneur privé. Le système utilise le moteur Alexa (racheté, depuis, par Amazon). Le site proposait, fin 2001, plus de 10 milliards de pages web (soit 100 tétra-octets) récupérées et archivées, ainsi qu'un outil associé, baptisé Wayback Machine. Celui-ci permet de consulter les sites tels qu'ils existaient lors de leur sauvegarde.

Dans le même registre, la Bibliothèque nationale de France (BNF), qui gère le dépôt légal, a expérimenté l'extension de sa mission aux sites sur le Web. Après avoir effectué la collecte automatique des sites web relatifs aux élections municipales de 2001, la BNF a entrepris en 2002 une collecte automatique d'une image (*snapshot*) du Web français. Ces archives comprendront donc tous les sites relatifs à l'élection présidentielle et aux élections législatives. Les problématiques des archives du Web sont proches de celles de la télévision en raison des contraintes liées au flux de données et à la méthode de l'échantillonnage qui semble la plus appropriée.

En dehors de ces aspects fonctionnels et techniques, le débat est également posé en France sur les rôles respectifs des Archives nationales, de la BNF et de l'Institut national de l'audiovisuel (INA), liés au dépôt légal.

Vivacité du marché des instruments de recherche

Le marché de l'accès à l'information sur les réseaux se complique, avec l'arrivée de nouveaux acteurs et une offre de plus en plus hétérogène. Disparitions (Excite France, Ecila, Excite ou Lokace ont disparu fin 2001), partenariats, changements de nom, nouveaux entrants (des moteurs cartographiques comme Mapstan et Kartoo... ou le moteur français Exalead), évolutions marketing (Northern Light se recentre sur les besoins en technologies des entreprises). Parallèlement, la production d'information électronique s'intensifie et les besoins d'orientation et de repérage se multiplient et se complexifient.

Ce panorama rapide de la situation nous permet de conclure qu'une exploitation optimum des ressources du Web par les professionnels de l'information suppose une veille et une évaluation (voir p. 45) aussi bien des contenus (sources, informations) que des outils disponibles, ainsi qu'une bonne connaissance de leurs apports technologiques et fonctionnels.

Le volume et la diversité de l'information disponible sur les réseaux et la qualité des outils de recherche peuvent être très variables, en fonction du contexte professionnel (tous les secteurs professionnels n'ont pas investi pareillement Internet) et des besoins particuliers à sa mission (territoire linguistique, géographique, niveau de langage...). Aussi convient-il d'effectuer, dans son propre environnement professionnel, une recension et une évaluation des sources existantes et de leur qualité, sur le Web visible et invisible.

Mais la complexité de la situation impose aux spécialistes de l'information d'adopter, pour le repérage et l'évaluation des ressources accessibles via Internet, des méthodes rigoureuses adaptées à ce contexte.

FONCTIONNALITÉS ET TECHNOLOGIES

Introduction : méthodes, techniques et / ou outils

Nous exposons ici les éléments de nature technique³⁰ relevant du domaine de la recherche d'information et exploités par les outils disponibles via Internet (indexation, recherche, exploitation des résultats).

En effet, il nous a semblé important, pour bien positionner les outils de recherche et apprécier les mouvements qui s'opèrent régulièrement sur ce secteur (changement de *crawler*, prise en compte d'un annuaire comme ODP plutôt que d'un autre...), d'effectuer un repérage des techniques proposées récemment et d'identifier leur contexte d'application.

Les techniques mises en œuvre sur le Web sont adaptées à la diversité, aux masses et aux flux de ressources et d'utilisateurs. Certaines des technologies les plus récentes – en particulier celles qui relèvent du traitement automatique des langues ou celles qui sont applicables aux documents multimédias³¹ – restent peu exploitées sur le Web, en comparaison d'espaces plus fermés (portails d'information, extranets, intranets), en raison de contraintes fortes pénalisant les performances et les ressources nécessaires. L'étude des techniques mises en œuvre dans les outils de recherche via Internet [17] [18] [19] peut être réalisée en suivant les différents composants de ces outils :

- le module de collecte automatique des données (*spider*, *crawler*) ;
- les modules d'indexation et de recherche, souvent appelés « moteurs de recherche », gérant les questions des internautes. On trouve parfois les termes « base de données web » ou « index », qui privilégient la structure de l'espace informationnel manipulé, et non l'interface de recherche.

Aujourd'hui, ces outils se sont enrichis de fonctionnalités qui peuvent être étudiées de façon autonome :

- un sous-composant du module de recherche, la présentation des résultats ;
- un nouveau composant en aval de la chaîne : le module d'exploitation des résultats avec les outils de cartographie et de navigation.

Ces deux derniers composants sont réunis dans ce qu'il est convenu d'appeler l'« interface homme-machine pour la recherche d'informations ». Celle-ci, espace de travail de l'utilisateur, fait l'objet de nombreux travaux visant à améliorer l'efficacité de l'activité même de recherche : simplicité croissante (graphique, formulaire, opérateurs implicites, ergonomie de surface améliorant la lisibilité... [53], [54]), articulation de plusieurs modes de recherche (navigation et interrogation directe), et développement d'accès personnalisés (filtres, agents...). Les développements technologiques de ces dernières années concernent l'un ou l'autre des quatre composants précités et, comme le soulignait déjà P. Le Loarer [28] en 1994, « c'est [...] la combinatoire subtile de ces différentes fonctionnalités qui caractérisera la richesse des systèmes de demain ».

Présentation de quelques technologies

Nous présentons quelques-unes de ces technologies ; d'autres sont plus particulièrement exposées dans les chapitres 4 et 7.

Les thèmes traités ici, et présentés par ordre alphabétique, correspondent soit à des techniques à proprement parler (classification automatique), soit à des fonctionnalités (affinement de la requête) orientées utilisateur et renvoyant à plusieurs techniques :

- affinement de la requête ;
- agent intelligent *versus* agent de recherche ;
- classification automatique ;
- documents images et sonores ;
- estimation automatique de pertinence ;
- filtrage ;
- filtrage collaboratif : le point de vue des internautes ;
- infrastructure hypertextuelle du Web : site de référence et site pivot ;
- modèle économique ;
- *peer-to-peer* ;
- présentation des résultats : listes ;
- présentation des résultats : représentation graphique ;
- prise en compte de la polysémie ;
- résumé automatique ;
- tolérance aux fautes d'orthographe ;
- traitement du langage naturel ;
- traitements différenciés suivant la nature de l'information.

Affinement de la requête

Les résultats fournis par les moteurs de recherche sur le Web sont en général très nombreux, même si la requête est précise et formulée correctement. Les moteurs de recherche ont développé différentes techniques pour aider l'utilisateur à affiner sa requête après qu'une première requête a été réalisée.

Globalement, il s'agit de techniques de classification (voir ci-après) :

- soit à priori, à partir des catégories issues d'un annuaire : fonction Fast Topic d'AllTheWeb à partir de l'annuaire Open Directory, catégories de Yahoo! pour Google ; résultats proposés dans la zone Guide Mondial sur Voilà ;
- soit à posteriori.

La sélection d'un *cluster* affine, de fait, la question (voir annexe IV, p. 67-68). Le moteur peut également exploiter des propositions issues de requêtes d'internautes. Altavista constitue une importante base de données des requêtes saisies par les internautes ; lorsque l'on saisit des termes trop génériques, Altavista propose (en anglais) les expressions les plus souvent demandées par les internautes, dans quelque langue que ce soit.

Agent intelligent vs agent de recherche

Selon la norme AFNOR, un agent intelligent est un objet de nature logicielle utilisant les techniques de l'intelligence artificielle pour adapter son comportement à son environnement et pour mémoriser ses expériences. Il se comporte comme un sous-système capable d'apprentissage. Il enrichit le système qu'il utilise en ajoutant, au cours du temps, des fonctions automatiques de traitement, de contrôle, de mémorisation ou de transfert d'information.

L'agent de recherche, quant à lui, interroge à votre place, après paramétrages, des outils de recherche et / ou des ressources identifiées. Ces outils sont utilisés pour des activités de veille (voir chapitre 8).

Classification automatique

La classification automatique suppose une automatisation du processus de répartition des objets dans des classes, un même document pouvant être classé dans une ou plusieurs classes. Ces techniques sont exploitées par certains moteurs de recherche pour organiser le lot de résultats de la recherche, offrant ainsi la possibilité d'affiner ou d'étendre la question.

Deux types de classification automatique sont possibles :

- le classement des éléments dans des classes connues à priori (femme / homme, liste de thèmes...). On parle alors de « classification par apprentissage supervisé ». La qualité du traitement dépend pour partie de l'élaboration de cette liste de classes et de son suivi. Le moteur de recherche Voilà offre ainsi, en recherche approfondie, la possibilité d'effectuer une recherche en la limitant à un ou plusieurs grands domaines préétablis (sport, arts / culture, administration / politique...);
- le regroupement à postériori sur la base de similarités trouvées dans tout ou partie des documents, similarités inconnues au départ (mots eux-mêmes, proximité entre mots...); il s'agit ici de créer des groupes homogènes au sein du corpus résultant de la recherche. Puis l'algorithme affecte les documents aux groupes créés; les documents sont triés par pertinence au sein du groupe. On parle alors de « *clustering* » (classes construites à la volée) ou d'« apprentissage non supervisé ». La difficulté consiste à trouver automatiquement et rapidement les groupes.

Dans ce deuxième type de regroupement, les algorithmes utilisés déterminent une relation de ressemblance floue entre documents. Pour l'élaboration des *clusters*-résultats, ils se fondent sur une fonction de comparaison ou de similitude entre documents par caractéristiques statistico-sémantiques. Une fonction détermine le terme ou l'expression qui donnera son nom au groupe³² (voir annexes IV et V, p. 67-69). Les regroupements peuvent être

hiérarchisés (voir annexe VI, p. 70), ce qui facilite l'exploitation et la navigation au sein de ces *clusters*. Le nombre de classes générées est variable et dépend des relations que l'algorithme trouve au sein du corpus traité. Ce nombre est un indice de la variété du corpus.

Ces techniques de classification automatique sont très utiles dans l'étape de sélection des documents dans un lot de résultats, car elles proposent une information sur les documents trouvés (grâce à l'énoncé des classes thématiques) complémentaire à celle fournie par l'utilisateur lors de la formulation de la requête. Elles facilitent l'élimination des corrélations inintéressantes, évidentes ou déjà connues. Certains regroupements peuvent également donner des idées et des orientations nouvelles.

Ces techniques sont fréquemment mises en œuvre dans des environnements plus restreints sur des intranets et pour des applications de *text mining* (voir chapitre 8), et elles peuvent être complétées par d'autres techniques, en particulier celles de représentation graphique (voir p. 41).

Documents images et sonores

La visibilité sur le Web des fonds images – fixes ou animées (vidéo) – ou sonores est de plus en plus grande ; d'une part les producteurs et créateurs, profitant des améliorations technologiques (débit, qualité) mettent ces ressources à disposition ; d'autre part les moteurs ou annuaires, généralistes ou spécialisés, offrent aux internautes des accès dédiés. Ainsi Google annonce 330 millions d'images... mais la recherche s'effectue essentiellement en anglais, pour l'instant.

Actuellement, les modalités de recherche exploitent les quelques informations textuelles disponibles : la légende, le titre ou le texte de l'URL [43] [46] [48]. Pour pouvoir améliorer l'efficacité de ces recherches, deux axes sont possibles : d'une part la recherche sur un texte descriptif, élaboré manuellement, de ces documents, et d'autre part des traitements automatiques d'analyse de leur contenu image ou son (voir chapitre 7 et [45] [47] [48] [50]). Le premier axe, qui nécessite une description préalable sous forme textuelle des contenus, n'est possible que sur des bases sélectionnées et des fonds relativement maîtrisés ; le deuxième axe suppose des algorithmes appropriés qui sont exposés dans le chapitre 7 du présent ouvrage. D'autres problèmes – juridiques et techniques³³ – freinent la mise à disposition plus large de ces ressources [46].

