ACADEMIE DE PARIS

UNIVERSITE RENE DESCARTES

Faculte de medecine cochin port royal

Année 2002 N°

THESE

Pour le

DIPLOME D'ETAT

DE DOCTEUR EN MEDECINE

(Résident)

PAR

HUYNH
Thanh Liem
Né le 11/08/1967 à Saigon (VIETNAM)

PRESENTEE ET SOUTENUE PUBLIQUEMENT LE :

RECHERCHE D'INFORMATIONS MEDICALES SUR INTERNET DANS L'EXERCICE QUOTIDIEN DE LA MEDECINE GENERALE

Une étude comparative entre deux outils complémentaires:
Moteur de recherche et Annuaire thématique.

PRESIDENT : M. A. VENOT PROFESSEUR
DIRECTEUR DE THESE : M. H. FALCOFF

VU LE DOYEN

J.F. DHAINAUT

VU ET PERMIS D'IMPRIMER
LE PRESIDENT DE L'UNIVERSITE

P. DAUMARD

TABLE DES MATIERES

INTRODUCTION

En octobre 2001, sur une des listes de discussion dédiées aux échanges entre médecins était publiée ce cas d'école :

Un homme de 26 ans sans ATCD particulier consulte pour les symptômes suivants survenus en plein repas du soir : hypersialorrhée et hypersudation intenses, trouble de l'accomodation visuelle, douleur abdominale, vomissements, et difficulté à la miction.

1) Identification de la présence de champignons dans le repas

2) Interrogation de Google (un moteur de recherche sur Internet www.google.com) : [intoxication champignon]

3) Consultation du premier lien http://perso.wanadoo.fr/sftg.pn/rev.pres.intoxication.champignons.9910.htm notamment le tableau sur les durées d'incubation courte : identification d'un syndrome muscarinien ou cholinergique

Liens sur les champignons cassés

4) Interrogation Google [syndrome muscarinien champignons]

le premier lien (Doctissimo) est cassé

le deuxième donne les champignons en cause : www.multimania.com/sms/initiation/intoxications.htm

5) Interrogation Google [inocybes veneneux]

aucun résultat pertinent

6) Interrogation Google Images (par http://images.google.com/advanced_image_search?) essais successifs inocybe, puis clitocybe (à noter que le clitocybe dealbata est présent en bas de la page de résultats pour inocybe)

7) Identification du clitocybe dealbata par le patient ou la famille grâce aux images

8) Appel au centre anti-poison

Durée totale : environ 3 minutes.

Autre exemple cité, plus proche de la pratique quotidienne :

Une patiente interroge son médecin traitant : « mon mari est suivi pour une maladie de Berger, est-ce héréditaire ? »

Interrogation de Google [maladie Berger] : le premier lien donne sur www.esculape.com qui lui-même pointe vers www.invivo.net/f2n/pro/glomerulonephrites/gniga.htm où la réponse est donnée : il existe quelques formes familiales.

Ainsi, « En 2001, Internet est devenu une source majeure d'informations scientifiques et médicales [Schatz97] (1) », comme on peut lire désormais sur le site du CISMeF (Catalogue et Index des Sites Médicaux Francophones : www.chu-rouen.fr/cismef) de l'université de médecine de Rouen, qui fait autorité en matière d'information médicales sur Internet. Le contenu du document (2) dont est issue cette citation peut se résumer ainsi : Bien qu'Internet soit devenu une source majeure d'informations scientifiques et médicales, y trouver la réponse satisfaisante à ses interrogations reste pour le professionnel de santé une tâche malaisée, tant du point

de vue de la pertinence des documents extraits que de celui de leur validité. Le CISMeF est un projet de type annuaire (ou catalogue de sites web) qui a été entrepris afin de répondre à ce besoin de trouver des documents à contenu médical à la fois pertinent et validé, basé sur une indexation manuelle par des documentalistes humains.

Toutefois, la nature même du CISMef ainsi que de tous les annuaires thématiques (constitution par indexation manuelle humaine) rend certaines recherches effectuées par son intermédiaire moins aisées et parfois moins satisfaisants que ceux obtenus par l'utilisation d'un moteur de recherche (indexation automatisée). Annuaire et moteur de recherche sont ainsi complémentaires dans leur approche de l'indexation de l'information : à la spécificité de l'annuaire répondent la simplicité, l'exhaustivité et la réactivité du moteur de recherche.

Une fréquentation assidue des listes de discussions de médecins utilisateurs d'Internet m'a permis de constater qu'il est communément admis par cette communauté qu'un moteur de recherche tel que Google suffit la plupart du temps, dans une recherche banale, à obtenir la réponse satisfaisante à leur question. L'objectif de ce travail est de tenter de vérifier, selon une démarche aussi objective que possible, le bien-fondé de l'utilisation préférentielle du moteur de recherche par rapport à l'annuaire thématique, et d'essayer de dégager les principes qui permettraient d'améliorer la pertinence des réponses obtenues par son intermédiaire, en se plaçant dans le cadre pratique d'un exercice quotidien de la médecine générale en ville.

I/ INTERNET ET INFORMATIONS MEDICALES

A) Présentation d'Internet

1. Description globale

1.1) Bref rappel historique

Fin des années 60 : débuts de la commutation par parquets. Les ordinateurs commencent à communiquer sur le mode peer-to-peer.

1969: ARPA-Net est créé par le département de la défense aux E.U. (ARPA= Advanced Research Projects Agency). Il s'agit d'un réseau fonctionnant sur le principe de la commutation par paquets. Premiers sites connectés: UCLA, SRI (Stanford Research Institute), UCSB et l'Université d'Utah. Le courrier électronique devient vite populaire, les conversations sont peu protocolaires.

1970-1980: divers autres réseaux se créent (THEORYNET, CSNET), basés sur d'autres protocoles.

1973: premières connections internationales à ARPANet: *University College* de Londres et le *Royal Radar Establishment* de Norvège.

1979: Début de Usenet, réseau de support de forums de discussion (basé sur le protocole UUCP - Unix-to-Unix Copy Protocol.

1980: Vinton Cerf propose un plan pour connecter CSNET et ARPANET de manière transparente (mêmes protocoles du point de vue de l'utilisateur).

1982: passage d'ARPA-Net aux protocoles TCP (Transfer Control Protocol) et IP (Internet Protocol), généralement utilisés conjointement (TCP/IP). De là sont nés les

termes d'**internet** pour signifier un ensemble de réseaux interconnectés utilisant TCP/IP et d'**Internet** pour signifier l'ensemble complet de tous ces internets. Le département de la défense (DoD) adopte TCP/IP comme standard.

1983: subdivision d'ARPA-Net (militaires+recherche) en ARPANET (purement recherche) et MILNET (purement militaire). Premier serveur de noms à l'Université du Wisconsin (l'utilisateur - ou sa machine - n'a plus besoin de connaître le chemin exact pour arriver au destinataire).

1986: NSFNET est créé. Il s'agit d'une infrastructure de communication à 56Kilobits par seconde (Kbps) connectant 5 super-centres de calcul (Princeton, Pittsburg, UCSD (Uni. Calif. San Diego), UIUC (Uni. Illinois Urbana Champaign) et Cornell). Cela provoquera un grand essor de connectivité, surtout des universités.

1986: l'*Internet Engineering Task Force* (IETF) est créé pour coordonner le développement technique de nouveaux protocoles. (Pour en savoir plus cf Annexe 1)

1986-1987: le trafic Usenet passe sur le réseau ARPANET et remplace le protocole UUCP par NNTP (Net News Transfer Protocol).

1988: NSFNET passe à 1,544 Mégabits par secondes (Mbps).

1989: Lors du tremblement de terre de San-Francisco, l'Internet est resté en action alors que le téléphone et les autres moyens de communication étaient hors-service, démontrant ainsi la robustesse de son architecture décentralisée.

1989-1990: Tim Berners-Lee conçoit le World-Wide Web (WWW).

1990: ARPANET est remplacé par NSFNET. Un premier prototype WWW est développé sur station de travail NeXT par Tim Berner's-Lee, au CERN.

1991: Durant le coup d'état en URSS et les évènements de Tienanmen, l'Internet a permis de passer des nouvelles et des informations à travers le monde entier, démontrant ainsi son aptitude à déjouer les contrôles de type censitaire et coercitif.
1991: NSFNET passe à 44,736 Mégabits par secondes (Mbps).

1991: Le "Stanford Linear Accelerator Center" installe le premier serveur Web pour permettre l'accès à leur base de données de publications de physique.

Janvier 1992: le premier "butineur" (browser) WWW ne supportant que du texte est disponible. Le code pour développer un serveur HTTP est rendu publique.
1992: l'*Internet Society* (ISOC) est créée.

Mars 1993: Eric Bina et Marc Andreessen , travaillant au NCSA (National Center for Supercomputing Applications) créent Mosaic, le premier "butineur" WWW graphique, supportant textes et images, et le mettent à disposition du public.

1994: Jim Clark et Marc Andreessen fondent Netscape Communications, Inc.
Novembre 1994: des fournisseurs privés (MCI, ANS et Sprint) reprennent l'infrastructure réseau de l'Internet et prennent peu à peu le relais de NFSNET.

Décembre 1994: Netscape Navigator, butineur graphique, est disponible. Son succès est foudroyant.

Décembre 1995: Java fait son apparition.

1996: l'infrastructure principale (backbone) passe à 155 puis 622 Mégabits par seconde (Mbps).

Juillet 1998 : plus de 36 millions de machines sont connectées à l'Internet, abritant plus de 4 millions de domaines formant un ensemble de plus de 500 millions de pages web. Toutes les statistiques concernant l'Internet continuent de décrire une courbe exponentielle. Il en est encore ainsi en octobre 2000 , date du début de la rédaction de ce travail.

1.2) Internet : un réseau d'ordinateurs interconnectés

Internet réalise la synergie entre l'informatique et les télécommunications : l'informatique permet un traitement automatisé de l'information à des vitesses toujours croissantes, et les télécommunications permettent l'acheminement des ces informations sur l'ensemble du globe à des vitesses elles aussi toujours croissantes.

Afin de pouvoir coopérer, les ordinateurs nécessitent d'être connectés, cette connexion devant être assurée par la conjonction deux facteurs : une liaison physique et un protocole commun d'échanges des données. La liaison physique est assurée par l'infrastructure de télécommunications, en croissance constante tant pour le territoire couvert que pour la vitesse de transport des données. Le protocole d'échange des données est le TCP/IP : il fonctionne selon le mode de commutation par paquets.

A l'origine structuré autour d'ARPANET, Internet est devenu un réseau international d'ordinateurs communiquant entre eux grâce à ce protocole d'échange de données standard : les protocoles TCP et IP formant la « couche » TCP/IP. TCP/IP permet aux différents ordinateurs branchés au réseau Internet de communiquer ensemble de façon transparente pour l'usager, indépendamment des types d'ordinateurs utilisés (Mac, PC, Unix ou autres) et des systèmes d'exploitation installés (Mac OS, Windows, Unix, Linux ou autres), pour peu qu'on dispose des logiciels appropriés.

Les ordinateurs connectés à l'Internet se répartissent en deux catégories: d'une part ceux qui sont connectés en permanence, dont la plupart font office de serveurs, et qui appartiennent le plus souvent à des institutions ou des personnes morales (associations à but non lucratif, entreprises, etc…); et d'autre part ceux qui appartiennent à des particuliers qui pour la plupart se connectent au réseau de façon ponctuelle pour leur usage personnel ou professionnel.

Afin de se représenter Internet, on peut imaginer une sorte de réseau similaire à celui du métro parisien, mais avec beaucoup plus de stations de correspondance et qui s'étendrait sur toute la surface du globe terrestre, et dont chaque « nœud » (c'est

à dire chaque station selon notre analogie) serait constitué d'un réseau informatique plus restreint. L'information qui y circule est découpée en « paquets » standardisés dont chacun circule d'un nœud à l'autre en tentant d'emprunter le meilleur trajet pour parvenir au nœud final. A l'arrivée, tous les parquets sont réassemblés pour reconstituer l'information qui a été envoyée au niveau du point de départ.

Ces opérations se font à une vitesse telle qu'elles paraissent quasi-instantanée la plupart du temps, surtout pour de petites quantités d'informations. Ainsi, à l'aide de logiciels adéquats, chaque ordinateur connecté à Internet peut communiquer de façon quasi-instantanée avec n'importe quel autre ordinateur connecté lui aussi, quel que soit l'éloignement géographique, et échanger des informations « à la volée» (c'est ce qu'on désigne par « en temps réel »).

1.3) Internet : un espace de libre expression

Afin d'être connecté à l'Internet, il suffit pour l'utilisateur de disposer de:

- Un contrat d'abonnement chez un fournisseur d'accès à Internet

- Un canal d'échanges de données entre l'ordinateur de l'utilisateur et un des ordinateurs du fournisseur d'accès : ligne téléphonique usuelle, ligne Numéris, câble, voie ADSL, réception satellite, et bientôt réseau hertzien ou encore réseau électrique (ces deux dernières techniques étant encore au stade de l'expérimentation actuellement)

- Un ensemble de logiciels permettant d'assurer les opérations souhaitées, dont la plupart peuvent être obtenus gratuitement : navigateur (ou browser), client e-mail, client de newsgroups, logiciel de transfert de fichiers (ou client FTP), logiciel de dialogue en direct, etc…

Une fois effective cette connexion à l'Internet, tout utilisateur peut mettre à la disposition de quiconque en fera la demande (à travers Internet) tout fichier informatique en sa possession, chaque fichier pouvant être le produit de sa création et reflétant ses propres opinions : Internet permet ainsi à tous ses utilisateurs de s'exprimer quasiment sans aucune censure ni régulation autre que l'audience qu'il remporte auprès de la communauté des gens connectés.

Ce qui aboutit parfois à des dérives extrêmes dont la prévention et la répression est l'un des enjeux actuels d'Internet (cf plus loin).

1.4) Internet : une bibliothèque virtuelle

Un ensemble de techniques logicielles permet à chaque opérateur d'un système ou d'un réseau local de mettre à disposition de chaque système connecté au réseau Internet une partie bien déterminée (et fixée par l'opérateur lui-même) de l'information dont il dispose sous forme de fichiers de formats variés (voir la partie

versant information dans le chapitre Principaux services d'Internet) : textes, images, sons, séquences vidéo, programmes informatiques ou données informatiques, etc…

Ainsi, à l'instar d'une bibliothèque qui met à la disposition des lecteurs qui s'y rendent l'ensemble de ses ouvrages destinés au public, Internet offre à tout ordinateur connecté la possibilité d'accéder à l'ensemble des fichiers pour lesquels l'administrateur de chaque réseau a autorisé l'accès au public, sans condition ou sous certaines conditions (identification de l'utilisateur demandant l'accès par des moyens divers). Et de même qu'une bibliothèque ne possède pas tous les livres existants et ne met pas forcément à la disposition de tout public tous les ouvrages qu'elle possède, Internet ne contient pas l'intégralité du savoir humain codifiable (puisqu'il faut en passer par une procédure de codification qui n'est pas systématique ni immédiate et encore moins sans générer de coût en termes de temps et d'argent), ni l'intégralité du savoir humain codifié (puisque la codification d'un savoir ne conduit pas forcément à un document sous forme utilisable par l'informatique), et un utilisateur quelconque n'a pas forcément accès à toutes les ressources disponibles sur Internet. L'accès à certaines ressources peut ainsi être soumis au paiement d'un certain droit de consultation, ou encore à la présentation par l'utilisateur d'une certaine authentification (par exemple, il faudra justifier de son appartenance à l'Ordre des Médecins pour accéder à certaines informations concernant la santé).

Une autre caractéristique de l'Internet est la redondance de l'information qui y est contenue. Comme l'Internet est une structure décentralisée, chaque réseau formant un de ses nœuds possède sa propre règle de fonctionnement. une même information peut donc être présente au sein de plusieurs réseaux, et souvent sous des formes légèrement différentes, rendant difficile l'élimination de ces redondances par des processus automatiques simples.

Il serait ainsi plus juste de représenter Internet comme l'aggrégation de milliers de bibliothèques différentes dont chacune aurait son propre catalogue, l'avantage étant que les contraintes géographiques n'existent plus et que tout se passe comme si toutes étaient réunies au même lieu.

2. Principaux services d'Internet

Internet offre plusieurs services à ses usagers. Pour y avoir accès, l'usager doit disposer des logiciels adéquats : autant de fonctionnalités, autant de logiciels. Internet est en évolution constante et il arrive fréquemment que certaines ressources d'apparition récente nécessite l'usage de logiciels plus récents que ceux en possession de l'utilisateur. C'est d'ailleurs une problématique générale à l'ensemble du domaine de l'informatique. Les logiciels les plus évolués vont vers une intégration des différentes fonctionnalités et le tout devient graduellement plus convivial. Jusqu'à présent néanmoins, pour la plupart des fonctions courantes d'Internet il existe des logiciels gratuits dont les versions les plus récentes (les mises à jour) restent gratuits.

Les services qu'offre Internet peuvent être divisés en deux catégories : ceux qui permettent à deux utilisateurs humains d'échanger des informations entre eux –

c'est le versant communication ; et ceux qui permettent à un utilisateur humain d'accéder à une information destinée à l'usage commun, et de la rapatrier sur son propre système –c'est le versant information.

2.1) Le versant communication

2.1.1) Le courrier électronique

Le courrier électronique (e-mail en Anglais et courriel dans le langage officiel) fonctionne de la même façon que le courrier postal, il n'y a que le moyen de transport qui change, avec tout de même des différences notables en termes de coût et de délais. Son utilisation nécessite un logiciel adéquat qu'on appelle « client e-mail ». Certains de ces logiciels sont payants, cependant il en existe qui sont gratuit et qui offrent autant de fonctionnalités que les premiers. Parmi ces logiciels gratuits, on peut citer Pegasus Mail (téléchargement sur http://www.pegasus.usa.com/) qui est très complet mais complexe à l'utilisation, ou Eudora (téléchargement sur http://www.eudora.com/) qui représente un bon compromis entre richesse fonctionnelle et simplicité d'utilisation. Il existe un site dédié à tous les aspects de ce service d'Internet : http://www.arobase.org/

Comme avec le courrier postal, l'expéditeur commence par rédiger un message à l'aide du logiciel. Ensuite, l'envoi du courriel peut se faire par l'intermédiaire de la connexion Internet : le courriel est « encapsulé » avec tous ses paramètres d'acheminement et transféré vers un ordinateur du fournisseur d'accès, depuis lequel il est acheminé vers tous les destinataires spécifiés lors de l'encapsulation, et arrive dans les boîtes aux lettres électroniques des destinataires qui sont situés au niveau du réseau de fournisseur d'accès de chaque destinataire. Et, à l'instar du courrier postal, le destinataire ne prend connaissance du courriel que lorsqu'il décide de consulter sa boîte aux lettres électronique par l'intermédiaire de sa propre connexion Internet.

Les choses se passent donc de façon très similaire à un envoi postal. L'informatique et les télécommunications apportent cependant quelques améliorations qui sont notables et qui sont détaillées ci-dessous.

Le délai d'acheminement : il est de l'ordre de la minute, entre le moment où le courriel est expédié et le moment où il arrive dans la boîte aux lettres électronique du destinataire.

La fonction de relevé automatique de courriel : la plupart des logiciels de courriel comportent une fonction de relevé automatique de la boîte aux lettres : le logiciel relève automatiquement la boîte aux lettres de l'utilisateur de façon périodique, la durée de cette période étant réglable de façon très souple par l'utilisateur.

Il résulte de ces deux caractéristiques que le courriel ressemble beaucoup au téléphone pour la rapidité (qualité apportée par son véhicule : les télécommunications), tout en restant aussi discret que le courrier (possibilité pour le destinataire de prendre connaissance de son message au moment qu'il choisit), et en apportant la richesse documentaire d'un imprimé voire d'un document multimédia (texte imprimé mais aussi images voire séquence audio ou vidéo en pièce jointe : qualité inhérente au traitement informatique de la fonctionnalité).

Par ailleurs, le coût d'envoi d'un tel courriel est minime : l'envoi de l'équivalent d'une page A4 de texte prend moins de trente secondes soit moins de 13 centimes au tarif téléphonique actuel de 0,25 Frs par minute, et ce coût est le même lorsqu'un même courriel est envoyé à plusieurs destinataires (quel que soit le nombre de destinataires) et quelle que soit l'éloignement du destinataire.

2.1.2) Les services dérivés du courrier électronique

2.1.2.1) Les listes de diffusion

La plupart des logiciels de courriel offrent à l'utilisateur la possibilité de constituer des listes de destinataires: cette fonction évite d'avoir à retaper toutes les adresses des destinataires lorsqu'on prévoit plusieurs envois de courriels à un même liste de destinataires: il suffit de regrouper les destinataires sous une liste et de la nommer; une fois cette manoeuvre accomplie il suffit de préciser le nom de la liste dans la rubrique « destination » du courriel pour que toutes les adresses des destinataires soient spécifiées au moment de l'encapsulage du courriel. C'est un progrès de plus dans la commodité des communications entre personnes, parfois très éloignées les unes des autres.

Par exemple: on peut constituer une liste nommée "Associés" et y faire figurer les adresses de tous ses associés:

claude.bernard@academie-medecine.org
sigmund.freud@psychanalysts.com
florent.coste@chu-cochin.fr

un courriel envoyé à "Associés" sera automatiquement envoyé aux trois destinataires dont les adresses figurent ci-dessus.
On peut également définir une autre liste nommée "Famille" comprenant

frère1@amiens.fr
frère2@ottawa.ca
soeurette@london-city.uk
oncle1@sydney.au
etc...

qui servira pour tous les courriels qui seront envoyés aux membres de la famille.