Estimation automatique de pertinence

L'estimation automatique de la pertinence (également nommée « contrôle de pertinence », « rétroaction de pertinence », « pertinence rétroactive » ou

« remontée de pertinence » ; en anglais, *relevance feedback*) est le nom générique donné à des techniques d'évaluation de la pertinence (*relevance*) des documents retrouvés par rapport à la requête posée.

L'identification des critères pris en compte par les différents moteurs est difficile, le fonctionnement de ces algorithmes restant en général assez opaque [12] [15]³⁴. De plus, les instruments de recherche modifient ces algorithmes et leur combinaison, ce qui rend leur évaluation encore plus difficile.

Ces techniques donnent une valeur aux pages indexées, qui est soit absolue (indépendante des recherches), soit relative (dépendante des recherches). La pondération (calcul d'un poids) et l'appariement entre requête et documents (similarité) que cette mesure autorise permettent alors d'ordonner de façon automatique les documents du lot de résultats.

Similarité et calcul de poids forment la base des algorithmes sur lesquels se fondent les systèmes de recherche d'information. Ces algorithmes traitent soit la requête, soit les documents au sein de la base d'index du moteur, soit le lot de résultats.

Les plus classiques utilisent :

- le poids d'un mot en fonction de sa place dans le document : titre, début du texte (Altavista) ; un poids plus important peut être donné à un mot en majuscule ou en gras (typographie sur Google) à l'intérieur d'un texte, ou à un mot appartenant à une liste de mots contrôlée ;
- le poids d'un mot en fonction de son occurrence dans la base : les mots peu fréquents dans le corpus sont favorisés, les mots « vides » sont soit éliminés soit sous-évalués ;
- la densité d'un mot en fonction de l'occurrence dans le document par rapport à la taille du document. Dans le cas où deux documents ont une occurrence identique pour un même mot, le document le plus petit en taille aura une meilleure pondération ;
- la correspondance de l'expression, la proximité et / ou l'ordre des termes de la requête, dans les documents.

Ces fonctions d'estimation automatique de pertinence s'enrichissent d'algorithmes nouveaux, en complément ou en remplacement des algorithmes plus classiques :

- pondération plus forte pour des pages de référence et des pages pivots (indice de popularité de Google – voir ci-après) ;
- retrait des pages ayant de très grandes fréquences des mêmes mots, considérées comme du *spamming* ;
- pondération plus forte pour des pages cliquées par les internautes (indice de clic) ou des pages sponsorisées.

Filtrage

Dans les interfaces de recherche, un certain nombre de filtres peuvent être mis en œuvre par l'internaute, et sauvegardés pour les sessions suivantes. Ces filtres opèrent soit au moment du traitement de la requête, soit pour paramétrer la présentation des résultats. Ils peuvent porter sur les types de ressources pris en compte : *news*, vidéo, MP3... Pour les informations d'actualité, en particulier, le filtre peut porter sur le type de ressources, mais également sur la date de mise en ligne et les domaines (sport, actualité internationale...). Le moteur Voilà propose de limiter la recherche sur une encyclopédie, Hotbot sur des pages personnelles.

Les autres critères classiquement proposés sont :

- le territoire régional ou linguistique (Web mondial ou francophone, etc.) ;
- la langue de la ressource ou le nom de domaine du serveur ;
- les dates ou périodes de mise à jour des ressources ;
- le nombre de pages affichées pour un même site ;
- le choix des champs à explorer, à afficher ;
- la prise en compte ou non d'une correction orthographique automatique³⁵ ;
- le nombre de réponses par pages, etc.

Ces fonctions de filtrage sont proposées soit sur l'écran standard (HotBot), soit, plus fréquemment, dans un écran de recherche approfondie (AllTheWeb, Altavista).

Filtrage collaboratif : le point de vue des internautes

La particularité du filtrage collaboratif, contrairement à la recherche d'information, est de se baser sur les évaluations des documents faites par les utilisateurs³⁶, en ignorant partiellement ou totalement les techniques automatiques sur les contenus. L'idée d'exploiter un système de filtrage dit « collaboratif » fait profiter une communauté d'utilisateurs des efforts d'évaluation réalisés par ses membres, en signalant les documents jugés intéressants par certains à d'autres utilisateurs dont les intérêts exprimés par la question sont proches.

Par exemple, la technologie WPS, utilisée par l'outil MapStan Search exploite des algorithmes de filtrage collaboratif pour constituer le lot de résultats en associant les résultats proposés par un moteur classique et ceux proposés par MapStan Search lui-même, se basant sur les résultats de requêtes d'autres internautes.

Sur d'autres systèmes³⁷, l'avis demandé aux internautes de façon interactive sur les sites enrichissent le calcul du poids de ces sites.

**Infrastructure hypertextuelle du Web :
site de référence et site pivot**

Les techniques déployées actuellement pour l'indexation, la recherche et l'exploitation des résultats s'appuient beaucoup sur des méthodes et des pratiques développées dans des dispositifs de recherche de références organisées en base de données. La prise en compte des potentialités et particularités du Web en tant que réseau de liens est relativement récente et a nécessité de nombreuses recherches et expérimentations.

Sur le Web, les objets documentaires ne sont pas considérés de façon autonome les uns par rapport aux autres, mais au sein d'un graphe dont les sommets sont les pages HTML et les arêtes les liens. Certaines approches s'appuient sur le principe suivant : un lien de parenté sémantique existe entre deux pages reliées [18] [19]. Les expérimentations menées vérifient cette hypothèse, les liens d'une autre nature comme les « renvois à la page d'accueil » ou les liens publicitaires ne remettent pas en cause ce principe de base.

On peut distinguer schématiquement quatre types de fonctionnalités.

EXPLOITATION DES LIENS ÉTABLIS PAR LES AUTEURS ET / OU ÉDITEURS

• *Pour la recherche d'information*

Le contenu des liens (URL), voire le contenu des pages liées, sont pris en compte lors de l'indexation (élargissement des zones indexées) et dans les critères de calcul de pertinence des résultats.

Deux types de pages sont pondérées plus fortement :

- celles constituées de nombreux liens (par exemple, un répertoire de signets), appelées « pages pivots » ;
- celles vers lesquelles pointent de nombreux liens de qualité en provenance d'autres pages, appelées « pages de référence » (ou *backlinks*, liens à l'arrivée).

Un deuxième principe intervient ici, déjà exploité dans d'autres environnements professionnels en veille avec des méthodes scientométriques et de *text mining* [13] : des pages de référence sont citées fréquemment par des sites « pivots », renforçant leur pondération.

Google a été le premier des moteurs de recherche commerciaux à prendre en compte cette fonctionnalité nommée « indice de popularité », qui mesure l'importance d'une page au sein de sa base, par la technologie PageRank, puis de leur pertinence par rapport aux termes de la requête. Avec ces techniques (indice de popularité au sein du corpus et adéquation par rapport

à la question), une recherche comportant des noms de marque comme mots clés sera orientée préférentiellement sur les sites officiels : sur Google France, le mot clé « éducation » (avec accent) vous renvoie, avec la fonction « J'ai de la chance », sur le site de l'Éducation nationale (.gouv.fr).

Le système analyse donc les liens croisés entre les documents de la liste obtenus en réponse à la question : plus un site est référencé par les autres, meilleure sera sa position dans la liste réponse.

Ce principe évite le *spamming*, mais pénalise les ressources nouvelles.

- *Pour la navigation dans les résultats*

Les représentations cartographiques des résultats des moteurs de recherche et les cartes de connaissances rendent également compte de ces progrès (voir chapitre 6). Il s'agit ici d'établir une représentation graphique à la place d'une représentation linéaire des résultats de la recherche, à partir de liens établis dynamiquement entre les documents.

EXPLOITATION DES ACTIVITÉS DES UTILISATEURS

- *Exploiter des liens à partir d'un intérêt exprimé par les utilisateurs*

L'intérêt de l'internaute est évalué à travers des actions soit automatiques – clic (indice de clic, CPC), affichage, utilisation de mots clés de la question (voir également « Filtrage collaboratif », *supra*) –, soit manuelles (notation de la part de l'internaute).

- *Exploiter les chemins (adresses) parcourus*

Une autre méthode consiste à conserver les parcours des utilisateurs, et à les proposer à ceux qui ont des préoccupations ou des besoins communs. C'est ainsi que l'on peut trouver des agents cartographes ou des historiques de chemins parcourus.

La mesure d'audience par DirectHit, exploitée par de nombreux moteurs, comme HotBot

Le principe consiste à pondérer les pages en fonction du nombre de visites reçues. Le système analyse le comportement d'un internaute (pages visitées, temps passé, parcours) lors de l'utilisation d'un moteur de recherche en suivant son parcours, et établit les pages les plus « populaires ». Ces systèmes fonctionnent en règle générale en tâche de fond sur un moteur existant.

Comme l'indice de popularité, cette méthode pénalise les ressources récentes mais contourne les effets du *spamming*.

Modèle économique

Depuis les difficultés économiques rencontrées pendant l'année 2000 par de nombreux acteurs du marché de l'Internet, les instruments de recherche multiplient les solutions susceptibles de constituer des nouvelles sources de revenu. Le modèle économique basé sur une composante publicitaire et / ou promotionnelle forte s'est renforcé et enrichi, et l'on assiste au passage du tout gratuit vers le payant.

MOYENS PROMOTIONNELS

Les moyens promotionnels proposés actuellement concernent :

- l'affichage d'une bannière publicitaire lors de la sélection d'une rubrique ou du choix d'un mot clé ;
- le soumissionnement payant, garantissant ou non le référencement. Ce type de soumission permet de référencer des ressources HTML, mais également des pages dynamiques en ASP, PHP ou en Flash (offre d'Inktomi, par exemple). Dans le cas où le référencement est validé par l'instrument de recherche, la soumission payante garantit la présence d'un certain nombre de pages d'un site dans la base de données du moteur de recherche, ainsi qu'un rafraîchissement des documents dans des délais courts et garantis ; mais le positionnement dans les résultats n'est, quant à lui, pas garanti. Cette procédure, qui est obligatoire pour certains moteurs (HotBotUS, par exemple), permet également de limiter le *spam* qui s'était accentué avec la soumission en ligne ;
- le positionnement payant dans le lot de résultats, qui s'effectue en fonction des mots clés de la question, se décline de deux façons :
 - certaines offres garantissent une présence en tête des résultats de recherche, ou en tête d'une catégorie ou rubrique. Cette démarche est souvent accompagnée d'un signe distinctif, mais elle peut également être difficile à distinguer (Tiscali) pour l'internaute,
 - d'autres offres, appelées « liens sponsorisés » ou « liens commerciaux », garantissent une présence dans la page de résultats en fonction des mots clés de la question. Ces liens sont placés dans une zone déterminée et visible de l'écran, ou en début de liste. En moyenne, le nombre de ces liens sponsorisés est de 3, mais il est de 2 pour Google, 5 pour Yahoo! et... 20 pour Nomade. En juin 2002, seules 20 % des requêtes proposent des liens promotionnels³⁸.

Le principe de ces liens sponsorisés est le suivant : des sociétés comme Spotting ou Overture proposent un système de vente aux enchères de mots clés ; le site ayant eu la plus forte enchère à un moment donné apparaîtra en première position sur la liste des résultats des instruments de recherche qui auront adopté cette formule. En règle générale, ce sont les premiers résultats qui sont ainsi affichés sur les pages de résultats. Google, avec son offre

Adwords, associe pour un même mot clé l'enchère proposée, mais également le taux de clics sur le lien en question. Les versions américaine, canadienne, anglaise et allemande d'Altavista exploitent les liens d'Overture avec trois liens présentés en début des réponses du moteur, introduits par la mention Sponsored Listings ou par Products and Services. Sur Altavista France, ce sont deux liens d'Espotting France.

On peut noter également la solution proposée par le moteur SearchEngine Colossus [68] avec la technologie Rolist³⁹. Le moteur de recherche utilise le principe d'une liste tournante pour classer en premier les sites qui se sont affiliés au système.