L'inconvénient de ce système réside dans le caractère excentré et statique de la liste: il faudrait que chaque destinataire maintienne dans son propre logiciel de courriel la même liste pour pouvoir envoyer à son tour un courriel aux autres membres de la liste; chaque ajout ou retrait d'une adresse devrait aussi être exécuté par chacun des membres de la liste sur son propre logiciel de courriel. Cette façon de procéder devient à la longue lourde et fastidieuse, surtout lorsque le nombre des membres est élevé comme cela peut être le cas par exemple pour une liste regroupant des médecins d'un même département par exemple.

Cette gestion de la liste des destinataires est une des fonctions qu'offrent les serveurs de listes: ce sont des entreprises qui proposent, en général en contrepartie de l'insertion d'un message publicitaire au début ou à la fin de chaque message, de s'occuper de toute l'intendance des mises à jour au quotidien des listes. Ainsi, la liste créée est nommée puis domiciliée chez le serveur de liste, et il suffit pour tout envoi

collectif de préciser l'adresse fournie par le serveur de liste au moment de son enregistrement par le propriétaire de la liste.

Exemple: afin de constituer un organe qui permet aux mêmes médecins de communiquer de façon simple et instantanée entre, et d'échanger sans devoir se déplacer des documents multimédia, un organisateur volontaire se connecte à Internet et se rend sur le site d'un serveur de listes, comme par exemple http://www.egroups.fr . Ensuite, il procède à l'enregistrement de la liste en lui donnant un nom, par exemple "medecins92". Une fois la liste enregistrée, il peut soit inscrire des destinataires d'office, soit les encourager à venir le faire eux-mêmes en allant sur le site http://www.egroups.fr . Désormais, chaque courriel adressé à l'adresse medecins92@egroups.fr sera ensuite réacheminé vers toutes les adresses contenues dans la liste qui est gérée par le serveur du site: inscriptions et désinscriptions se font de façon centralisée et instantanée. Ainsi s'est créée une liste de diffusion, ainsi nommée parce qu'elle permet de diffuser à une large audience tout message considéré comme d'intérêt général.

Outre cette gestion centralisée des inscriptions et désincriptions, la domiciliation chez un serveur de listes permet un ensemble d'autres fonctions, variable selon le serveur considéré, mais qui peuvent consister en:

- Le contrôle des inscriptions: le propriétaire de la liste peut décider de soumettre toute inscription d'un nouveau membre à son approbation préalable. Elle peut également décider la radiation d'un membre de la liste.
- La modération des messages: le propriétaire de la liste peut choisir l'option qui consiste à subordonner à son approbation tout message envoyé à la liste, en lui soumettant au préalable le contenu de chaque message destiné à la liste, dont il devient ainsi le modérateur.
- L'archivage des anciens messages: si cette option est choisie, le serveur gardera en archive tous les messages qui ont été envoyés par son intermédiaire, permettant ainsi aux membres qui se sont inscrits à une date postérieure à celle de la création de la liste d'avoir accès aux messages postés avant leur inscription.
- La mise en commun de fichiers de tous types: c'est une fonction qui permet d'éviter l'attachement systématique de fichiers volumineux aux courriels.

2.1.2.2) Les forums de discussion

Il existe sur Internet des serveurs qui s'occupent de stocker et de mettre à la disposition de tout internaute qui en fait la demande l'ensemble des messages à caractère public envoyés sur Internet par d'autres internautes : ce sont les serveurs de news. L'ensemble de ces serveurs de news est connu sous le nom d'Usenet et tous communiquent en utilisant le protocole NNTP.

Tout se passe comme si Usenet était composé d'un grand nombre de panneaux d'affichage (virtuels) où chacun peut venir écrire son message, qu'il s'agisse d'une déclaration, une question à la cantonade ou une réponse à une question posée. Les panneaux sont classés selon un procédé hiérarchique qui détaille les catégories et sous-catégories jusqu'au groupe qui comprend l'ensemble des messages sur un sujet donné.

Ci-dessous, un extrait de la hiérarchie .fr :

…
fr.announce.seminaires
fr.bienvenue
fr.bienvenue.questions
fr.bio.biomol
fr.bio.general
fr.bio.genome
fr.bio.logiciel
fr.bio.medecine
fr.bio.medecine.veterinaire
fr.bio.paramedical
fr.bio.pharmacie
fr.biz.d
fr.biz.produits
…

Bien que les messages soient envoyés sur Usenet en utilisant le protocole NNTP (et non le SMTP comme pour l'envoi des courriels), tout se passe pour l'utilisateur comme s'il envoyait un courriel, et il a d'ailleurs la possibilité d'envoyer un courriel privé à un des intervenants d'un groupe. C'est pourquoi on peut considérer que ce service est assimilable à un service dérivé du courriel.

L'avantage de ce mode de fonctionnement réside dans la taille importante du public potentiel auquel s'adresse un message : les questions les plus inattendues trouvent parfois des réponses venant d'intervenants de qualité inconnus jusque là de l'utilisateur. L'inconvénient qui découle de cette large audience potentielle est le risque de recevoir un très grand nombre de réponse à sa question, parmi lesquelles une grande majorité est dénuée d'intérêt.

2.1.3) Le dialogue en direct (IRC)

Les ordinateurs connectés échangeant les données en temps réel, il est possible pour plusieurs internautes connectés au même moment de converser entre eux grâce à l'utilisation d'un logiciel dédié: c'est ce qui est communément désigné sous le terme de « chat », terme anglais signifiant « bavarder » ou encore d'IRC (pour Internet Relayed Chat).

Sur le plan technique, le logiciel, une fois lancé, se charge de détecter la connexion à Internet de l'ordinateur sur lequel il est implanté et dès cette connexion établie, envoie un signal à un serveur centralisé comportant l'identification (codée) de l'utilisateur, puis récupère sur ce serveur le statut de tous les identifiants connus (qui lui auront été préalablement indiqués par l'utilisateur) : connecté ou déconnecté. La mise à jour se fait de façon automatique et tous les internautes connectés au serveur qui auront indiqué l'identifiant de celui qui vient de se connecter seront avertis de l'arrivée de celui-ci sur le réseau.

Ainsi, chaque internaute utilisant le logiciel aura en permanence sous les yeux le statut de ceux qu'il connaît : connecté ou déconnecté. Lorsqu'une personne est connectée, il devient possible de lui envoyer un message et d'engager la conversation, les échanges de messages se faisant en temps réel. La conversation engagée sera alors semblable à une conversation téléphonique, à ceci près que les échanges se font par écrit, et qu'il est possible d'envoyer à son correspondant toutes sortes de documents numériques, depuis les sons, les images jusqu'aux URL.

Une autre fonction offerte par ces logiciels est la fonction « conférence », où un grand nombre d'internautes connectés peuvent engager une conversation à plusieurs, chaque message étant envoyé simultanément à tous les membres de la conférence en temps réel, permettant notamment la tenue de réunion virtuelles entre participants ne pouvant se trouver physiquement ai même endroit pour des raisons diverses (par exemple géographiques).

2.2) Le versant information

En informatique, les informations sont regroupées en unités structurées appelées fichiers. L'Internet étant un espace entièrement régi par l'informatique, les informations qui sont mises à disposition par son intermédiaire se présentent essentiellement sous forme de fichiers, parfois sous forme d'un ensemble de fichiers : c'est alors une base de données.

2.2.1) Les informations multimédia disponibles sur Internet

2.2.1.1) Les fichiers texte

Historiquement parlant, la première ressource à être mise à disposition sur Internet était la reproduction sous forme électronique de textes écrits. Historiquement parlant également, la transmission du savoir s'effectuant de manière stable (sans déformation majeure) se faisait également par l'écrit. C'est ce qui explique sans doute pourquoi l'information qui circule sur Internet comprend une grande partie de textes. L'autre explication tient probablement au fait qu'en dépit de la forte croissance des débits qu'ont connu les infrastructures d'Internet, les autres

ressources d'informations nécessitent à volume informationnel égal le transport d'une quantité de données beaucoup plus importante, entraînant des latences de réponse non négligeables au vu des capacités actuelles d'Internet.

Un fichier texte est un ensemble structuré d'informations lisible par un logiciel de traitement de textes. Cependant, avec la complexification des logiciels et l'ajout de nombreux enrichissements (polices de caractères, graisses et styles des caractères, formats de mise en page, etc…), chaque logiciel possède actuellement sa propre façon de coder un fichier de texte, et le fichier ainsi formé est rarement lisible par un autre logiciel.

Il est donc nécessaire le plus souvent de posséder le même logiciel que celui qui a servi à produire le fichier texte que l'on veut consulter. Les logiciels de traitement de texte les plus usités à l'heure actuelle sont Word ou Works sous Windows pour les utilisateurs de PC, et Word ou AppleWorks pour les utilisateurs de Macintosh.

Cependant, si l'on excepte les enrichissements de texte qui l'agrémentent mais sans en changer le sens, à chaque caractère de l'alphabet occidental correspond un code informatique universel, reconnu par quasiment tous les logiciels : il donc toujours possible de convertir un fichier texte que l'on veut mettre à disposition du plus grand nombre en un fichier au format standardisé qui sera lisible par n'importe quel logiciel (au prix de quelques pertes d'enrichissement) : le fichier texte brut (extension .txt) ou le fichier Rich Text Format (extension .rtf) un peu plus évolué.

Enfin, la société Adobe commercialise un logiciel nommé Adobe Acrobat qui permet de produire des documents visuels (écrits +/- enrichissements graphiques) dans un format lisible sur toutes les machines par un logiciel distribué gratuitement : Adobe Acrobat Reader.

2.2.1.2) Les fichiers images

La numérisation a beaucoup progressé et désormais il est possible avec un appareillage simple et relativement peu coûteux de générer un fichier informatique qui contient la représentation d'une image, laquelle pourra être visualisée grâce à un logiciel spécifique de visualisation ou de traitement d'images.

Le codage de cette représentation a également beaucoup progressé et il existe désormais des algorithmes de compression d'image qui permettent de diminuer considérablement le volume de données nécessaire à la représentation d'une image sans en altérer sensiblement la qualité. Il est devenu ainsi possible de faire transiter à travers Internet des images de dimensions tout à fait correctes (format A5 par exemple) avec des délais acceptables (de quelques secondes à moins d'une minute dans la plus grande majorité des cas).

Ainsi, les algorithmes de compression les plus populaires sont ceux qui génèrent les formats GIF (extension .gif) et JPEG (extension .jpg ou .jpeg). Ces deux

formats sont lisibles par la plupart des logiciels de visualisation ou de traitement d'image, ainsi que les navigateurs usuels.

2.2.1.3) Les fichiers audio

Le codage sous forme numérique d'une séquence sonore existe depuis déjà un certain temps, et comme toute donnée numérique il était possible de les faire transiter d'un endroit à un autre par l'intermédiaire d'Internet. Toutefois, pendant longtemps les algorithmes de codage étaient conçus de telle façon qu'une minute de séquence audio nécessitait pour sa représentation sous forme numérique pas moins d'environ 10 méga-octets (Mo). Faire transiter un tel volume de données par Internet avec les débits d'alors était difficile à envisager, puisqu'il fallait plus de 10 voire 20 minutes de téléchargement pour disposer d'une minute d'enregistrement audio.

Plus récemment, de nouveaux algorithmes ont fait leur apparition, permettant d'encoder des séquences audio qui jusque là étaient trop volumineuses pour pouvoir transiter sur Internet avec des délais raisonnables.

L'algorithme le plus populaire pour cette compression des données est celui du MPEG-1 Layer 3, plus connu sous le nom de MP3 (extension de fichier .mp3). Afin d'en avoir un ordre de grandeur, il faut savoir qu'une minute de séquence audio à une qualité proche de celle produite par une chaîne HiFi grand public œcupe environ 1 Mo en format MP3. Lorsque la qualité HiFi n'est pas requise, le volume des données pourra encore être diminué d'un facteur 3-4 supplémentaire.

Récemment, un algorithme dérivé du MP3 vient d'être annoncé qui réduit encore d'un facteur 2 la taille du fichiers obtenu sans altération supplémentaire notable de la qualité sonore : le MP3 Pro.

Cependant, sachant que le débit habituel d'une liaison téléphonique est d'environ 46 kbps et que la vitesse de téléchargement excède rarement les 4-5 Ko/s (sauf pour les connexions à haut débit), il faudrait encore entre 1 et 2 minutes pour télécharger une séquence audio d'une minute. Même en utilisant des artifices de téléchargement qui permettent de réduire de 50-65% le temps nécessaire, il reste qu'il faut patienter de longues minutes avant de pouvoir entendre la séquence audio.

Un autre procédé a été mis au point afin de contourner cette difficulté : la diffusion en « streaming ». L'astuce consiste à exploiter le mode de fonctionnement multitâche préemptif de l'ordinateur (disponible en standard sur la plupart des ordinateurs équipés de systèmes d'exploitation récents) et à remplir une mémoire « tampon » (buffer en anglais) qui contient les données nécessaires à l'audition de quelques secondes de séquence audio, puis à lire cette mémoire-tampon pour produire la séquence audio en temps réel, tout en continuant pendant la production des sons à télécharger la suite : la mémoire-tampon se vidant ainsi d'un côté au fur et à mesure du déroulement de la séquenœ audio, mais se remplissant en même de l'autre grâce au téléchargement simultané continu. Un réglage est éventuellement nécessaire pour obtenir le bon compromis entre la qualité du son produit et le volume de la mémoire-tampon qui dépend de la vitesse de téléchargement. Les standards pour ce mode de diffusion sont essentiellement le RealAudio et le WMA.

Cette diffusion de données audio en mode streaming permet d'avoir sur l'ordinateur (pour peu que l'utilisateur ait l'équipement adéquat : carte-son et haut-parleurs) exactement les mêmes fonctions qu'avec une radio sur ondes hertziennes, à cette différence près qu'il est possible pour chaque utilisateur de choisir lui-même le contenu qu'il veut écouter et le moment où ce contenu est diffusé, avec rediffusion possible à la demande.

La compression audio est un domaine en évolution constante et rapide actuellement et il apparaît régulièrement de nouveaux formats. Il est donc difficile de prédire quel format s'imposera à l'avenir et deviendra le standard universel.

2.2.1.4) Les fichiers vidéo

Le codage des séquences vidéo existe lui aussi depuis un certain temps, mais pour l'heure force est de constater que le volume de données reste lourd et le débit de l'Internet est encore trop lent pour autoriser l'échange de longues séquences vidéo.

Il existe plusieurs algorithmes de codage vidéo. L'objectif de ce travail n'étant pas d'en traiter les subtilités, citons simplement les plus répandus qui sont MPEG-1, MPEG-2 et Quicktime, ce dernier étant le seul format lisible par les plate-formes PC et Macintosh. A titre indicatif, le format utilise environ 10 Mo pour coder 1 minute de séquence vidéo de qualité équivalente à celle d'une TV classique.

A l'instar de la diffusion audio en mode streaming, c'est le format RealVideo qui est le plus répandu à l'heure actuelle pour la diffusion en mode streaming des séquences vidéo. L'autre format en concurrence avec RealVideo est le VDOLive.

Comme toujours, la lecture de telles séquences vidéo (qui sont des fichiers informatiques produits dans un standard donné) nécessite les logiciels adéquats, sachant cependant qu'il en existe un certain nombre qui sont soit gratuit soit fournis avec le système d'exploitation de l'ordinateur, tant le multimédia est en passe de se banaliser dans l'usage quotidien de l'informatique.

La compression vidéo est également un domaine en évolution constante et rapide actuellement et il apparaît régulièrement de nouveaux formats. Il est donc difficile de prédire quel format s'imposera à l'avenir et deviendra le standard universel.

2.2.2) Les bases de données bibliographiques

Si les revues médicales cliniques (« Le Concours Médical », « Prescrire », etc...) offrent le plus souvent de bonnes synthèses sur un sujet donné, les publications scientifiques représentent la documentation primaire essentielle sur laquelle le praticien peut fonder sa réflexion et dans laquelle il pourra puiser de la matière pour nourrir sa propre recherche.

Les bases de données bibliographiques sont une ressource précieuse à ce titre qu'elles permettent une vision d'ensemble de la littérature scientifique. Elles sont le résultat d'une procédure d'indexation manuelle par des opérateurs humains. Certains d'entre elles sont d'accès gratuit, d'autres sont payantes. Aucune ne couvre l'ensemble des publications scientifiques existantes.

La plupart du temps (surtout s'agissant des bases de données bibliographiques à accès gratuit), une requête à une telle base de données retourne des résultats sous forme de fiches d'indexation donnant les références des articles correspondant aux critères de la requête, et il est possible de consulter le résumé de l'article tel qu'il a été rédigé par ses auteurs. L'accès au texte intégral de l'article est presque toujours subordonné au paiement d'un droit supplémentaire. Une fois le paiement (qui peut se faire par l'intermédiaire d'Internet) effectué, le contenu de

l'article peut être envoyé sous forme matérielle (version papier) et/ou sous forme électronique (version numérique : un fichier texte).

La particularité de ces bases de données bibliographiques est de ne pas être accessible à partir d'un moteur de recherche généraliste: elles font partie de ce qu'il est convenu de qualifier par « web invisible ». Il est donc nécessaire de les consulter selon un mode de recherche par navigation.

2.2.3) Les pages HTML et le Web

Afin que la consultation des différentes ressources mises à disposition (les fichiers multimédia et fiches des bases de données cités plus haut) se fasse de manière transparente (ie aussi simple que possible pour l'utilisateur) et unifiée (ie sous une interface unique), il existe un standard de mise à disposition qui est (à quelques détails près) commun à tous les ordinateurs et à tous les logiciels de navigation : le HTML (sigle de HyperText Markup Language).

Comme l'indique son nom, HTML fut au départ un langage de description d'affichage qui permet la navigation par le concept de l'hypertexte : une page écrite en HTML et lue par un navigateur se présente comme une page de journal ou de livre, avec du texte agrémenté d'images (voire de sons pour certaines et si l'ordinateur dispose de périphériques sonores), et quelques zones où le simple fait de cliquer avec la souris provoque automatique l'affichage dans la même fenêtre ou une nouvelle fenêtre du navigateur une autre page HTML qui peut être située n'importe où dans l'ensemble des pages disponibles sur l'Internet. L'utilisateur n'a ainsi plus le souci de se préoccuper de la localisation des différentes pages qu'il est amené à consulter lors d'une séance de «surf ». Cette zone spéciale et les informations qu'elle contient est désignée sous le nom de « lien hypertexte ».

L'ensemble des pages HTML accessibles est désigné sous le nom de « World Wide Web », ou plus communément le « Web ».

2.2.4) Le world wide web

Au départ, le world wide web ne devait constituer qu'une partie d'Internet, mais sa simplicité d'utilisation lui a permis de s'imposer progressivement, au fur et à mesure qu'un nombre croissant de gens y faisait appel et qu'en réponse de plus en plus de ressources y étaient mises à disposition. On peut estimer qu'à l'heure actuelle il n'est quasiment plus jamais nécessaire d'explorer d'autres espaces d'Internet que le Web pour ce qui est du versant information.

2.2.4.1) Structure du Web : le HTML et la notion d'URL

Afin de regrouper toutes les ressources disponibles sur le Web sous une forme qui soit centralisée (souci de simplicité) et identique quel que soit l'ordinateur utilisé (souci de standardisation), le langage HTML (pour HyperText Markup Language) a été développé puis remanié plusieurs fois depuis sa création (nous en

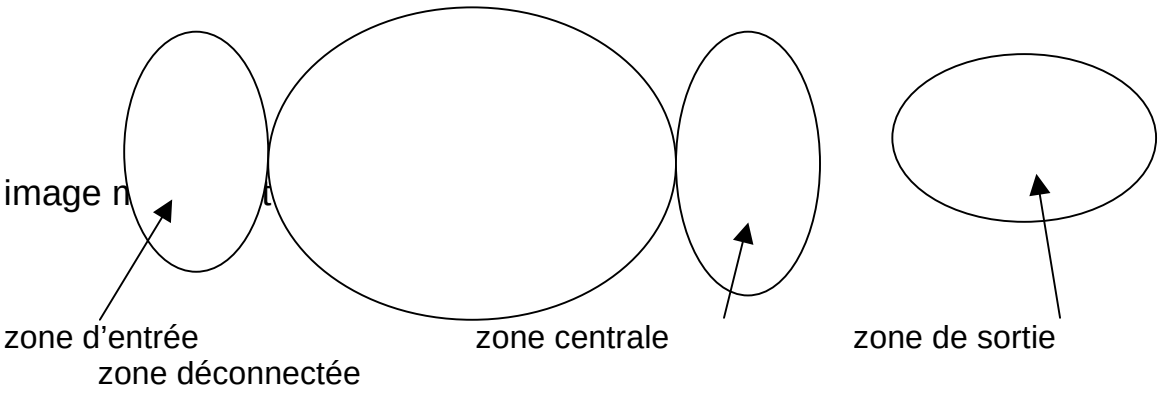
sommes actuellement à la version 4), incluant de plus en plus de fonctionnalités jusqu'à englober désormais la quasi-totalité des ressources disponibles sur Internet.

Le concept central de ce langage est la notion de « lien hypertexte » : c'est un lien virtuel qui relie deux documents différents au moyen de deux « ancres » : l'ancre de départ est matérialisée par une zones définie dans le document HTML affiché par le navigateur, l'ancre d'arrivée est représentée par un autre document vers lequel pointe ce lien hypertexte. Le fait de cliquer avec la souris sur l'ancre de départ provoque l'affichage par le navigateur du document représenté par l'ancre d'arrivée, et ce quel que soit l'éloignement de ce dernier document. Si à aucun moment l'utilisateur n'a eu à préciser les références de ce document vers lequel pointe le lien hypertexte, c'est grâce à l'usage de l'URL (pour Uniform Resource Locator) qui attribue à chaque document HTML une «adresse » unique, reprise dans le lien hypertexte.