Chaque fois qu'un nouveau visiteur effectue une recherche sur un mot ou une phrase, le système Rolist effectue une rotation de la liste des sites enregistrés, plaçant à tour de rôle les sites affiliés ayant inscrit, dans les métadonnées, le mot clé recherché. Tous les membres peuvent ainsi espérer être au moins une fois en tête de liste ! Les membres de Rolist contrôlent la validité des mots clés que le postulant a proposés lors de son autoréférencement. Les résultats fournis par la Rolist, placés en tête de liste, sont distingués clairement des autres résultats.

MODE DE FINANCEMENT

Le principe de la gratuité, aussi bien pour les utilisateurs que pour les producteurs d'information référencés, est fortement remis en question. Les modes de financement et de rémunération sont multiples : au nombre de clics (CPC), au nombre d'affichage (CPM), aux enchères, au forfait...

De ce point de vue, les instruments de recherche déploient des pratiques très variées : partenariats, subventions, bandeaux publicitaires, vente en ligne, abonnements, location de fichiers... Certains moteurs (Yahoo!, par exemple) adoptent le principe d'une rémunération du référencement par abonnement annuel.

Il est important de prendre en compte, dans l'évaluation des instruments de recherche, ces techniques qui peuvent bien évidemment perturber les résultats. On note par exemple que la soumission non payante ne garantit aucunement une présence dans les index, et de fait rallonge les délais de référencement. Les ressources ne pouvant ou ne voulant envisager ces solutions doivent faire confiance au travail de veille⁴⁰ et de référencement effectué par les documentalistes des moteurs et annuaires.

Un autre moyen de rémunération des outils de recherche réside dans la location de fichiers des annuaires aux moteurs de recherche, par exemple Yahoo! est exploité par Google, et Indexa (sites professionnels français) par Yahoo!.

Peer-to-peer

Le *peer-to-peer*⁴¹, littéralement « de pair à pair », est un système d'échange direct de ressources entre machines interconnectées popularisé en 2000 par Napster, un site web américain qui permettait de télécharger des fichiers musicaux MP3 provenant de millions d'ordinateurs connectés. Il constitue une technologie intéressante du point de vue économique puisque, dans une architecture *peer-to-peer*, les ordinateurs peuvent agir à la fois comme clients et comme serveurs (donc comme des pairs). Ce fonctionnement permet l'échange de fichiers ou de programmes, voire la mise en commun de la puissance de calcul des machines, etc. Le premier réseau de réseaux, Arpanet, fonctionnait sur ce principe ; il permettait de décentraliser les contenus, et d'alléger l'administration des dispositifs. Les nouvelles applications logicielles *peer-to-peer*, comme Gnutella, se distinguent de celle mise en œuvre sur Napster ; elles exploitent totalement l'infrastructure réseau et intègrent les contraintes actuelles (pare-feu, adresses dynamiques...).

Aujourd'hui, le modèle de réseau du *peer-to-peer* est remis en cause pour des raisons légales, économiques et techniques (problèmes de contrôle des données échangées, résultats différents en fonction de l'état de connexion des machines).

Présentation des résultats : listes

L'organisation et le contenu des pages de résultats, sous forme de listes, des moteurs et annuaires se sont enrichis.

LA STRUCTURE DE LA PAGE

La page de résultats présente différents types d'information dans des zones distinctes : rappel de la requête avec plus ou moins de précision sur ses composantes (HotBot), liste des résultats, catégories issues d'un annuaire complémentaire⁴², autres outils complémentaires. La zone résultat est elle-même découpée en plusieurs sous-zones distinctes en fonction de la nature du tri résultat : résultats « naturels », liens sponsorisés, type d'information (vidéo, son...).

LA REPRÉSENTATION DE CHAQUE DOCUMENT

Cette représentation du document doit être claire, précise, non ambiguë, suffisante pour permettre d'apprécier l'intérêt d'un document et ne pas imposer, à cette étape de sélection, la lecture du document pour effectuer un

tri. On peut regretter, par exemple, que la date – qui, même si elle ne correspond qu'à la date d'indexation, reste un critère de sélection important – ne soit pas proposée sur Google. Par contre, Google fournit les catégories de son annuaire correspondant à la recherche.

La présentation des références obtenues par traitements publicitaires, lorsqu'elles ne se distinguent pas de celles obtenues automatiquement, constitue un point négatif de certains instruments de recherche.

Google

Portail 2, le portail des portails : accueil

Mesurez votre audience. **Portail 2. Portail 2, le premier portail facile.....** Les news: Actualité. **Portail 2 est votre portail.** Nous...

Description: Un **portail** des portails pour faciliter les recherches pour débutants et professionnels.

Catégorie: World > Français > Informatique > Internet > Portails

www.portail2.com/ - 21k - En cache - Pages similaires

HotBot

17. Portail 2

Un **portail** des **portails** pour faciliter les recherches pour débutants et professionnels.

URL: <http://www.portail2.com/>

Date: 2002/04/28 - Taille du fichier: 20k - Nom de domaine: www.portail2.com

[ouvrir dans une nouvelle fenêtre](#) | [voir les résultats de ce site uniquement](#) | [envoyer à un ami](#)

FONCTIONS POUR POURSUIVRE LA RECHERCHE

Les fonctionnalités permettant d'affiner ou d'étendre la recherche sont associées à la référence elle-même, comme le montrent les exemples précédents, pour en faciliter l'usage.

Présentation des résultats : représentation graphique

Ces techniques de représentation graphique de l'information existent depuis de nombreuses années, mais leur exploitation était freinée par les exigences en ressources machine que supposaient leur mise en œuvre, et également en raison de difficultés liées à leur usage.

Les années 2000 ont vu apparaître des métamoteurs généralistes sur le Web, offrant des représentations graphiques des résultats de recherche.

Kartoo (voir annexe III, p. 66-67) présente les résultats de la recherche par une carte dont les nœuds représentent un document et les liens, des relations entre ces documents. La taille des nœuds est plus ou moins grande en fonction du degré de pertinence du document par rapport à la question. Celle-ci était jusque-là calculée suivant l'indice de popularité (nombre de liens pointant vers lui) ; aujourd'hui il semble que ce principe ait été modifié. Plusieurs cartes sont nécessaires pour visualiser les documents résultant de la question ; le nombre de documents par cartes est paramétrable. Des fonctions spécifiques permettent de se déplacer dans une carte (haut, bas, gauche, droit ; zoom) et entre les cartes ; le passage du curseur sur un nœud permet de visualiser les éléments d'information sur le document.

Le choix de l'outil Mapstan [66] s'est porté, quant à lui, sur une représentation synthétique de type « plan de quartier » : les places représentent les documents, les rues représentent les relations entre documents. Utilisant des fonctions de catégorisation, les nœuds peuvent représenter un ou plusieurs sites (*cluster*). Cet outil, doté d'une iconographie particulière, évoque un tableau de bord de navigation d'un corpus de documents. Actuellement, cette technologie est proposée en s'appuyant sur les résultats du moteur de recherche de Google. On peut citer également Malaspina⁴³ ou Vivisimo, qui ont opté pour une représentation en arborescence.

Prise en compte de la polysémie

MÉTHODES STATISTIQUES

Les moteurs de recherche prennent en compte certains phénomènes de polysémie, ceux-ci pénalisant fortement les recherches, par différents moyens (analyse automatique des résultats de la recherche, *feedback* thématique...), en exploitant des algorithmes de nature statistique (voir le chapitre 4).

Les techniques de catégorisation permettent, également, de lever l'ambiguïté due à la polysémie de certains termes, en les remettant dans leur contexte (voir p. 32).

REALNAMES : LISTE DE PERSONNES MORALES OU DE MARQUES

Il est délicat de parler d'une fonctionnalité (plutôt que d'une technologie) qui vient d'être supprimée, mais le principe de base des *Internet keywords* de la société RealNames semble intéressant à clarifier.

Les *Internet keywords* sont des mots, et non des noms de domaines, assignés à des sites particuliers. Par exemple, le mot Renault était associé au site du groupe Renault. Le principe était assez proche de celui des marques déposées. La base de données, gérée par la société RealNames, contenait des objets de type nom propre (société, produit...), décrits par un nom (*Internet keyword*), une adresse (URL), la langue dans laquelle était écrite la page web en question et quelques mots clés. Par exemple :

Internet keyword: Renault Scénic
URL: <http://scenic.renault.fr/>
Language: français

Chaque terme de la requête posée par un internaute était comparé à cette liste. Dans le cas où le système trouvait une réponse dans la base RealNames, le site associé obtenait une pondération maximale ; sinon, l'internaute était redirigé directement vers le site approprié. Cette fonctionnalité était implémentée sur le navigateur Internet Explorer et certains moteurs de recherche. Le modèle économique reposait sur l'achat de mots clés par une société, et le partage des gains entre RealNames et le moteur de recherche. (La société RealNames a annoncé son retrait en juin 2002⁴⁴, après que Microsoft eut renoncé à son partenariat.)

Résumé automatique

Les quelques lignes présentées par les moteurs de recherche lors de la présentation des résultats correspondent soit à la technique d'extraction d'une partie de phrase dans laquelle se trouvent les mots recherchés, soit aux premières lignes du document. On ne peut donc pas véritablement parler de « résumé » pour ce qui les concerne, puisqu'il n'y a pas reformulation. Mais c'est le terme aujourd'hui utilisé pour l'offre sur Internet.

Le résumé automatique permet de fournir une représentation synthétique du contenu d'un document. Les technologies employées identifient et extraient les concepts clés d'un texte pour en générer un « résumé » composé des phrases les plus marquantes du document original. La longueur du résumé peut, dans le cas de Copernic Summarizer⁴⁵, être modifiée en temps réel. Copernic Summarizer traite des textes en français, en anglais, en espagnol ou en allemand.

Actuellement l'offre de la société Pertinence⁴⁶ permet de résumer, via Internet, des textes en français au format MsWord (.doc, .rtf), texte, HTML ou PDF. Le principe est le suivant : des tournures linguistiques sont étiquetées et affectées de valeurs dépendant de leur capacité d'extraction informationnelle, par le biais de l'analyse du discours. Puis les phrases dont le poids est nul ou très faible sont supprimées (« résumé maximal »). Il est possible de choisir un indice de compression. Les phrases ainsi extraites ne sont pas

reformulées ; la cohésion est donc incomplète, ce qui est parfois préjudiciable à la compréhension, mais reste satisfaisante en terme de recherche informationnelle. Des améliorations peuvent être apportées de ce point de vue par l'aspect visuel, en présentant dans son contexte la phrase extraite.

Tolérance aux fautes d'orthographe

Les statistiques montrent que de nombreuses requêtes contiennent des erreurs orthographiques qui réduisent fortement la qualité des résultats. Or, en raison des importants volumes qu'ils manipulent et des techniques mises en œuvre, les moteurs généralistes (et même spécialisés) proposent toujours des documents en réponse (il y aura toujours un document qui contiendra l'un des termes demandés), même si la requête comporte une erreur, ce qui laisse l'internaute dans l'ignorance de celle-ci. Dans une requête ne comportant pas ces erreurs, ces documents se seraient trouvés à une place éloignée de la liste des résultats, alors même qu'ils apparaissent ici en début de liste. Des algorithmes intégrant la notion d'orthographe approximative peuvent être mis en œuvre pour pallier ce problème, en extrayant de l'index du moteur les termes proches.

Il est possible de paramétrer le niveau de tolérance sur l'imprécision : selon la souplesse souhaitée, l'outil tolérera la recherche partielle ou n'admettra que l'omission d'une lettre ou son remplacement par une autre. On peut, à titre d'exemple, citer l'agent de recherche d'AntiSearch⁴⁷ sur un annuaire (AS@Mail).

Traitement du langage naturel

Le principe de l'indexation automatique repose sur les postulats suivants : les mots du texte sont des indices formels de son contenu, et l'exploitation intelligente de ces indices par un outil peut remplacer l'indexation manuelle et permettre la recherche. Mais l'indexation automatique accorde à tous les mots la même importance, générant dans le cas du Web beaucoup d'insatisfaction pour des recherches de type thématique, principalement en raison des phénomènes de polysémie [9, p. 161]. Dans ce type de recherche en texte intégral, on utilise, pour formuler la requête, des règles difficiles à appliquer efficacement par un utilisateur non professionnel de l'information (voir annexe II, p. 65).