Les pages du Web sont donc pour chacune repérées par une adresse unique : l'URL, et contiennent en leur sein des lien hypertextes pointant vers certaines de ces adresses (donc d'autres pages HTML). Le tout forme un réseau gigantesque et planétaire dont la topographie reste encore mal connue, et qui fait l'objet d'études actuelles.

2.2.4.2) La topographie du Web

Certains auteurs ont tenté d'approcher cet espace virtuel sous l'angle topographique, notamment des scientifiques des équipes de recherche d'Alta Vista, Compaq et IBM, dont les études sur plus de 200 millions de pages HTML ont abouti à une représentation similaire à l'image d'un noeud papillon, plus une zone déconnectée :



La zone d'entrée est composée de pages contenant beaucoup de liens pointant vers celles de la zone centrale.

La zone centrale est composée de pages contenant essentiellement des liens pointant vers d'autres pages de la même zone. Selon certaines évaluations, la richesse des liens est telle que lorsqu'on se trouve dans cette zone centrale, il suffit d'explorer et donc de cheminer le long de sept liens en moyenne pour retomber sur la même page.

La zone de sortie est composée de pages contenant peu de liens, représentant donc des « impasses » virtuelles.

La zone déconnectée est composée de pages accessible uniquement en tapant directement l'URL dans la fenêtre du navigateur, et ne contient pratiquement pas de lien pointant vers d'autres pages.

Il existe donc une partie du Web d'accès difficile puisqu'il faut en connaître l'URL exact afin de pouvoir y accéder.

2.2.4.3) La taille du web

Eu égard à la facilité et au coût minime de la publication sur le Web, les pages HTML se sont multipliées de façon exponentielle depuis sa création :

Début 1998 : Steve Lawrence et C. Lee Giles publient dans Science leur estimation de la taille du Web à 320 millions de pages à la date de décembre 1997 (3).

Le 08/07/1999 : Steve Lawrence et C. Lee Giles publient dans Nature que le Web grand public comporte 800 millions de pages selon leur nouvelle estimation en février 1999 (4).

Le 18/01/2000 : Inktomi et Nec Research Institute publient une étude selon laquelle le Web compterait 1 milliards de pages (5).

Le 11/07/2000 : la société Cyveillance évalue la taille du Web à plus de 2 milliards de pages, avec une croissance de 7 millions de pages supplémentaires par jour (6).

La méthode utilisée par Steve Lawrence et C. Lee Giles est une méthode statistique, qui utilise l'extrapolation à partir de requêtes: leur première estimation était obtenue à partir d'un pannel de requêtes aléatoires exploitant les moteurs de recherche, la deuxième estimation à partir d'un pannel d'adresses IPv4 aléatoires. Dans les deux cas, l'estimation des pages porte sur la taille du Web « indexable », c'est à dire l'ensemble des pages qui sont accessibles aux moteurs de recherche. Cependant, il existe par ailleurs une partie non négligeable du Web qui n'est pas accessible aux moteurs de recherche : ce sont notamment les bases de données accessibles par formulaire de recherche et les fichiers textes sous une autre forme que le HTML.

2.2.4.4) Le « Web invisible »

Cette partie du web qui est non accessible aux moteurs de recherche communs est évaluée à environ 40% du total des pages web (7) (certains l'évaluent à bien plus mais de par sa définition son étendue exacte est non mesurable) et est communément désignée sous le nom de « web invisible » ou « web profond » (deep web en Anglais). Bien qu'il existe de plus en plus de ressources logicielles qui tentent de donner à l'utilisateur un accès le plus exhaustif possible à ce web invisible, on peut considérer qu'il est nécessaire pour accéder au web invisible de procéder par navigation.

Annexe web invisible (cf Medline, Nosobase, Pediadol, biliothèque Cochrane notamment)

3. Principaux enjeux d'Internet

3.1) La manipulation de l'information

Le terme « manipulation » revêt ici les deux facettes de son double sens : celui du maniement objectif, c'est à dire la production, la reproduction, le stockage et la diffusion de l'information ; mais aussi celui du dévoiement plus ou moins délibéré de l'information afin d'induire en erreur un public cible plus ou moins étendu.

3.1.1) Le maniement objectif de l'information

3.1.1.1) La normalisation

L'abondance de l'information est devenue telle que l'enjeu s'est désormais déplacé du souci de sa production vers celui de la restitution de l'information adéquate, posant ainsi le problème de son indexation.

Jusqu'à l'apparition de l'informatique, cette indexation était assurée par des humains, suivant un processus d'indexation sémantique (a). De façon analogue, il existe sur Internet des annuaires (b) qui sont des outils d'indexation mis en place et alimentés par des opérateurs humains, mais la croissance exponentielle du volume d'informations de l'Internet a rapidement abouti à un allongement considérable du délai entre la publication d'un document et son indexation par un opérateur humain, justifiant ainsi pour des raisons d'exhaustivité et de fraîcheur des informations le recours de plus en plus fréquent aux moteurs de recherche, qui reposent sur un indexation lexicale (c) automatique assurée par des logiciels informatiques, avec les avantages d'une très grande rapidité, une reproductibilité parfaite, et une absence de besoin de périodes de repos permettant l'indexation d'une bien plus grande quantité de pages par unité de temps. En revanche, à la différence de l'indexation sémantique, cette indexation lexicale qui repose sur le principe d'une comparaison caractère à caractère est souvent pris en défaut par la polysémie de la langue usuelle et pèche par la faible pertinence (d) des résultats fournis.

Une tentative pour améliorer la qualité du travail d'indexation des moteurs de recherche a été entreprise par le biais de la définition des « méta-données » : le rédacteur de la page web est ainsi mis à contribution pour indexer lui-même sa page, suivant le principe de l'indexation sémantique. Malheureusement, comme l'utilisation des ces balises n'est pas réglementé, le désir d'apparaître aussi souvent que possible dans les résultats de recherche a conduit certains auteurs indélicats à inclure des mots-clefs n'ayant aucun rapport avec le contenu de sa page dans les balises méta. Cette tendance est devenue si envahissante que les concepteurs des moteurs de recherche les plus récents ont fini par abandonner le recours aux balises méta.

Cependant, il est important de connaître et surveiller les initiatives qui traitent de ce sujet, dans l'espoir que dans un avenir proche la discipline redevienne la règle et qu'une indexation faisant appel aux méta-données redevienne possible. Il faut citer dans ce domaine le Dublin Core Metadata Initiative (DCMI) (http://dublincore.org) dont les travaux servent de référence au CISMeF qui lui même fait autorité en matière d'Internet médical en France.

3.1.1.2) Les droits de la propriété intellectuelle

Comme toute oeuvre originale, les documents publiés sur Internet (sous forme numérique) devraient en toute logique être soumis aux mêmes droits de propriété intellectuelle que les documents matériels. Malheureusement, la rapidité, la modicité et la facilité du processus de reproduction des données numériques quelles qu'elles soient sont

(a) L'indexation sémantique consiste à indexer le document en fonction de la signification de son contenu.
(b) L'annuaire web est un site web où les URL sont classés en fonction d'une hiérarchisation arborescente.
(c) L'indexation lexicale consiste à indexer le document en fonction des mots présents dans son contenu.
(d) La pertinence d'un document est caractérisée par sa capacité à répondre à la question posée par l'utilisateur.

aussi mises à profits par certains pour effectuer des copies, parfois à visée commerciales, parfois à visée plagiaire. Outre le fait qu'il est pour le moment très difficile de restreindre les possibilités de copies des documents électroniques, notamment à cause des logiciels de décryptage ou de contournement de codes anti-copie, force est de constater que la législation en matière de droit de propriété intellectuelle des documents numériques n'en est qu'à ses balbutiements et que ce vide juridique peut constituer un frein à certaines publications sous forme numérique, restreignant ou retardant ainsi la possibilité d'accès par Internet à la totalité des informations publiables.

Néanmoins, la culture Internet reste encore majoritairement non-commerciale, en raison de sa filiation puisqu'à l'origine créé et animé essentiellement par des scientifiques qui fonctionnent beaucoup selon la tradition de coopération et de mise à disposition non-onéreuse des informations.

La médecine, étant à la frontière entre science pure et technique appliquée, mais aussi entre préoccupations de soins (de la part des praticiens) et visées commerciales (de la part des laboratoires pharmaceutiques), est tiraillée entre ces deux logiques opposées que sont la publication à titre gratuit des informations et la rémunération par le droit de la propriété intellectuelle. En témoigne par exemple l'accès payant au contenu des articles de recherche clinique. Cependant, la tendance actuelle se dessine vers une mise à disposition à titre gratuit par les grands éditeurs d'articles ayant une certaine ancienneté.

3.1.2) La manipulation non-objective de l'information

La publication de l'information sur support papier est désormais bien structurée avec des instances de contrôle de sa pertinence et de sa véracité. Le coût d'une telle publication est par ailleurs non négligeable, obligeant chaque auteur à un minimum de réflexion avant de se lancer dans l'aventure. Des filtres sont ainsi bien en place pour assurer une relative fiabilité de l'information sur support papier. Il en est tout autrement pour l'information publiée sur Internet.

Les coûts de reproduction et de stockage numériques étant minimes, chaque particulier a la possibilité de publier sur Internet les informations qu'il désire, quelle que soit sa véracité ; et Internet étant un média encore relativement jeune, il n'existe pas encore d'instance de hiérarchisation de la fiabilité des contenus qui y sont publiés qui soit aussi bien établie que sur support papier. S'y côtoient ainsi, par exemple, les articles des plus prestigieuses revues de médecine clinique et les fiches

à la rigueur parfois plus que contestables rédigées par des particuliers, en passant par des contenus à visée commerciales des différentes firmes pharmaceutiques.

En ce contexte de grande porosité des instances de contrôle a priori de l'information, le contrôle de la fiabilité des informations publiées sur Internet doit se faire à l'heure actuelle de façon a posteriori, par l'utilisateur lui-même.

3.2) La fiabilité des informations publiées sur Internet

Ainsi que cela a été décrit précédemment, la facilité et la modicité qui caractérisent une publication sur Internet sont telles que les informations qui y figurent sont d'une exactitude et d'une rigueur très variables d'une source à l'autre, et les instances de contrôle sont pour le moment trop jeunes pour avoir la même assise que celle des publications papier. L'utilisateur est ainsi astreint à déterminer par lui-même quelle confiance il peut accorder aux informations qu'il est parvenu à trouver sur Internet.

Le contrôle a posteriori de la fiabilité de l'information récupérée sur Internet fait partir intégrante du processus d'acquisition de connaissances pour qui souhaite se documenter par le biais d'Internet, et doit devenir un réflexe systématique.

Une thèse de médecine a été soutenue sur ce sujet par de Dr Xavier Dumont en 1997 à l'université de Lille II (http://www.multimania.com/xdumont/info/index.htm (8)).

Cette vérification est aisée lorsqu'il s'agit de sources bien identifiées comme les sites web des éditeurs de revues médicales : les documents qui y sont publiées sont des versions numériques des publications écrites et bénéficient donc de la même fiabilité.

Elle est plus délicate dès qu'on sort de ce cadre et, dans le domaine médical, les tentatives de mise en place d'instances de contrôle de la fiabilité de l'information sur Internet ont vu l'émergence de trois principaux concepts : la charte Health On the Net (HON), le Netscoring, et en France le projet « Qualité des sites de e-santé » du Ministère de l'Emploi et de la Solidarité.

3.2.1) La charte HON

La fondation Health On Net (site www.hon.ch) a vu le jour pour essayer de trouver une solution à la question de la fiabilité des informations qu'on trouve sur les sites web. Il s'agit d'une fondation à but non lucratif dont le siège est à Genève (Suisse), qui propose notamment des outils de recherches dans les domaines médicaux et biomédicaux dont le moteur de recherche MedHunt et la charte « Health On Net ».

La charte « Health On Net », dénommée HONCode, est un document élaboré par la fondation du même nom, regroupant un ensemble de recommandations de bonne conduite qui n'ont aucun caractère obligatoire, conçues pour s'assurer que le lecteur connaisse toujours la source et le but des données qu'il consulte. En aucun cas ce n'est un outil destinée à évaluer la fiabilité des informations publiées. Ainsi, il ne s'agit que d'une charte destinée à «aider à unifier et normaliser la fiabilité des informations médicales et de santé sur le Web ».

Les sites manifestant leur adhésion à cette charte étant contrôlés par la fondation sur le respect des critères édictés dans la charte, la fiabilité de leur contenu peut être considérée comme relativement digne de confiance. Il ne faut toutefois pas perdre de vue que l'immensité du Web est telle qu'il est malgré tout impossible d'obtenir une vérification exhaustive et à intervalles rapprochés, c'est pourquoi la nécessité de l'évaluation de la justesse des informations fournies reste à la charge de l'utilisateur.

Les recommandations de la charte sont au nombre de 8 :

1. Tout avis médical fourni sur le site sera donné uniquement par du personnel spécialisé (diplômé) du domaine médical et des professionnels qualifiés, à moins qu'une déclaration explicite ne précise que certains avis proviennent de personnes ou d'organisations non médicales.

2. L'information diffusée sur le site est destinée à encourager , et non à remplacer, les relations existantes entre patient et médecin.

3. Les informations personnelles concernant les patients et les visiteurs d'un site médical, y compris leur identité, sont confidentielles. Le responsable du site s'engage sur l'honneur à respecter les conditions légales de confidentialité des informations médicales applicables dans le pays dans lequel le serveur (ainsi que les éventuels sites- miroir) est situé.

4. La source des données diffusées sur le site est explicitement citée avec, si possible, un hyperlien vers cette source. La date de la dernière modification doit apparaître clairement sur la page Web (par exemple: en bas de chaque page).

5. Toute affirmation relative au bénéfice ou à la performance d'un traitement donné, d'un produit ou d'un service commercial, sera associée à des éléments de preuve appropriés et pondérés selon le principe 4. ci-dessus.

6. Les créateurs du site s'efforceront de fournir l'information de la façon la plus claire possible, et fourniront une adresse de contact pour les utilisateurs qui désireraient obtenir des détails ou du soutien. Cette adresse (e-mail) doit être clairement affichée sur les pages du site.

7. Le support d'un site doit être clairement identifié, y compris les identitées d'organisations commerciales et non-commerciales qui contribuent au financement, services ou matériel du site.

8. Si la publicité est une source de revenu du site, cela sera clairement établie. Le propriétaire du site fournira une brève description de la règle publicitaire adoptée. Tout apport promotionnel ou publicitaire sera présenté à l'utilisateur de façon claire afin de le différencier de l'apport uniquement créé par l'institution gérant le site.

Les personnes activement impliquées dans l’organisation HON sont :

- M. Guy-Olivier SEGOND (Président d’Honneur)

Conseiller d’Etat, Suisse.

- Pr Jean-Raoul SCHERRER (Président Exécutif)

Ancien Directeur de la Division d’Informatique Médicale, Hôpital cantonal universitaire de Genève (Suisse)

- Pr Ron D. APPEL

Institut Suisse de Bio-informatique, Genève (Suisse)

- Dr Marion J. BALL

Professeur-adjoint de l’Ecole d’Infirmier de l’Université John Hopkins, vice-présidente de « First Consulting Group », et ancienne présidente de « International Medical Informatics Association » USA.

- Pr Jan VAN BEMMEL

Vice-président de l’Institut d’Informatique Médicale, Université ERASMUS, faculté de science et médecine, Rotterdam (Pays-Bas).

- M. Michel CARPENTIER

Ancien directeur général DG XIII de la Commission Européenne, Bruxelles (Belgique).

- Ing. Maria LAIRES

Coordinatricede l'EHTO Observatoire Européen de la Télématique dans le domaine de la Santé et de ses Réseaux de Sites Affiliés de Langues Nationales (SDALN), Lisbonne (Portugal).

- Dr Donald B. LINDBERG

Directeur à la National Library of Medicine, Bethesda (USA) et directeur du Bureau National de Coordination des Performances Informatiques et de Communications au « Office of Science and Technology Policy » (1992-1995).

- M. Jean-Claude PETERSCHMITT

Ancien directeur général à «Digital Equipment Corporation ».

- Dr Mathias TSCHOPP

Membre du comité consultatif, Genève (Suisse).

3.2.2) Le Netscoring

Allant encore plus loin dans cette direction, un groupe de travail a été formé comprenant des médecins, ingénieurs, bibliothécaires et juristes (dans le cadre de Centrale Santé, groupement professionnel destiné à réunir autour d'un projet fédérateur des centraliens intéressés par la santé et des professionnels de la santé), chargé d'élaborer « un ensemble de critères qui peuvent être utilisés pour évaluer la qualité de l'information de santé sur Internet ».

Au total, 49 critères ont été énoncés, répartis en 8 catégories : crédibilité, contenu, hyper-liens, design, interactivité, aspects quantitatifs, déontologie, et accessibilité. Chaque critère est pondéré en critère essentiel (noté de 0 à 9), critère important (noté de 0 à 6) ou critère mineur (noté de 0 à 3). Le total de ces critères donne le score global du site (avec un maximum de 312 points).

L'utilisateur dispose ainsi d'un véritable moyen quantitatif d'évaluer la qualité de l'information médicale qui lui est présentée.

Pour plus de précisions, se reporter à l'URL www.chu-rouen.fr/netscoring/

Cependant, l'objectif de ce travail étant essentiellement de déterminer les stratégies destinées à améliorer la pertinence des résultats de recherche, nous ne ferons qu'un survol de la question de la fiabilité des informations médicales. Par ailleurs, l'expérience montre que la plupart du temps, un praticien est à même d'apprécier avec une relative exactitude la qualité de l'information qu'on lui délivre.

3.2.3) Le projet « Qualité des sites de e-santé »

Lancé au printemps 2000, le projet « Qualité des sites de e-santé » initié sous la tutelle du Ministère de l'Emploi et de la Solidarité a pour objectifs :

- de dégager un certain nombre de règles qui pourraient permettre à l'internaute de se faire lui-même une opinion sur la qualité des sites qu'il consulte
- d'assurer à l'usager que les sites qui se réclament de ces règles les règles les respectent bien
- de favoriser la mise en ligne de contenus de qualité et de développer les services offerts en e-santé.

Les actions envisagées sont :

- inciter les promoteurs de sites à s'inscrire dans une démarche qualité
- mettre en place un espace d'information et de formation à destination des internautes
- à moyen terme, qualifier la qualité de l'information de santé.

Toutefois, la visite de l'URL www.sante.gouv.fr/htm/dossiers/qualite/sommaire.htm permet de se rendre compte que le projet est à l'arrêt depuis juin 2001, date à laquelle a eu lieu la dernière mise à jour du document.

3.3) La sécurisation des informations en transit sur Internet

A l'heure où la télé-médecine devient une réalité avec la première d'une opération d'un patient en France réalisée par une équipe de chirurgiens aux Etats-Unis (9) (le 07/09/01), et où la transmission des résultats d'examens complémentaires via Internet est d'actualité, la sécurisation des informations qui transitent sur Internet devient cruciale afin d'en assurer l'intégrité et la confidentialité, conformément aux impératifs de fiabilité des processus télé-médicaux et de respect du secret médical.

Cependant, c'est un enjeu qui sort complètement du cadre de ce travail et qui ne sera donc pas abordé ici.

B) Accéder à l'information sur Internet

1. La recherche par navigation

La recherche par navigation consiste à renseigner directement le navigateur en lui indiquant l'adresse URL de la page web à afficher: soit en la saisissant dans la fenêtre appropriée du navigateur, soit en cliquant sur un lien hypertexte qui contient cette adresse URL.

Hormis quelques rares occurrences où l'utilisateur connaît à l'avance l'adresse URL d'un site fréquemment visité (ex. www.nejm.com pour le site du New England Journal of Medicine), la multiplicité et la complexité des intitulés des adresses URL rendent nécessaire le recours à des sortes de «répertoires » qui font l'inventaire des sites ou pages web intéressantes avec leurs adresses URL exactes.

1.1) Les annuaires

Les annuaires sont le résultat de la systématisation de tels répertoires, dont l'ambition est de balayer l'ensemble (ou du moins la plus grande partie possible) du Web. Les annuaires sont constitués d'adresses URL de pages qui ont été visitées par des opérateurs humains, afin d'en connaître le teneur du contenu et de pouvoir en effectuer la classification. Cette classification se fait de manière arborescente et hiérarchique, suivant des catégories et sous-catégories successives qui sont ajoutées au fur et à mesure des besoins.

Un exemple de classification (catégories de l'annuaire www.yahoo.fr) :

Actualités & Médias
Arts & Culture
Commerce & Economie
Divertissement
Enseignement & Formation
Exploration géographique
Informatique & Internet
Institutions & Politique
Références & Annuaires
Santé
Science & Technologie
Science humaines
Société
Sports & Loisirs

avec sous la catégorie Santé :

48 sous-catégories dont

Diététiques & Nutrition
Droit & Législations
Economie de la Santé
Gériatrie & Vieillesse
Maladies & Pathologies
Médecine

et sous la mention Médecine :

65 sous-catégories dont
Anatomie
Anatomie Pathologique
Cancérologie & Oncologie
Cardiologie & Angiologie
etc...

Il est possible d'effectuer une recherche dans un annuaire de deux manières : la démarche traditionnelle qui était disponible depuis les débuts des annuaires, et la méthode par moteur de recherche qui a été implantée plus récemment, essentiellement pour contrer le succès croissant remporté par les nouveaux moteurs de recherche.