Diverses fonctionnalités et techniques sont aujourd'hui mises en œuvre pour limiter les biais et les effets négatifs de cette technologie (voir *infra*), sans toutefois répondre à tous les problèmes que pose la recherche. Aussi est-il intéressant d'utiliser des techniques de traitement automatique du langage (voir chapitre 4) dans les systèmes de recherche d'information. Sur

le Web, seuls AskJeeves⁴⁸ et Infoclic⁴⁹ (basé sur le moteur en langage naturel de la société Sinequa) ont aujourd'hui intégré ces technologies. Par contre, sur un espace plus restreint, de nombreux sites proposent une interface en langage naturel pour l'exploitation de tout ou partie de leur contenu⁵⁰ [21]. Ces techniques supposent un travail sur les ressources linguistiques qui seront exploitées par le moteur. Ces activités s'apparentent aux activités liées à la constitution et à la maintenance des langages documentaires ; elles sont rendues complexes par l'exploitation pas toujours maîtrisable qui pourrait en être faite sur le Web !

Traitements différenciés suivant la nature de l'information

Pour améliorer la pertinence des résultats sur les instruments de recherche, la nature hétérogène des ressources disponibles, en termes de formats de fichiers ou de contenu, peut être mieux prise en compte. Établir des distinctions dans les traitements à opérer, comme la réduction des délais de traitement pour les documents d'actualité⁵¹ ou la prise en compte des accents pour des documents reconnus dans leur version linguistique, constitue un choix pragmatique et devrait devenir un axe d'amélioration efficace dans l'avenir.

En conclusion

Les techniques et fonctionnalités développées tout au long des années 1990 s'enrichissent et se déploient sur la majorité des instruments de recherche. Mais elles font référence à des pratiques (éditoriales, techniques, d'usage) que l'on pourrait qualifier de « classiques », la prise en compte des liens n'étant encore bien souvent considérée que comme utile à la navigation à postériori, et non comme une clé d'accès initiale.

Le développement des concepts d'XML et du Web sémantique, ainsi que les possibilités fonctionnelles et ergonomiques offertes par les cartes sémantiques devraient modifier plus en profondeur le paysage des outils de recherche dans les années à venir.

ÉVALUER LES SOURCES D'INFORMATION ET LES OUTILS DE RECHERCHE

Introduction

Toute recherche d'information suppose l'existence de stratégies de recherche adaptées à la problématique posée, au demandeur ainsi qu'aux ressources exploitables. Ces activités supposent des pratiques d'évaluation :

pour repérer et évaluer les ressources et les outils susceptibles d'être utilisés, les pratiques et les besoins des utilisateurs, les activités mêmes de recherche d'information, etc. Les résultats d'une surveillance sur ce thème montrent le fort intérêt suscité par les méthodes et les actions d'évaluation. Mais force est de constater que, dans la pratique professionnelle, ces méthodes et même les principes d'évaluation et de jugement sont moins clairement établis, si ce n'est négligés, faute de temps, d'intérêt ou de formation.

De façon générale, la notion de pertinence renvoie à des critères de jugement de valeur. Même si ces jugements sont relatifs à certains types de besoins ou d'attentes, leur énonciation fait appel à un être humain ; aussi ces notions doivent-elles être manipulées avec beaucoup de prudence. Par exemple, de nombreuses expériences d'évaluation ont montré qu'une information jugée par un expert de qualité médiocre pouvait être évaluée comme pertinente par un utilisateur, c'est-à-dire utile et utilisable dans son contexte de travail.

Les finalités peuvent être diverses : évalue-t-on la performance technique des outils (moteurs) ? la qualité de l'information ? dans un contexte précis ? la réponse à des besoins ? les facilités d'utilisation et d'exploitation, etc. ?

Ces notions méritent quelques éclaircissements que nous nous proposons d'aborder ici.

Notion de pertinence en recherche d'information⁵² ?

La pertinence, que l'on peut schématiquement définir comme un degré de corrélation entre une question et la réponse apportée, est un concept clé en recherche d'information depuis les années 1950. Depuis, cette notion a été traitée sous plusieurs aspects, comme le montre [33] dans son étude complète sur ce sujet. Ces aspects sont essentiellement :

- les documents, leur représentation à travers un substitut (notice, méta-données) et les éléments informatifs des documents (contenu) ;
- le problème à résoudre, le (ou les) besoin(s) d'information qui en découle(nt), la demande représentant le problème posé par l'utilisateur, et la requête (*query*) représentant le besoin informationnel dans le langage système ;
- le facteur temps, qui caractérise les évolutions pouvant se produire entre le moment de l'expression du problème initial et sa résolution.

La mise en relation de ces ensembles montre la diversité de la notion de pertinence : pertinence de la représentation du document par rapport à la requête, pertinence d'un document trouvé par rapport à une requête⁵³, pertinence d'une information reçue par rapport à un besoin exprimé... Cela explique pourquoi « un document retenu comme conforme à la demande (*relevant*) peut ne pas être pertinent pour l'utilisateur ».

D'autres éléments sont intrinsèquement contenus dans la définition de la notion de pertinence : le thème ou sujet⁵⁴ (*topic*) qui revêt, pour les professionnels de l'information et de la documentation, une valeur particulière ; la tâche visée, qui correspond à l'activité pour laquelle l'utilisateur éprouve un besoin d'information ; et le contexte, qui affecte la méthode de recherche ainsi que l'évaluation des résultats.

On peut dire que, classiquement, les systèmes de recherche d'information majoritairement informatisés et les outils d'évaluation se concentrent essentiellement sur le composant « thème » (valeur intrinsèque du document), peu sur la tâche, et encore moins sur le contexte. Or, une réponse peut-elle être jugée pertinente si elle n'est d'aucune utilité pour la tâche à exécuter, ou si elle est déjà connue (contexte) ?

De plus, la majorité des études font référence à un système généralement informatisé et associant, sans toujours les distinguer, la pertinence du contenu et la pertinence de l'outil de manipulation – à savoir le logiciel.

Cette typologie, proposée par [33] puis reprise par [26], montre la distinction entre la pertinence vue sous l'angle du système de recherche (pertinence-système), la pertinence vue sous l'angle de l'utilisateur, son besoin, son contexte (pertinence-utilisateur), et la pertinence du point de vue du thème (pertinence-sujet).

Nous présentons ci-après, de façon succincte, quelques exemples de types de pertinence.

Pertinence-système

PERTINENCE DES MESURES DE CLASSEMENT

Cette pertinence mesure la capacité qu'a le système de retrouver et de classer des documents en réponse à une requête. La mesure de la pertinence s'appuie fréquemment sur les mesures de rappel (documents pertinents récupérés, par rapport à l'ensemble des documents pertinents du fonds) et de précision (documents pertinents récupérés par rapport à l'ensemble des documents récupérés), ce qui pose un problème important dans le cas du Web, où le rappel est difficile à évaluer.

PERTINENCE DE L'INDEXATION AUTOMATIQUE

La pertinence de l'indexation automatique pour une visée de recherche (voir p. 33) est probablement celle qui a fait l'objet du plus grand nombre d'études. Les protocoles d'évaluation du programme TREC (*Text Retrieval Conference* [41]) peuvent être assimilés à cette catégorie.

Pertinence-utilisateur

PERTINENCE DE LA FORMULATION DE LA REQUÊTE

Ce type de pertinence orientée utilisateur fait appel à deux types d'analyse de pertinence : d'une part, celle de l'analyse du problème posé et de sa traduction claire et précise par l'utilisateur en question ; d'autre part, la qualité de la formulation de la requête à travers l'interface du système, c'est-à-dire la traduction de la question en requête qui sera traitée par le système. Pour ce deuxième type d'analyse, les évolutions fonctionnelles et ergonomiques des interfaces de recherche tentent de faciliter la formulation et la reformulation de la question par l'utilisateur. Comme exemple, on peut citer les interfaces standard des moteurs de recherche, qui restent essentiellement adaptées aux pratiques d'un public non professionnel de la recherche d'information, et qui peuvent être considérées, par ce public, comme pertinentes selon ce critère.

L'exploitation des fichiers .log (impossible pour les recherches via Internet) est d'une grande aide pour réaliser ce type d'évaluation [21]. L'évaluation des interfaces exploite des grilles qui ont fait l'objet de travaux de normalisation [54] (et d'évaluation des grilles elles-mêmes !) qui pourraient être plus fréquemment utilisées par les professionnels de l'information et de la documentation.

PERTINENCE DE LA PRÉSENTATION DES RÉSULTATS

La pertinence de la présentation des résultats peut s'évaluer à partir de plusieurs indicateurs, tels les moyens fournis pour apprécier l'adéquation des résultats à la question, ou encore les moyens pour juger de l'opportunité ou non de consulter l'intégralité de la ressource (étape de sélection de l'information). Elle doit donc fournir des indications sur chacun des documents et sur la valeur de chaque document par rapport à l'ensemble des documents proposés. Ce dernier point est particulièrement pris en compte dans les interfaces cartographiques (Kartoo ou Vivisimo) et avec les techniques de clustérisation du lot de résultats. Pour ces méthodes, des critères de mesure de cette pertinence restent à construire.

PERTINENCE DU DOCUMENT PAR RAPPORT AU BESOIN

De nombreux paramètres liés aux utilisateurs entrent ici en ligne de compte. En particulier, le critère d'utilité immédiate peut être estimé prépondérant, plutôt que celui de la précision ou de l'exhaustivité informationnelle, voire de la crédibilité de la source.

Cette évaluation se mesure à partir d'enquêtes utilisateurs par questionnaire, observation ou entretien.

Pertinence-thème (sujet)

PERTINENCE LIÉE À LA VALEUR DES DOCUMENTS

Il s'agit d'évaluer le document hors contexte pour sa pertinence par rapport au thème. Mais l'absence de chaîne éditoriale sur le Web⁵⁵ a laissé la place au développement de types particuliers d'évaluation de la valeur des ressources du point de vue du thème, comme « l'indice de la popularité »⁵⁶ (voir p. 36-37), en plus des valeurs attribuées à posteriori (après leur insertion dans la base d'index) par un professionnel de l'information ou un expert du domaine. Certains de ces critères, de par la variabilité des jugements de pertinence portés par des individus et surtout l'absence de méthodes et outils normalisés, sont bien sûr à utiliser avec prudence.

PERTINENCE DU CHOIX DE L'INSTRUMENT DE RECHERCHE

Ce critère s'intègre dans les « bonnes pratiques » liées à une démarche de recherche documentaire. Quel instrument de recherche (base, moteur généraliste ou spécialisé, portail...) exploiter pour tel sujet ? Les grilles d'évaluation de ressources web mises en place par de nombreux professionnels de l'information entrent dans cette catégorie.

Grilles d'évaluation des outils de recherche

L'évaluation suppose la mise en place de méthodes et outils spécifiques qui se déclinent différemment, mais complémentaires, s'il s'agit d'évaluer une ressource d'information ou un moteur d'indexation.

Mais les évolutions fonctionnelles proposées dans les systèmes de recherche d'information, en particulier les instruments spécialisés, ont abouti à une imbrication de plus en plus forte entre contenu et fonctionnalités de manipulation de ce contenu, imposant le développement de grilles complètes et complexes.

Aujourd'hui, plusieurs tendances se dégagent :

– des **répertoires de ressources web**, généralistes ou spécialisés. Ceux-ci couvrent des champs très divers et intègrent des descriptifs plus ou moins signalétiques, plus ou moins précis. Ces outils permettent aux usagers de s'orienter vers des ressources. Les descriptions ou analyses ne suivent pas toujours de trame particulière, et l'énonciation de la qualité est très variable (« riche, mais on s'y perd souvent », « intéressant »...). Souvent, l'irrégularité du contenu de l'évaluation et l'hétérogénéité de leur rédaction rendent l'exploitation de ces outils difficile dans un autre environnement. Ces outils valident la pertinence-sujet, et parfois certains aspects de la pertinence-système ;

– d'autres **guides** (voir [37], [38], [39], [40], [41], [42]) font un travail plus poussé, établissant des notations sur les critères d'évaluation, permettant des études comparatives. La méthode consiste à évaluer la qualité des ressources par le biais d'un score. Certains de ces outils sont établis en intégrant des usagers lors de la phase de conception, voire pour l'évaluation elle-même (pertinence-utilisateur). Ces grilles d'évaluation s'accompagnent d'un manuel méthodologique pour assurer la qualité de la mesure établie par des évaluateurs différents. Ils peuvent être utilisés pour établir des répertoires de ressources ou de moteurs.

Exemples de grilles d'évaluation

Exemple 1 - Critères de classement des moteurs de recherche [38]

Marc Duval. (Page créée le 28 juillet 2001). Classement des automates de recherche. [En ligne]. Longueuil. Québec, ©2001.

<http://www.dsi-info.ca/moteurs-de-recherche/classement-requetes-criteres.html>

Classement d'une vingtaine de moteurs de recherche (Google, Altavista, MSN...)

Le site contient : le protocole de recherche, les outils pour l'évaluation (formulaires utilisés, grille d'évaluation des formulaires de recherche) ainsi que la liste, les critères et la grille d'évaluation des requêtes, puis le classement des sites de recherche.

Onze critères ont été retenus pour la notation : le bruit, le champ sémantique (mot correspondant à la requête mais non pertinent à la question), le code d'erreur, le doublon, la dysfonction (dysfonctionnement) du logiciel de recherche, la non-correspondance entre le terme recherché et la page de référence, le rang du document pertinent, le silence et l'adresse URL inactive.

Protocole de recherche :

- Requêtes administrées à partir de formulaires en français ou en anglais.
- Champ langue utilisé systématiquement pour toutes les recherches.
- Requêtes posées en français avec accents et majuscules.
- Recherche sur 12 sujets, chacun ayant 2 questions.
- Difficultés : combinaison d'un mot et d'un nombre, majuscules, accents, homographes, champ domaine, opérateur de proximité immédiate « " " » et recherche d'images. Pour la recherche d'images, une autre difficulté a été ajoutée en utilisant une demande enfantine, soit la recherche d'images du lapin Bunny.

Exemple 2 - NetScoring : critères de qualité de l'information de santé sur l'Internet [40]

Dernière mise à jour le: 19 septembre 2001 (version 4) < //www.chu-rouen.fr/nescoring>

Ensemble de critères utilisables pour évaluer la qualité de l'information de santé sur l'Internet, construit par Centrale Santé, un groupement professionnel destiné à réunir autour d'un projet fédérateur des centraliens intéressés par la santé et les professionnels de la santé. Le champ de l'évaluation comporte à la fois les sites Internet et intranets destinés soit aux professionnels de santé soit au « grand public ».

L'organisation produisant cet outil offre en quelque sorte une « labellisation » aux sites soumis à la notation.

Liste des critères :

- **Crédibilité** (sur 99 points): source, révélation, mise à jour, pertinence / utilité, existence d'un comité éditorial, cible du site, qualité de la langue, métadonnées.
- **Contenu** (sur 87 points): exactitude, hiérarchie d'évidence et indication du niveau de la preuve, citations des sources originales, dénégation, organisation logique, facilité de déplacement dans le site, exclusions et omissions notées, rapidité de chargement du site et des différentes pages, affichage clair des catégories d'informations disponibles.
- **Hyper-liens** (sur 45 points): sélection, architecture, contenu, liens arrières, vérification régulière de l'opérationnalité des hyper-liens...
- **Design** (sur 21 points): design du site, lisibilité du texte et des images fixes et animées, qualité de l'impression.
- **Interactivité** (sur 18 points): mécanisme pour la rétroaction, forums / chat, traçabilité.
- **Aspects quantitatifs** (sur 12 points): nombre de machines, de citations de presse, de productions scientifiques issues du site.
- **Aspects déontologiques** (sur 18 points): responsabilité du lecteur, secret médical.
- **Accessibilité** (sur 12 points): présence dans les principaux répertoires et moteurs de recherche, adresse intuitive du site.

Chaque critère est pondéré en trois classes: critère essentiel (coté de 0 à 9), important (coté de 0 à 6), mineur (coté de 0 à 2). Chaque critère est jugé par une échelle de Lijert à cinq occurrences. La note maximale est de 312 points. Par exemple: la « distinction hyper-liens internes et externes » est un critère mineur ; « En cas de modification de structure d'un site, les liens entre les anciens documents HTML et les nouveaux » est considéré comme un critère important ; « Contexte: source de financement, indépendance de l'auteur » ou « Citations des sources originales » sont des critères essentiels.

En conclusion

Les indices ou critères servant à l'évaluation⁵⁷ des instruments de recherche ou des ressources s'enrichissent pour s'adapter aux évolutions technologiques et sociologiques (de production, d'usage) en cours, se complexifiant par voie de conséquence.

Les méthodes et pratiques utilisées pour les systèmes de recherche interactifs doivent, dans ce contexte, être revues, d'autant plus qu'il s'agit parfois d'évaluer des outils qui seront manipulés par l'utilisateur final. Les questions d'ergonomie, d'utilité immédiate (*searchability*⁵⁸), peuvent alors prévaloir sur l'efficacité technique du moteur ou la qualité intrinsèque de l'information⁵⁹.

Certaines difficultés et contraintes apparaissent lors de la mise en place de procédures d'évaluation. On peut citer les points suivants :

- ces grilles d'analyse et d'évaluation supposent un travail préalable de définition de la politique de sélection et d'évaluation des ressources ;
- le contour de l'objet à analyser, c'est-à-dire le niveau de granularité de la ressource (le portail ou le site web, la page web, l'article sur la page, la base de données accessible depuis la page, le forum...) ou la catégorie de moteur d'indexation est essentiel et délicat à déterminer ;
- l'adéquation entre les finalités de l'évaluation et le choix des critères et des unités de mesure qui président à la définition de la pertinence est un aspect préparatoire essentiel dans la réussite de ces activités⁶⁰ ;
- bien sûr, les procédures, les outils et les modalités mêmes de cette évaluation doivent être clairement et officiellement établis [35] pour minimiser la subjectivité de la notion de pertinence et de l'activité d'évaluation.

Le travail à fournir pour établir et exploiter ces outils est lourd et complexe mais, professionnellement, il répond à un enjeu fort. Le développement des travaux de normalisation en sont un indicateur parlant. Mais la mutualisation des méthodes et des résultats reste indispensable dans ce contexte.

CONCLUSION

Postulat : évolution des pratiques et compétences

Beaucoup des méthodes et techniques documentaires ont été conçues pour des recherches thématiques dans des bases de références bibliographiques – recherches effectuées par des spécialistes de la recherche d'information. Ce sont, pour une grande part, ces pratiques qui ont été transposées et étendues à l'exploitation de documents électroniques en texte intégral d'abord, puis, depuis une dizaine d'années, aux ressources du Web par Internet.

L'amélioration continue des instruments de recherche – dont la maîtrise reste indispensable – rencontre aujourd'hui certaines limites. Ceci pose la question de l'efficacité réelle des outils de recherche qui, brassant des volumes toujours plus importants, cachent ces limites à l'utilisateur non averti [11].

Ces constats nous amènent à envisager d'autres solutions aux problèmes évoqués de surinformation [5], de silence, de déperdition... L'une des solutions semble être, de façon assez intuitive, de faire évoluer nos pratiques professionnelles, évolutions engagées chez certains depuis quelques années. Mais évoluer dans quelles directions ?

Participer à la production d'information

La prise en compte, dans les systèmes documentaires, de l'information produite par les entreprises, a déjà amené les professionnels de l'information et de la documentation à prendre une place dans la chaîne de production de l'information, mais toujours avec une approche traditionnelle (relation de l'écrit / auteur, les métadonnées se substituant aux notices bibliographiques, modèle base de données). Actuellement, presque tout le contenu du Web est destiné à être lu avec une contrainte supplémentaire [3] : la mise en forme des ressources est souvent peu appropriée à la lecture à l'écran et à la navigation, ou à l'exploitation par des outils d'analyse et cartographiques. De plus, le passage de la lecture à l'écran a fait naître de nouvelles pratiques de lecture, lourdes de conséquence [3] : « L'unité de sens n'est plus le contexte, c'est-à-dire le livre, l'article, l'auteur [...] mais le fragment "pertinent" considéré en lui-même [...]. C'est pourquoi le photocopillage, technique de prélèvement des fragments qui nous sont le plus directement utiles, est devenu une pratique tout à fait "naturelle". » [2]

L'objectif de ce travail en amont est de participer à une rationalisation de certains gisements d'information pour répondre au problème de la surinformation [5], et d'orienter les modalités de production de l'information en intégrant les éléments (structure, contenu) utiles à la recherche. Ce travail consiste également à être imaginatif dans les solutions proposées, en se détachant des modèles connus (base de données, document)⁶¹ [11] [45].

Participer à la production d'information, dans ce contexte du Web, avec une visée d'optimisation de la recherche sous-entend d'assimiler les concepts d'hypertexte, de granularité de l'information, et de sémantique.

Participer à la co-construction de dispositifs

Il s'agit ici d'étendre les pratiques déjà anciennes des réseaux documentaires en les adaptant à ce nouveau contexte. Les questions d'interopérabilité entre les systèmes et de prise en compte des normes⁶² seront de plus en

plus d'actualité. L'ordinateur étant utilisé pour mener (ou assister) certaines tâches répétitives (acquisition d'information, recherche), des activités d'évaluation (des besoins, des outils, des ressources) doivent se développer. Dans ce contexte, les approches collaboratives incluant l'utilisateur, ainsi que les approches normatives, sont alors indispensables.

Assister l'utilisateur à trouver et exploiter l'information

Le mythe de la recherche d'information a fait perdre de vue aux professionnels de l'information et de la documentation la finalité de toutes ces activités : l'exploitation et l'appropriation par un utilisateur de l'information trouvée. Les usages de l'information sont très nombreux et variés ; de plus, les besoins des usagers évoluent. Mais les enquêtes d'utilisation de l'Internet montrent que les moteurs de recherche sont utilisés pour des recherches d'information qui, pour une part importante, ne sont pas celles auxquelles répondaient les systèmes documentaires « traditionnels », à savoir la recherche de documents.

Le syndrome de l'utilisateur « chercheur scientifique », qui était de fait le référent pour la construction des systèmes documentaires, postule d'une part que la recherche documentaire porte sur « un article relatif à un sujet précis », et d'autre part que l'identification du besoin a été effectuée préalablement, lors de la démarche de recherche scientifique. Or, on sait aujourd'hui que les besoins sont beaucoup plus variés pour un chercheur, tant dans sa finalité que dans ses contraintes, et que ceci est multiplié pour les autres publics des systèmes d'information. En fonction des activités à réaliser – établir un cahier des charges, rechercher une méthode d'analyse, préparer un cours (professeur), préparer un exposé (élève), étudier l'offre commerciale pour l'achat d'un produit, préparer des travaux innovants dans son domaine d'expertise, surveiller les activités d'une entreprise concurrente, etc. –, on peut vouloir :

- retrouver une information ou un document connu, d'où l'importance accordée à la précision ;
- souhaiter des informations précises, de type question / réponse à l'intérieur d'un corpus, pour laquelle un seul document pertinent suffit : le document de référence (le plus demandé ? le plus référencé ?) ;
- réaliser une recherche plus exploratoire supposant un questionnement. Dans ce dernier cas, la réponse doit viser l'exhaustivité ou, du moins, une certaine représentativité du problème posé. Ici, le rappel est important.

En prenant en compte tous les types de recherche qui pourraient être utiles à un certain public, on note que seule la recherche de documents suppose un dispositif basé sur une formulation du thème [9, p. 91], cette dernière étant source principale d'ambiguïté et de bruit, les autres types de recherche supposant un accès direct au renseignement ou au contexte.

Des remarques de deux ordres différents peuvent alors être faites pour conclure. La conception de systèmes doit permettre de répondre à l'ensemble des besoins informationnels des utilisateurs [21]. Comprendre leurs pratiques et mieux connaître leurs besoins et contraintes reste une tâche complexe mais indispensable.

NOTES

(Toutes les localisations des adresses Internet ont été contrôlées en juin 2002.)