La démarche traditionnelle s'apparente à ce qu'on pourrait qualifier de «top-down » : l'utilisateur « descend » progressivement dans l'arborescence hiérarchique de la classification, en progressant de lien hypertexte en lien hypertexte, en essayant d'affiner le mieux possible le domaine sur lequel porte sa recherche, puis en fin d'arborescence dispose d'une liste d'URL qu'il peut explorer par navigation en cliquant sur les liens hypertextes présentés. Cette démarche possède l'avantage de son caractère global : on est en principe devant une liste représentative des pages web qui traitent du sujet indiqué par la classification. Il faut cependant se garder du mirage de l'exhaustivité : comme on l'a vu les délais d'indexation par les annuaires sont de plus en plus longs et la proportion indexée du Web par ces annuaires décroît sans cesse au fur et à mesure que la taille du Web augmente de façon exponentielle.

La méthode du moteur de recherche n'est à signaler qu'à titre anecdotique, car elle se fait de manière identique à celle de la recherche par interrogation des moteurs de recherche, mais uniquement sur la partie du Web qui a été indexée par l'annuaire.

Les annuaires dont la base est la plus étendue sont Yahoo (www.yahoo.com ou www.yahoo.fr), et Open Directory (www.dmoz.org). Il existe par ailleurs beaucoup d'autres annuaires qui apparaissent régulièrement, dont il est nécessaire d'évaluer le degré d'exhaustivité (ie proportion du Web indexé) mais aussi de pertinence (ie qualité de la grille de classification et de l'indexation par les opérateurs).

Rapidement, il est apparu qu'étant donné la nécessité d'ajouter de plus en plus de niveaux de sous-catégories àl'arborescence de classification hiérarchique, la navigation devenait progressivement plus longue et plus malaisée. Afin de pallier à cet inconvénient, des « annuaires » plus restreints ont été construits, avec des liens uniquement vers les pages traitant d'un domaine spécifique : ils reçoivent l'appellation de « portails ».

Le Web devenant de plus en plus investi par des acteurs économiques à la logique marchande, les portails se sont vus progressivement enrichir d'autres ressources qui ne sont plus strictement des liens vers d'autres pages web mais aussi des publicités, des offres de commerce en ligne, etc…

1.2) Les portails thématiques

Les portails thématiques sont des sites web où ont été regroupés pour des raisons de commodité tous les liens utiles pour un domaine spécifique. Ils sont le plus souvent créés et contrôlés par une ou plusieurs personnes experts dans ce domaine et offrent donc une grande pertinence pour ce qui est des liens présentés.

Cependant, il faut veiller à faire la distinction entre portails « universitaires » et portails « commerciaux », car pertinence n'équivaut pas toujours à fiabilité de l'information présentée.

Les portails les mieux pourvus dans le domaine médical sont le CISMeF de l'université de Rouen (www.chu-rouen.fr/cismef) et Medical Matrix (www.medmatrix.org). Une liste plus extensive des portails est donnée à titre indicatif en annexe.

2. La recherche par interrogation

2.1) Principe

Comme indiqué précédemment, l'inflation rapide de la taille du Web a fini par rendre impossible son indexation « manuelle » par des opérateurs humains avec une exhaustivité et une actualisation satisfaisantes. Par opposition à cette indexation sémantique, les moteurs de recherche sont des outils informatiques chargés d'effectuer une indexation lexicale des pages web : chaque page est considérée comme résultat plausible d'une recherche dès lors que les mots-clefs indiqués dans la requête de recherche sont présents dans le contenu de la page, et ce indépendamment du sens général de la page. Cette méthode a l'avantage de l'exhaustivité (relative néanmoins) puisque l'index (a) n'est plus constituée par des opérateurs humains mais des programmes informatiques qui ont une plus grande puissance de travail : elle a aussi l'inconvénient d'une plus faible pertinence puisque l'indexation ne se fait pas sur le sens du contenu.

Un moteur de recherche est composé d'un « spider » (b) (ou « crawler », les deux termes étant synonymes) qui est chargé de visiter toutes les pages web, d'en lire le contenu et d'en repérer les liens hypertextes ; d'un « index » (a) qui contient toutes les pages visitées par le spider ainsi que les pages issues du référencement volontaire par les auteurs de pages web ; et l'interface de recherche qui traite les requêtes des utilisateurs en lui renvoyant les pages de son index suivant un algorithme de repérage et de hiérarchisation propre. Il arrive que plusieurs moteurs utilisent un index commun mais diffèrent par leurs algorithmes de hiérarchisation des résultats.

En novembre 2000, le nombre approximatif des moteurs de recherche était estimé à 1500 toutes catégories confondues. Les statistiques montrent que 95% des internautes utilisent moins de 11 moteurs de recherche dans leur pratique courante.

2.2) Les moteurs de recherche généralistes

Les moteurs de recherche généralistes sont chargés d'indexer l'ensemble des pages web visibles par le spider. Ils ont en cela le souci de l'exhaustivité puisqu'ils

tentent d'indexer la plus grande partie du Web possible. En pratique, les index des moteurs les plus exhaustifs n'excède guère le quart du Web dit « visible ». Cependant, l'expérience concrète montre que cela suffit pour la plupart des requêtes usuelles.

(a) L'index d'un moteur de recherches désigne la base des documents web qu'il a indexés.
(b) Le spider ou crawler est le robot-programme chargé de parcourir les documents du web afin de les indexer.

Les moteurs actuels rivalisent entre eux pour ce qui est de la taille de leur index et signalent régulièrement l'extension de la taille de celui-ci. Il est donc difficile d'avoir une idée précise des exhaustivités comparées des moteurs de recherche entre eux. Cependant, un recul d'environ 15 mois (depuis le début de ce travail) permet de se rendre compte que la classification est désormais relativement stable et consiste en (en termes d'étendue de leur index, classement effectué sur www.searchengineshowdown.com/stats/size.shtml/ en août 2001):

1. Google www.google.com
2. All the Web www.alltheweb.com
3. Wisenut www.wisenut.com
4. Northern Light www.northernlight.com
5. HotBot www.hotbot.com
6. AltaVista www.altavista.com

entre autres.

Chaque moteur de recherche offre des caractéristiques différentes d'interrogation qui sont décrites, notamment, dans l'URL www.searchengineshowdown.com/features/ .

Internet étant encore un média en remaniement constant et rapide, il apparaît régulièrement de nouveaux moteurs de recherche et il en disparaît aussi certains au fil du temps. La taille de l'index des différents moteurs de recherche existants est elle aussi appelée à évoluer de façon rapide, condition nécessaire pour accompagner la croissance exponentielle du nombre de pages web existants. Il est ainsi nécessaire de pratiquer une veille régulière afin d'être certain d'utiliser toujours les instruments les mieux adaptés aux besoins de la recherche entreprise.

2.3) Les moteurs spécialisés

De même que précédemment, l'univers des moteurs de recherche spécialisés est en remaniement constant et rapide : une veille est ainsi nécessaire pour être certain d'utiliser l'instrument le plus adapté au moment où l'on entreprend la recherche.

2.4) Les méta-moteurs

Comme on a pu le voir précédemment, la taille du Web est telle que désormais aucun moteur de recherche ne peut se targuer d'en indexer plus d'une petite partie. Il arrive ainsi qu'une requête n'aboutisse à aucun résultat pertinent, voire aucun résultat du tout lorsqu'on interroge un seul moteur de recherche, parce

que son index ne contient pas la page web pertinente alors qu'elle existe bel et bien sur le Web.

Il est possible pour l'utilisateur d'interroger plusieurs moteurs de recherche l'un après l'autre, mais deux écueils se font rapidement jour : il faut se connecter à chaque moteur de recherche manuellement, et les résultats retournés sont souvent redondants d'un moteur de recherche à l'autre. Ce qui aboutit à une grande perte de temps pour l'utilisateur.

Afin de répondre à ce besoin d'interroger plusieurs moteurs de recherche, sont apparus le marché des méta-moteurs.

Les méta-moteurs sont des logiciels qui sont soit installés sur le disque de l'utilisateur soit installés au niveau du serveur et qui ont pour fonction d'interroger plusieurs moteurs de recherche, et de trier les résultats obtenus pour éliminer les redondances. Le chevauchement des index permet ainsi d'élargir la zone de recherche.

L'inconvénient reste le nombre de résultats obtenus qui sont souvent plus nombreux que celui obtenu avec un seul moteur de recherche, et parmi lesquels figurent beaucoup de résultats non pertinents.

De même que précédemment, l'univers des méta-moteurs est en remaniement constant et rapide : une veille est ainsi nécessaire pour être certain d'utiliser l'instrument le plus adapté au moment où l'on entreprend la recherche.

2.5) Les bases de données

Les bases de données bibliographiques regroupent des documents secondaires issus d'une indexation réalisée par des documentalistes professionnels sur des documents primaires. Elles font partie de cette portion du web qui n'est pas indexable par les moteurs de recherche : le « web invisible ». Les plus importantes en sciences biomédicales sont le Medline (accessible à http://www4.ncbi.nlm.nih.gov/PubMed , Pascal dont Pascal Biomed (accessible à www.inist.fr)

L'interrogation de Medline par l'intermédiaire de PubMed doit se faire sur le site de la National Library of Medicine pour bénéficier de toutes les fonctionnalités de l'outil de recherche.

Bien qu'ayant acquis une certaine stabilité en comparaison à l'univers des moteurs de recherche et méta-moteurs, celui des bases de données est lui aussi en perpétuel remaniement (hormis les bases de données bien installées comme Medline) et la veille reste utile afin de repérer l'apparition de nouvelles bases de données susceptible d'améliorer les recherches.

3. La synthèse : les « starting-points »

Chaque utilisateur a la possibilité par l'intermédiaire du langage HTML de se construire une page web qui contient toutes les ressources qu'il a l'habitude d'exploiter pour mener ses recherches sur Internet. L'affichage par la suite de cette

page par le navigateur constituera pour lui le point de départ pour la plupart de ses recherches et les liens hypertextes qui figurent déjà sur la page permettent d'un simple clic de souris de se rendre sur le site du moteur de recherche requis ou sur le site qui contient les ressources intéressantes : le gain de temps peut se révéler considérable et la mémoire complètement déchargée de la tâche ingrate de retenir les noms exacts des adresses URL. Une telle page peut être enregistrée sur le support de stockage de l'ordinateur de l'utilisateur, mais aussi être publiée sur Internet et servir de point de départ pour une recherche à l'utilisateur qui se trouve hors de l'endroit où se trouve son ordinateur, ou encore par tout autre utilisateur qui a accès à cette page web.

Une telle page peut être qualifiée de « starting-point » et son utilité est d'autant plus perceptible qu'elle peut aisément être mise à jour, réactualisée, modifiée au fur et à mesure que les outils informatiques progressent ou que la veille aura détecté des changements notables dans les instruments habituellement utilisés.