1. Sous l'angle des techniques de télécommunication et d'informatique liées à l'Internet et à la société de l'information [6].
2. Date de validation des informations : juin 2002.
3. Définition établie à partir de celle mise au point par les formateurs Internet ADBS, en 2001 (document interne).
4. <http://www.quigo.com>
5. <http://www.invisibleweb.com>
6. <http://www.searchengineshowdown.com/stats/freshness.html> (voir [59]).
7. <http://www.about.com>
8. <http://vlib.org>. Pour plus d'information : To the Virtual Library catalog. Gerard Manning. <http://vlib.org/AboutVL.html/>. Dernière modification : 22 janvier 2001.
9. <http://www.aol.fr>.
10. <http://www.indicateur.com>
11. <http://www.beaucoup.com>
12. <http://www.searchengineguide.com>
13. <http://www.searchability.com>
14. <http://www.searchpdf.adobe.com>
15. <http://www.soros.org/openaccess/view.cfm> (visité le 5 juillet 2002).
16. Nous ne traitons pas ici des logiciels ayant des fonctions de veille, paramétrables hors ligne, tels des produits comme Copernic (Copernic Technologies) ou BullsEye (Intelliseek), qui laissent plus de liberté pour la sélection des sources exploitées et le paramétrage des requêtes et des fonctions de surveillance (voir chapitre 8).
17. <http://searchenginewatch.com/sereport/01/05-metasearch.html> (voir [59]).
18. <http://www.question.fr>
19. <http://www.webhelp.fr/direct/>
20. <http://answers.google.com/answers/main>
21. <http://www.portail2.com> : « Un portail est une page qui regroupe un ensemble de renseignements divers, utiles et pratiques, permettant à l'utilisateur de trouver rapidement une information dans le secteur qu'il souhaite. On appelle un "portail spécialisé" une page regroupant un ensemble d'informations abordant le même thème. »

22. <http://www.educlie.fr>
23. <http://www.medisite.fr>
24. Cette définition peut s'appliquer aux portails sur les intranets et / ou extranets.
25. <http://www.dmoz.org/> Pour plus d'information : http://dmoz.org/Computers/Internet/Searching/Directories/Open_Directory_Project/Sites_Using_ODP_Data/
26. Nombre de visiteurs sur l'ensemble des quatre sites qui composent cet anneau.
27. <http://cogsci.soton.ac.uk/~harnad>
28. <http://www.archive.org>
29. Des collections d'archives conservées dans une « bibliothèque publique de l'Internet » en libre accès. Julie Krassovsky. *Le Monde interactif*, 28 septembre 2001.
30. Nous excluons, dans ce chapitre, les techniques d'identification et de sécurité nécessaires à toutes les fonctions transactionnelles (achat, etc.).
31. Les moteurs d'indexation et de recherche utilisés par les moteurs de recherche via Internet sont utilisables sur les intranets / extranets. Par exemple, en France : Fast, Google, Inktomi ou Exalead.
32. On trouve fréquemment le terme d'« annotation ».
33. Des compromis doivent être trouvés entre la qualité des images diffusées et les contraintes de « poids » et de « débit » utile. Les techniques de *streaming* constituent actuellement une solution alternative.
34. Les manuels ou livres blancs fournis par les moteurs de recherche sont en général peu détaillés.
35. <http://options.ke.voila.fr/>
36. FRIC (Filtrage et recherche d'information collaborative). <http://www.mrim.imag.fr/> (visité le 15 juin 2002).
37. Voir <http://www.searchenginecolossus.com>
38. Voir <http://www.abondance.com>
39. <http://www.rolist.com>
40. Qui reste l'un des moyens de collecte développés, y compris pour les moteurs de recherche, dans le cadre de la prise en compte du Web invisible.
41. Voir Peer-to-peer ou l'art de partager l'information, Kareen Frascaria, ZDNet France, 15 novembre 2001, <http://www.zdnet.fr/techreport/peer-to-peer/intro.html> (visité le 2 juillet 2002), ou encore Source : Présent difficile et avenir radieux pour le point à point, Gueric Poncet, *Le Monde interactif*, 26 juin 2002.
42. Fast Topic d'Alltheweb à partir d'OpenDirectory, catégories de Yahoo! pour Google... Pour plus d'information : http://dmoz.org/Computers/Internet/Searching/Directories/Open_Directory_Project/Sites_Using_ODP_Data/
43. <http://www.mala.bc.ca>
44. RealNames To Close After Losing Microsoft. Danny Sullivan, The Search Engine Report, June 3, 2002, <http://searchenginewatch.com/> (visité le 25 juin 2002).
45. <http://www.copernic.com/fr/products/summarizer/index.html>

46. <http://www.pertinence.net>

47. <http://www.antisearch.net>

48. <http://www.askjeeves.com>

49. <http://www.infoclic.fr>

50. Par exemple, recherche sur la « classification internationale des brevets » sur le site de l'INPI (<http://www.inpi.fr>) ; exploitation d'une base de questions / réponses sur le bricolage sur le site de Leroy Merlin avec Intuition (<http://www.leroymerlin.fr>) ; consultation des Lettres d'observation des chambres régionales des comptes ou interrogation du site internet de RTL (<http://www.rtl.com>) avec Spirit de TGID.

51. Fast / AllTheWeb ou Google indexent en temps réel les sources d'information d'actualités.

52. Cette partie s'appuie essentiellement sur [25] [29] [32] [33] [34] [35]. La présentation des différents types de pertinence prend appui sur [26].

53. Le terme anglais *relevant* renvoie à cette notion de conformité du document trouvé par rapport à la requête lancée dans le système.

54. Pour une définition de la conception d'un sujet, matière, contenu informatif d'un objet documentaire, voir [9, p. 115-140].

55. Le développement d'une édition électronique de forme différente mais de nature équivalente à l'édition traditionnelle répond en partie à cette question. Voir [44] [48] [49].

56. L'indice de popularité peut également relever de la pertinence-utilisateur.

57. Rappel, précision, silence et bruit, confrontés aux besoins des utilisateurs.

58. « Capacité qu'a une information à être trouvée, à être extraite le plus facilement et rapidement possible de la masse d'information circulant sur le Web. » [11].

59. Le succès de Google – le moteur le plus utilisé par le plus grand nombre de personnes – s'explique, bien sûr, par la performance de ses techniques de recherche, mais également par sa simplicité d'utilisation et son adaptation aux spécificités d'un pays (options de langue directement accessibles).

60. Par exemple, les méthodes mises en œuvre dans les protocoles TREC, développées pour des moteurs texte intégral, pénalisaient l'analyse des moteurs en langage naturel. Autre exemple : les grilles utilisées pour les banques de données ne sont pas toujours exploitables pour des ressources web (voir chapitre 4 et [30]).

61. On peut citer, par exemple, le travail réalisé par le centre de documentation du Centre national de la danse dans son approche novatrice, s'appuyant sur les concepts sous-jacents au Web sémantique, pour la conception d'une base de connaissances sur la danse mêlant toute nature et tout type d'information et de documents. [Journée d'étude ADBS, 2002].

62. Pour un point sur les normes, voir [4] et [7].

RÉFÉRENCES

Classement thématique, puis alphabétique par titre. (Toutes les localisations sur le Web ont été contrôlées fin juin 2002.)

A - Généralités

B - Recherche d'information : méthodes et techniques

C - Recherche d'information : outils et guides

D - Évaluation : méthodes et techniques

E - Évaluation : outils et guides

F - Conception de systèmes de recherche d'information et traitements documentaires

G - Ergonomie

H - Surveillance du secteur de la recherche via Internet

I - Instruments de recherche sur le Web, cités et exploités

A - Généralités

[1] Jean-Michel SALAÛN, Alain MARTER, Benoît EPRON, Stéphane BÉLLINA. Étude économique et juridique d'un portail pour les revues françaises en sciences humaines et sociales. Novembre 2001. http://isdn.enssib.fr/otr_pg/archiv.htm

[2] Jean-François BARBIER-BOUVET. Internet, lecture et culture de flux (entretien). *Esprit*, décembre 2001, p. 20-34

[3] L'écrit et l'écran. *Captain Doc*, mars 2002, n° 6. <http://www.ftpresskiosque.com/www/arc/captaindoc-txt/2002-03/thrd1.html#00000>. [Les rapports de l'écrit et de l'écran ; un entretien avec Brigitte Juanals : « Accès aux savoirs, de la page du livre à la page-écran » ; dossier complet <http://www.captaindoc.com/dossiers/dossier07.html>]

[4] Ghislaine CHARTRON. Standards - Normes - Documents numériques. Introduction générale [dossier électronique pédagogique]. Urfist de Paris, janvier 2000. web.ccr.jussieu.fr/urfist/presse/standard/coursintro.htm

[5] Anne SANOUILLET. La surcharge informationnelle. 2000. <http://www.urfist.cict.fr/lettres/lettre24/lettre24-22.html>

[6] Jean-François ABRAMATIC. Développement technique de l'Internet. Mis à jour le 25 juin 1999. <http://mission-dti.inria.fr/>

[7] Vincent QUINT. Normes et documents numériques. In : Les Jeudis du numérique, Grenoble, 6 juin 2002. <http://isdn.enssib.fr/archives/normes/normes.htm>

B - Recherche d'information : méthodes et techniques

[8] 14 outils de recherche sur le Web invisible. *Netsources*, 1999, hors-série n° 3

[9] Jacques MANIEZ. Actualités des langages documentaires : fondements théoriques de la recherche d'information. Paris, ADBS Éditions, 2002. [En particulier les premiers chapitres (III, IV et V) : notions d'objets informationnels et de document ; typologie des systèmes de recherche d'information]

[10] Ghislaine CHARTRON. Les outils de recherche sur le Web. In : La recherche d'information sur les réseaux : cours INRIA, 1996, Trégastel. Paris, ADBS Éditions, 1996

- [11] Hubert FONDIN. La recherche d'information dans les mémoires électroniques : l'enjeu documentaire. *Documentaliste - Sciences de l'information*, 1999, vol. 36, n° 4-5, p. 242-248
- [12] Philippe LEFÈVRE. La recherche d'informations, du texte intégral au thésaurus. Paris, Hermès, 2000
- [13] Hervé Rostaing. Le Web et ses outils d'orientation : comment mieux appréhender l'information disponible sur l'Internet par l'analyse des citations ? *Bulletin des bibliothèques de France*, 2001, t. 46, n° 1, p. 68-75. <http://bbf/enssib/fr>
- [14] Jean-Louis BÉNARD. Les portails d'entreprise : conception et mise en œuvre. Paris, Hermès, 2002. [Caractéristiques du portail d'entreprise, en particulier des technologies mises en œuvre et des principaux acteurs du marché ; démarche de conception]
- [15] Jean-Pierre LARDY. Méthodes de tri des résultats des moteurs de recherche. Mis à jour le 28 décembre 2001. <http://www.adbs.fr/site/repertoires/sites/lardy/risi.htm>. [Sommaire : <http://www.adbs.fr/site/repertoires/sites/lardy/toc.html>]
- [16] Catherine LELOUP. Moteurs d'indexation et de recherche. Paris, Eyrolles, 1997
- [17] François BOURDONCLE. Panorama et perspectives des outils de recherche d'information textuelle sur Internet. In : IDT 1999 : textes des communications. <http://www.exalead.com/Francois.Bourdoncle/idt99.html>
- [18] MEMBRES DU PROJET CLEVER IBM. Recherche intelligente sur Internet. *Pour la science*, août 1999, n° 262, p. 52-58. <http://www.pourlascience.com/numeros/pls-262/art-4.htm>. [Voir également en anglais : The CLEVER project <http://www.alma.den.ibm.com/cs/k53/clever.html>]
- [19] S. BRIN, L. PAGE. The Anatomy of a Large-Scale Hypertextual Web Search Engine. <http://www7.scu.edu.au/programme/fullpapers/1921/com1921.htm>
- [20] M. KOSTER. The Web robots FAQ. <http://www.robotstxt.org/wc/faq.html>
- [21] Sylvie DALBIN, Bruno SALLÉRAS. Une expérience d'utilisation d'un système d'information documentaire en langage naturel. *Documentaliste - Sciences de l'information*, 2000, vol. 37, n° 5-6, p. 312-324

C - Recherche d'information : outils et guides

- [22] Laura COHEN. Boolean Searching on the Internet. May 2002. <http://library.albany.edu/internet/choose.html>
- [23] Jean-Émile TOSELLO-BANCAL, Annie LÉON (rev.). Les outils de recherche d'information sur l'Internet. Mis à jour le 18 juin 2002. <http://web.ccr.jussieu.fr/urfist/outsej.htm>
- [24] Martijn KOSTER. The Web Robots Pages. <http://www.robotstxt.org/wc/faq.html>
- [25] Béatrice FOENIX-RIOU. Recherche et veille sur le Web visible et invisible : agents intelligents, annuaires sélectifs, interfaces des grands serveurs, portails thématiques. Paris, Bases Publications, Éditions Tec&Doc, 2001. [Les outils de recherche (sélectifs, sites fédérateurs, serveurs, agents, etc.), en dehors des grands moteurs ou annuaires]

D - Évaluation : méthodes et techniques

[26] Brigitte SIMONNOT. De la pertinence à l'utilité en recherche d'information : le cas du Web. In : Recherches récentes en sciences de l'information : convergences et dynamiques. Actes du colloque MICS-LERASS, 21-22 mars 2002, Toulouse. Paris, ADBS Éditions, 2002. P. 395-410

[27] France BOUTHILLIER, Kathleen SHEARER. Étude comparative de systèmes de veille concurrentielle en regard du traitement de l'information. In : Filtrage et résumé automatique de l'information sur les réseaux. Actes du Colloque ISKO, 5-6 juillet 2001, Nanterre. Nanterre, Université Paris-X, Centre de recherche en information spécialisée, 2001. P. 265-273

[28] Pierre LE LOARER. Indexation automatique, recherche d'information et évaluation. In : Le traitement électronique du document, cours INRIA, 1994, Aix-en-Provence. Paris, ADBS Éditions, 1994

[29] Monica LANDONI, Steven BELL. Information retrieval techniques for evaluating search engines: a critical overview. *Aslib Proceedings*, March 2000, vol. 52, n° 3, p. 124-129

[30] Jose PEREZ-CARBALLO, Tomek STRZALKOWSKI. Natural language information retrieval: progress report. *Information Processing and Management*, 2000, vol. 36, p. 155-178

[31] Howard GREISDORF. Relevance: an interdisciplinary and information science perspective. *Informing Science* (Special issue on information science research), 2000, vol. 3, n° 2. [Aspects historiques de l'évaluation des systèmes de recherche d'information]

[32] Hongseok PARK. Relevance of science information: origins and dimensions of relevance and their implications to information retrieval (theory). *Information Processing and Management*, 1997, vol. 33, n° 3, p. 339-352

[33] Stefano MIZZARO. Relevance: the whole history. *Journal of the American Society for Information Science*, 1997, vol. 48, n° 9, p. 810-832

[34] Linda SCHAMBER, Judy BATEMAN. User Criteria in Relevance Evaluation: Toward Development of a Measurement Scale. ASIS 1996 annual conference proceedings, October 19-24/1996. <http://www.asis.org/annual-96/ElectronicProceedings/schamber.html>. [Établissement, à partir d'une expérimentation, d'une liste d'échelle de mesure de critères de pertinence du point de vue de l'utilisateur.]

[35] Sylvie LAINÉ-CRUZEL. Vers un nouveau positionnement des professionnels de l'information : quelle valeur ajoutée, pour quels systèmes ? In : Filtrage et résumé automatique de l'information sur les réseaux. Actes du Colloque ISKO, 5-6 juillet 2001, Nanterre. Nanterre, Université Paris-X, Centre de recherche en information spécialisée, 2001. P. 13-23

[36] Lise HERZHAFT. Évaluation de la recherche d'information : petite bibliographie. URFIST de Lyon. In : Lettre 24/25, 1^{er} et 2^e trimestres 2000. [http://www.urfist.cict.fr/lettres/lettre24/pdf24/recherche information.pdf](http://www.urfist.cict.fr/lettres/lettre24/pdf24/recherche%20information.pdf)

E - Évaluation : outils et guides

[37] Cerise : les questions de base à se poser. Urfist de Paris, 1999, dernière mise à jour : 16/01/2002 par Julia Jumeau, Claire Panijel. <http://web.ccr.jussieu.fr/urfist/cerise/p361.htm/>

[38] Marc DUVAL. Classement des automates de recherche. Longueuil, Québec, 2001. <http://www.dsi-info.ca/classement-introduction.html>

[39] SAPRISTI (Sentiers d'accès et pistes de recherche d'informations scientifiques et techniques sur l'Internet). Lyon, INSA, [s. d.] <http://csidoc.insa-lyon.fr/sapristi/digest.html>

[40] Net Scoring® : critères de qualité de l'information de santé sur l'Internet. Central Santé. Dernière mise à jour le 19 septembre 2001 (version 4), révisée en mai 2001. <http://www.chu-rouen.fr/netscoring/>

[41] Text Retrieval Conference (TREC). <http://trec.nist.gov/>

[42] Université Laval. www.fse.ulaval.ca/fac/href/grille/grille.gif

F - Conception de SRI et traitements documentaires

[43] Dossier « Images ». *Bulletin des bibliothèques de France*, 2001, t. 46, n° 5. <http://bbf.enssib.fr>

[44] Ghislaine CHARTRON. Évolution dans le modèle éditorial des articles scientifiques : analyse économique et stratégique. In : *Recherches récentes en sciences de l'information : convergences et dynamiques. Actes du colloque MICS-LERASS*, 21-22 mars 2002, Toulouse. Paris, ADBS Éditions, 2002. P. 371-392

[45] Yannick PRIÉ. Exploitation de documents audiovisuels numériques dans un système d'information audiovisuelle. 25 janvier 2000. <http://lisi.insa-lyon.fr/~yprie/these/node6.html>

[46] Cécile KATTNIG. Gestion et diffusion d'un fonds d'images. Paris, Nathan/VUEF, ADBS, 2002

[47] Chahab NASTAR. Indexation et recherche d'images : enjeux, méthodes et perspectives. In : *IDT 1999 : textes des communications*. <http://www-rocq.inria.fr/~nastar/nastar-idt99.html>

[48] Le numérique : apports et contraintes pour les archives. *Dossiers de l'audiovisuel* (Les archives télévisuelles à l'heure du numérique), octobre 2000, n° 93, p. 6-34

[49] Stevan HARNAD. Repenser la communication scientifique : l'auto-archivage par l'auteur. In : *Publication électronique des résultats de la recherche. Rencontre du 29 mars 2002*. <http://cogsci.soton.ac.uk/~harnad>

[50] Heting CHU. Research in image indexing and retrieval as reflected in the literature. *Journal of the American Society for Information Science and Technology*, October 2001, vol. 52, n° 12, p. 1011-1018

[51] Michèle HUDON. Structuration du savoir et organisation des collections dans les répertoires du Web. *Bulletin des bibliothèques de France*, 2001, t. 46, n° 1. <http://bbf.enssib.fr>

[52] Gabriel GALLEZOT, Ghislaine CHARTRON, Jean-Max NOYER. Une archive ouverte des publications en InfoCom. In : colloque Place et enjeux des revues pour la recherche en infoCom, Nice, mars 2002 (sur le site : <http://archivesic.ccsd.cnrs.fr>).

G - Ergonomie

[53] Xavier CASANOVA, Joëlle COHEN. L'écran efficace : une approche cognitive des objets graphiques. *Documentaliste - Sciences de l'information*, 2001, vol. 38, n° 5-6, p. 272-289

[54] J. M. Christian BASTIEN, Corinne LEULIER, Dominique L. SCAPIN. L'ergonomie des sites web. In : Créer et maintenir un site web : cours INRIA 1998, Pau. Paris, ADBS Éditions, 1998. P. 111-174

H - Surveillance du secteur de la recherche via Internet (chiffres clés, marché, technologie)

[55] Abondance : <http://www.abondance.com/> [Actu Moteurs (w) Service d'information (lettre *Actu Moteurs*, hebdomadaire et gratuite, actualités, fiches techniques, bibliographie, etc. Certaines parties du site sont payantes) sur les outils de recherche (annuaires, moteurs) francophones et mondiaux]

[56] Internet Planète : <http://winternet.planete.qc.ca/>

[57] Bibliothèque universitaire - Sciences - Université catholique de l'Ouest - Angers – 2002 : <http://www.uco.fr/services/biblio/cdps/index.html>

[58] *Netsources*. Éditeur FLA Consultants (<http://www.fla-consultants.fr/>). « Lettre bimestrielle consacrée à la recherche sur Internet (méthodologies de recherche, analyses d'outils performants, descriptions de sites dignes d'intérêt...) ».

[59] Search Engine : searchenginewatch.com (Ed. Danny Sullivan) et www.searchengineshowdown.com/ (Ed. Greg R. Notess) [Informations techniques, chiffrées, sur les moteurs de recherche, y compris des rapports sur des liens inactifs (*dead links*), le taux de recouvrement (*overlap*) entre moteurs, ou des problèmes rencontrés (Inconsistencies Reports)]

[60] Veille : <http://www.veille.com/> [Service d'information (lettre, actualités, fiches techniques, etc.) consacré aux moteurs de recherche sur Internet, surtout d'un point de vue technique]

I - Instruments de recherche sur le Web, cités et exploités. Documentation en ligne

[61] AltaVista : <http://www.altavista.fr>

[62] FAST : <http://www.alltheweb.com>

[63] Google : <http://www.google.com>

[64] HoBot : <http://www.hotbot.com>

[65] Kartoo : <http://www.kartoo.com>

[66] Mapstan : <http://www.mapstan.net>

[67] Voilà : <http://www.voila.fr>

[68] Search Engine Colossus : <http://www.searchenginecolossus.com>

INDEX DES AUTEURS RÉFÉRENCÉS

Les chiffres renvoient au numéro de référence bibliographique. Ont été exclus les sites web.

ABRAMATIC Jean-François, 6
BARBIER-BOUVET Jean-François, 2
BASTIEN J. M. Christian, 54
BATEMAN Judy , 34
BELLINA Stéphane, 1
BENARD Jean-Louis , 14
BOURDONCLE François, 17
BOUTHILLIER France, 27
BRIN S., 19
CASANOVA Xavier , 53
CHARTRON Ghislaine, 4, 10, 44, 52
CHU Heting, 50
COHEN Joëlle , 53
COHEN Laura , 22
DALBIN Sylvie , 21
DUVAL Marc, 38
EPRON Benoît , 1
FOENIX-RIOU Béatrice, 25
FONDIN Hubert , 11
GALLEZOT Gabriel , 52
GREISDORF Howard , 31
HARNAD Stevan , 49
HERZHAFT Lise, 36
HUDON Michèle, 51
IBM -
MEMBRES DU PROJET CLEVER , 18
INSA Lyon, 39
JUANALS Brigitte, 3
KATTNIG Cécile , 46
KOSTER Martijn , 20, 24
LAINE-CRUZEL Sylvie, 35
LANDONI Monica, Steven BELL, 29
LARDY Jean-Pierre, 15
LE LOARER Pierre, 28
LEFEVRE Philippe , 12
LELOUP Catherine, 16
LEON Annie (rev.), 23
LEULIER Corinne, 54
MANIEZ Jacques , 9
MARTER Alain , 1
MIZZARO Stefano , 33
NASTAR Chahab, 47
NOTESS Rogers, 59
NOYER Jean-Max , 52
PAGE L., 19
PARK Hongseok, 32
PEREZ-CARBALLO Jose , 30
PRIE Yannick , 45
QUINT Vincent , 7
ROSTAING Hervé , 13
SALAÜN Jean-Michel, 1
SALLERAS Bruno, 21
SANOUILLET Anne, 5
SCAPIN Dominique L., 54
SCHAMBER Linda , 34
SHEARER Kathleen , 27
SIMONNOT Brigitte, 26
STRZALKOWSKI Tomek , 30
SULLIVAN Danny, 59
TOSELLO-BANCAL Jean-Émile, 23

ANNEXE I - QUELQUES DONNÉES CHIFFRÉES ¹

Les quelques chiffres qui suivent visent à donner les tendances des évolutions de ces dernières années. Ils renseignent avant tout sur les contraintes auxquelles se trouvent confrontés les acteurs de l'accès à l'information du Web, mais aussi sur les difficultés que peuvent rencontrer les professionnels de l'information et sur la nécessité de se doter d'outils et de méthodes d'évaluation et d'exploitation adaptées.

- En ce début 2002, AllTheWeb et Google annonçaient avoir dépassé les 2 milliards d'objets référencés ; Altavista affichait un index de 30 millions de pages web en mai 1996. Pour sa base, Google annonce 1,465 milliard (73,1 %) de pages web indexées ; 500 millions (25 %) de pages non indexées ² ; 35 millions (1,75 %) de fichiers d'un autre type de format, et 3 millions (0,15 %) de pages réindexées quotidiennement. Fast traite plus de 3 000 sources de dépêches d'actualités quasiment en temps réel.

- Le taux de recouvrement entre moteurs de recherche, c'est-à-dire le nombre de documents identiques trouvés dans les premiers résultats, reste faible. Ainsi, d'après [59], environ la moitié des pages trouvées le sont par un seul moteur ; plus de 78 % par trois moteurs.

- 2 millions de sites web sont référencés sur l'annuaire Yahoo! ou Looksmart, et près de 3,42 millions de sites (600 000 sites de plus en 6 mois) pour l'ODP. Au niveau francophone, près de 150 000 sites sont répertoriés par Yahoo.fr, 70 000 sur les guides de Voila, de Lycos France ou de MSN ; environ 35 000 sites francophones sont gérés par l'ODP. Cela donne, au niveau mondial, des systèmes de catégories très importants : quelque 70 000 catégories pour Looksmart et 396 000 pour l'ODP ; Tiscali (Nomade) annonce 900 catégories.

- Les statistiques fournies montrent que, malgré la puissance des instruments de recherche, ils ne sont que faiblement représentatifs de la totalité du Web à un instant donné : sur 200 millions de pages web, liées entre elles, les moteurs de recherche se concentrent sur un « cœur » de 50 millions de pages ³.

- Concernant les Internautes : aujourd'hui, on estime à 9,10 millions le nombre de personnes connectées au réseau mondial en France, et 516,1 millions dans le monde ; 25 % des Français utilisent régulièrement Internet.

- 27,5 % du trafic sur les sites Internet francophones sont générés par les outils de recherche, moteurs et annuaires. Plus de 7 millions de requêtes sont effectuées chaque jour sur le site de Google.fr, qui draine 3,11 millions de visiteurs uniques par mois ⁴. « Le français est la cinquième langue la plus utilisée sur Google, et 20 % des entreprises ont recours au Web pour mener leurs achats professionnels. ⁵ »

1. Sources : www.searchenginewatch.com et www.notess.com (Greg R. Notess) ; www.abondance.com ; www.jmm.com/index.html (Jupiter Media Metrix) ; www.nielsen-netratings.com (Nielsen Netratings).

2. Les adresses ainsi que les textes des ancres sont indexés, mais pas le texte de la page.

3. Web : les secrets des moteurs de recherche. *Science et Vie*, novembre 2000, n° 998, p. 138-148.

4. Source Nielsen Netratings, avril 2002, http://www.nielsen-netratings.com/pr/pr_020311_france.pdf.

5. Résultats d'une étude menée par IDC pour le compte de la Fédération des entreprises de vente à distance (Fevad) [cité par : Premier ministre - Direction du Développement des Médias - France. L'Internet en France 2002. <http://www.internet.gouv.fr/francais/chiffcles/france.htm>].

ANNEXE II - RÈGLES GÉNÉRALES POUR FORMULER UNE REQUÊTE

La formulation d'une requête suit des règles précises basées sur des opérateurs et une syntaxe plus ou moins complexes. Même si la forme prise par les opérateurs et l'expression de la syntaxe d'une requête sont différentes d'un module de recherche à un autre, leurs fonctions restent identiques.

Un mot : chaîne de caractères alphanumériques

Guillemets pour caractériser les mots composés

Accentuation et Majuscules/minuscules (différenciés ou non)

Opérateurs :

+ devant un mot : présence obligatoire

- devant un mot : exclusion obligatoire

* opérateur de troncature

Recherche par champs : par défaut ou à préciser

Champs disponibles liés à la structure HTML : host, link, url, title, text, date, etc.

Opérateurs pour l'expression de la requête : AND, OR, NEAR, NOT Near

Ces techniques de recherche en « texte intégral » mises en œuvre actuellement dans la majorité des moteurs de recherche du Web sont, pour la plupart d'entre elles, identiques à celles exploitées par les logiciels d'interrogation des serveurs professionnels depuis trente ans.

ANNEXE IV - AOL/EXALEAD (AFFINEMENT)

Le système analyse automatiquement les documents du lot de résultats sur la base de techniques essentiellement – mais pas exclusivement – statistiques, et propose des expressions ou mots clés issus des documents, ici des groupes nominaux (écran A). Cette fonction permet de valider ou d'invalider rapidement la formulation utilisée : la sélection de l'une de ces propositions réduit la recherche initiale (écran B).

Écran A : recherche « effets et dangers du dopage dans le sport »

The screenshot shows a search interface with the following elements:

- Search Bar:** Contains the text "effets et dangers du dopage dans le sport" and a "Rechercher" button.
- Filters:** Radio buttons for "Internet français" (selected) and "mondial".
- Search Results:**
 - Header: "Vous cherchez : effets et dangers du dopage dans le sport"
 - Count: "112 documents trouvés en 0.18 secondes"
 - Refinement options: "Affinez avec les mots :", followed by a grid of terms:
 - ☑ Dangers du dopage
 - ☑ Problèmes du dopage
 - ☑ Lutte anti-dopage
 - ☑ Fins de dopage
 - ☑ Face au dopage
 - ☑ Définition du dopage
 - ☑ Lutte antidopage
 - ☑ Augmentation de la masse musculaire
 - Search results list:
 - Les dangers du sport : le dopage**
Les dangers du sport : le dopage Depuis quelques années, le ... n'est pourtant qu'un effet purement psychologique, comme les ...
[www.pratique.fr/actualites/2400.htm](#)
 - LA CREATINE ET LE DOPAGE DES SPORTS**
LA CREATINE ET LE DOPAGE DES SPORTS Claude LEVY-GALLIEN ... un dopage avéré. 6) Dans le contexte économique de "sport ...
[www.tous.com/la_creatine_et_le_dopage_des_sport.htm](#)
 - Le dopage et le sport : les risques pour la ...**
... Conférence Mondiale sur le dopage dans le sport, La déclaration ...
recherché. [http://www.sports.org/dopage/effets_recherches.htm](#) La réglementation : ...
[www.caducee.net/actualites/specialises/medecine-du-sport/dopage03.asp](#)
 - Les effets du sport**
Les effets du sport Le sport et l'appareil locomoteur ... Les effets psychologiques du sport Les dangers du sport L' ...
[www.medecine.com/2000/04/04/040404.htm](#)
 - Dopage et Sports**
... L'ENFANT FACE AU DOPAGE DANS LE SPORT L' Avis ... de la pratique de sport Dans la population pratiquant une ...
[www.unsports.com/avis_specialises/index_at_dopage-enfant.htm](#)
 - ACOPHRA - Pourquoi le dopage ? La position du médecin du ...**
5. Pourquoi le dopage ? La position du médecin du sport ... Où se situe le sport de haut niveau ? Le sport ...
[www.adph.org/acophra/091200-5.html](#)
 - Le dopage dans le cyclisme. Et qu'il se passe**
- Amazon.fr Widget:** Located on the right side, it includes the text "Livres, CD, Jeux..." and "avec amazon.fr". Below it is a search bar with the text "Rechercher : effets et dangers du dopage" and buttons for "OK" and "Tous les produits".

1 - Instruments de recherche sur le Web

Écran B : sélection de la catégorie « Problème du dopage »

The screenshot shows a search engine interface with the following elements:

- Search Bar:** Contains the text "effets et dangers du dopage dans le sport" and a "Rechercher" button.
- Language Selection:** Radio buttons for "Internet français" (selected) and "mondial".
- Search Results:** A list of 23 documents found, with the first result highlighted in grey. The highlighted result is titled "Problème du dopage" and includes a snippet: "Vous cherchez : effets et dangers du dopage dans le sport > Problème du dopage".
- Refinement Options:** A section titled "Affinez avec les mots :" with several filterable categories: "Liste des produits dopants", "Taux d'hématocrite", "Affaire Festina", "Lutte antidopage", and "Hormones de croissance".
- Document Snippets:** Several document titles and snippets are visible, including "Les dangers du sport : le dopage", "dopage", "EPO", "Le guide de prévention sportive et de lutte contre le...", and "Gagner le Tour à l'eau c'est possible!".
- Right-Side Panel:** A vertical sidebar on the right contains a search box with the text "L'urine, CO2, Jeu..." and a "Rechercher :" button. Below it are buttons for "Tous les produits" and "Tous les produits".

ANNEXE V - FAST TOPIC (REGROUPEMENT)

The screenshot shows the alltheweb search engine interface. At the top, there is a search bar with the text "alltheweb" and "all the web, at the time". Below the search bar, there are navigation links for "Web pages", "News", "Pictures", "Videos", "MP3 files", "FTP files", and "customize i hate". A language dropdown menu is set to "French". A search button labeled "SEARCH" is visible, along with a checkbox for "exact phrase".

The search results are displayed under the heading "Web pages found" and "FAST Topics". The "Web pages found" section shows a list of 140 results, with the first six items visible:

1. **le dopage : les débuts**
Actuellement, le dopage se définit comme l'utilisation de produits et de méthodes destinés à augmenter artificiellement la performance et dont les effets présentent des dangers supérieurs ou égaux à ceux sur la santé des personnes.
— www.viol-dalen.com/23-dopage/23page1.htm (32.8 kB)
2. **questions sur le dopage**
Qu'est-ce que le DOPAGE ? POURQUOI se dope-t-on ? Quels sont les BUTS RECHERCHES ? Le dopage améliore-t-il la PERFORMANCE ? Quels sont les EFFETS et les DANGERS du dopage ? Quelle est la RÉALITÉ du DOPAGE ? Comment est faite la LUTTE contre le dopage ?
http://www.lesports.org/lesports/question.htm (3.4 kB)
3. **dopage - les risques**
centre d'aide et d'accueil des toxicomanes et à leur familles) informations et prévention sur le dopage
http://www.multimedia.com/dopage/23page3.htm (33.2 kB)
4. **Medicine - Dépendances - Le dopage**
Les effets, les risques et les dangers des drogues (substances psychoactives) valent avant les produits et l'usage qu'on en fait.
— www.medicine.fr/dependances/23page1/46.148
5. **dopage**
Les produits dopants perturbent les équilibres physiologiques naturels, développent artificiellement les capacités du sportif et amoindrissent ses réactions de défense.
— www.lesports.org/lesports/23page1/23page1.htm (3.9 kB)
6. **SPORTS - Exercices - Drogues et sport / Drogues**
On doit rendre à Fiamme de Coubarin, officine de-matériaux sportives, d'avoir repéré ce qui menace le sport et doit par conséquent être sanctionné, dans une perspective de défense de celui-ci. Dans un document tenu à Rome en 1923, il...
http://www.ons.fr/23/23page1/23page1.htm (35.1 kB)

The "FAST Topics" section on the right lists several related topics:

- [Medicine - Dépendances - Drogues](#)
- [Medicine - Dépendances - Drogues](#)
- [Evidences](#)
- [Evidences - Sources - Drogues](#)
- [Site No. 10 - Site Members](#)
- [Sports - Drogues - Drogues](#)
- [Sports - Drogues - Drogues](#)
- [Sports - Drogues - Drogues](#)

ANNEXE VI - VIVISIMO (REGROUPEMENT HIÉRARCHISÉ)

L'utilisateur a une vision plus précise de l'ensemble des résultats grâce à une représentation hiérarchisée des regroupements (*clusters*) créés automatiquement par le moteur de recherche.