L'utilisateur qui a des bases en HTML peut facilement en créer une qui lui est personnelle, ou se connecter sur les pages de ceux qui l'ont publiée sur Internet comme :

- « Tous vos outils en une page » de Marie-Colette Fauré sur www.tours.inra.fr/tours/doc/ist1.htm
- « Vite… tous les outils en une page » de Jean-Pierre Lardy sur www.adbs.fr/adbs/sitespro/lardy/outils.htm
- « Atoute.org » de Dominique Dupagne sur www.atoute.org
- « Esculape » d'Hugues Raybaud sur www.esculape.com

II/ ETUDE COMPARATIVE ANNUAIRE VERSUS MOTEUR DE RECHERCHE

A. Préambule

L'analyse des principes de fonctionnement des deux approches d'extraction d'informations du web permet de bien saisir leur complémentarité: une approche de type arborescente (allant du plus général au plus spécifique) privilégiant la pertinence, caractéristique de la recherche par navigation; et une approche de type directe (examen du texte intégral des documents parmi lesquels on effectue sa recherche) privilégiant l'exhaustivité, caractéristique de la recherche par moteur (spécialisé ou non).

Outre cette dichotomie, il faut remarquer que pour un utilisateur sans culture ni formation particulères à l'Internet, le recours aux moteurs de recherche est plus naturel de par la simplicité de son utilisation, comparée à celle des annuaires (nécessité de maîtriser la séquence de recherche arborescente) et encore plus à celle des bases de données (nécessité de maîtriser la syntaxe d'interrogation de l'interface, voire du thésaurus associé).

Il ne faut donc guère s'étonner de fait que le recours aux moteurs de recherche soit prédominant par rapport à l'utilisation des annuaires thématiques et bases de données. Et ce d'autant plus que la plupart des médecins cliniciens interrogés estiment généralement satisfaisants les résultats obtenus en interrogeant leur moteur de recherche préféré (Google www.google.com) : cf archives du groupe de discussion Emilie http://groups.yahoo.com/group/emilie (10).

L'objectif de cette partie est de déterminer, avec le plus d'objectivité possible, les pertinences comparées des résultats obtenus par l'intermédiaire de l'une et l'autre des voies de recherche: annuaire ou moteur de recherche. La pertinence d'un document étant défini par : la capacité du document à répondre de façon satisfaisante à la question posée, sans préjuger de sa validité scientifique.

B. Méthodologie de recherche

Pour des raisons de maîtrise insuffisante des langues étrangères, les recherches qui seront conduites seront circonscrites aux documents rédigés en langue française.

L'évaluation des résultats des recherches menées portera sur la pertinence, c'est à dire sur la capacité des documents extraits à répondre de façon précise et satisfaisante à la question posée.

1. Matériels

1.1) Choix des outils de recherche

Dans un article de la Revue du Praticien Médecine Générale daté du 21/01/01 (11), Philippe Eveillard signale une étude publiée dans The Journal of Family Practice, numéro de novembre 2001, à propos du palmarès des banques de données médicales nord-américaines concernant leur capacité à renseigner les praticiens au cours de leur exercice quotidien (12). Cette étude porte sur l'interrogation des principales banques de données nord-américaines, triées dans un premier temps selon une procédure qui devait permettre de ne retenir que celles qui offrent une certaine qualité de contenu.

Parmi les sites recensés, n'étaient retenus que ceux qui se sont révélés capables de répondre correctement à au moins 4 des questions suivantes:

1. Quelles sont les causes de douleur thoracique ?
2. Quelles sont les causes d'anémie ?
3. Comment faire l'évaluation d'une lombalgie ?
4. Comment traiter une hypertension artérielle ?
5. L'amoxicilline est-elle efficace pour traiter une otite moyenne aiguë ?

Sur les 38 banques de données nord-américaines recensées, 14 ont pu remplir les conditions d'entrée dans l'étude.

Sur les 4 portails français dont l'accès est réservé aux professionnels de santé (www.33docpro.com , www.egora.fr , www.atmedica.com , www.medisite.fr), aucun ne passe le cap des 5 questions. Le CISMeF est le seul à faire un sans faute avec 5/5 et les polycopiés de santé lyonnais obtiennent un 4/5. Ainsi, le choix s'est naturellement porté sur le CISMeF qui dispose d'un moteur de recherche interne intégré pour l'évaluation envisagée dans ce travail.

Le choix du moteur de recherche à utiliser se révèle plus délicat dans la mesure où c'est un domaine très mouvant où les acteurs sont encore très nombreux. Le choix de Google a été finalement retenu dans la mesure où, outre son plébiscite par les médecins qui s'expriment sur les listes de discussions dédiées, il est donné comme disposant de l'index le plus exhaustif lors de la dernière évaluation par des instances indépendantes (dont www.abondance.com et www.searchengineshowdown.com).

1.1.1) Présentation du CISMeF

Le CISMeF, ou Catalogue Indexé des Sites Médicaux Francophones, est un projet de type annuaire : comme il est expliqué sur http://www.chu-rouen.fr/cismef.cismef.html , c'est un ensemble de documents online qui ont été indexés par une équipe de documentalistes, avec des critères très stricts de validité de l'information.

Le CISMeF recense exclusivement les ressources en Français de toutes provenances. L'étendue du CISMeF est d'environ 10 000 documents recensés en juillet 2001, avec un enrichissement de 40 nouvelles ressources en moyenne par semaine (en 2001).

Pour ce qui est de l'évaluation de la validité de l'information délivrée par les documents indexés, l'équipe du CISMeF utilise les critères du Netscoring qui a fait l'objet d'un paragraphe ci-dessus.

CISMeF est accessible à http://www.chu-rouen.fr/cismef et dispose d'une interface d'utilisation composée de 3 axes principaux : un moteur de recherches effectuant les recherches uniquement sur les documents indexés par l'équipe Cismef, un index alphabétique des mots-clefs qui ont servi à l'indexation des documents indexés par l'équipe Cismef, et un index thématique qui ventile par spécialités les documents indexés par l'équipe Cismef.

1.1.2) Présentation de Google

Google est un moteur de recherche dont l'interface d'interrogation est accessible sur http://www.google.com . Outre sa fonction de recherche de base, il possède de multiples fonctionnalités avancées : affinage des critères de recherche, recherches de documents de formats autres que HTML (Postscript, Word, PDF, etc…), recherches d'images, recherches dans les archives des forums de discussion, etc…

Ce qui distingue Google des autres moteurs de recherche tient en deux axes : celui de l'exhaustivité, avec un index qui est le plus large de tous les moteurs d'après l'évaluation d'un journaliste indépendant (http://www.searchengineshowdown.com/stats/size.shtml (13)); et celui de la pertinence des documents extraits, grâce à l'emploi de sa technologie brevetée PageRank.

PageRank étant une technologie brevetée, il est impossible d'obtenir des précisions détaillées sur son mode opératoire. Cependant, il est bien reconnu que son principe le plus important et le plus original est sa classification d'après le nombre de liens « entrants » (liens pointant vers le document considéré depuis un autre document) (14).

1.2) Présentation du pannel des questions-test

Afin de mieux permettre une mise en parallèle de cette étude et celle de l'équipe d'Alper BS, Stevermer JJ, White DS, Ewigman BG (12), la comparaison sera

opérée en utilisant le même pannel des questions que celui de leur étude, composé de 20 questions tirées au sort parmi plus de 1100 recensés par l'article de Ely J.W., Osheroff J.A., Ebell M.H., et al. (15).

La liste des questions se compose de:

1. Quelles sont les causes de dysurie chez une femme dont l'examen d'urines est normal ?
2. Comment déterminer la cause d'un prurit chronique ?
3. Quelles sont les causes des coccygodynies non traumatiques ?
4. Comment différencier un durillon d'une verrue plantaire ?
5. Quelles sont les causes de fièvres prolongées au-delà de 5 mois ?
6. Comment déterminer la cause d'un kyste situé en dessous du lobe de l'oreille chez une adulte ?
7. Qu'implique une élévation des antistreptolysines O chez un jeune homme de 18 ans ayant une hématurie ?
8. Qu'implique l'absence de flore intestinale normale dans une coproculture ?
9. Un sérodiagnostic de la maladie de Lyme est-il indiqué en présence d'un ou deux symptômes évocateurs ?
10. Quand faut-il envisager un test de grossesse sanguin à la place d'un test urinaire ?
11. Quelle est la dose de famciclovir nécessaire pour traiter une première poussée d'herpès génital ?
12. Quelle est la dose d'aspirine prescrite en prévention de la maladie coronaire et des accidents vasculaires cérébraux ?
13. A-t-on la preuve que Zoloft peut être la cause d'une douleur thoracique comme cela est indiqué dans le Physician's Desk reference (équivalent du Vidal).
14. Quels sont les médicaments pouvant être à l'origine de bouffées de chaleur chez une patiente ?
15. Quel est le traitement non chirurgical du canal lombaire étroit ?
16. Quand faut-il hospitaliser un patient souffrant de pneumonie ?
17. Quels sont les objectifs à atteindre (en termes de contrôle de la glycémie) chez une diabétique enceinte ou désireuse de l'être ?
18. Que faire pour un patient souffrant de cervicalgies à la suite d'un accident de la circulation, mais dont l'examen clinique est normal ?
19. Conduite à tenir devant une masse sensible d'un sein chez une femme allaitante ?
20. Conduite pratique chez une femme de 70 ans asymptomatique (sur le plan cardiovasculaire) chez laquelle un examen sanguin découvre un cholestérol à 2,34 g/L et un LDL-cholestérol à 1,75 g/L ?

2. Mise en oeuvre

Afin de permettre une éventuelle mise en parallèle de cette étude avec celle de l'équipe Alper BS, Stevermer JJ, White DS, Ewigman BG (12), la mise en œuvre a été conçue pour être la plus proche possible de celle qu'a adoptée l'équipe citée ci-dessus.

Pour chaque question, il est procédé à l'interrogation d'une part de CISMeF, et d'autre part de Google, en utilisant une stratégie simple et intuitive qui est supposée celle qu'utiliserait un internaute sans maîtrise pointue du sujet des recherches online. Il a été alloué pour chaque recherche un maximum de 10 minutes.

A chaque fois, dans CISMeF, il a été procédé directement à une recherche avancée par Doc'Cismef (donc sans utilisation de la recherche de base), puis en cas d'insuccès une recherche en parcourant l'index alphabétique, puis en cas de nouvel insuccès une recherche dans l'index thématique, en avançant dans l'arborescence indiquée aussi longtemps que le permet le temps alloué pour la recherche en cours.

C. Résultats

1. Les performances des deux outils en réponse aux 20 questions

1. Quelles sont les causes de dysurie chez une femme dont l'examen d'urines est normal ?

CISMeF : [((dysurie.mc ET etiologie.tc) ET femme.tc)] = aucun document extrait ; [dysurie.mc] = aucun document extrait; exploration de l'index thématique arborescence "gynécologie" = pas d'URL pertinent retrouvé; exploration de l'arborescence "urologie" = pas d'URL pertinent retrouvé par manque de temps.

Google : [cause dysurie femme] = 314 documents extraits, URL pertinent en 2è position http://cri-cirs-wnts.univ-lyon1.fr/Polycopies/Urologie/Urologie-11.html

2. Comment déterminer la cause d'un prurit chronique ?

CISMeF : [((prurit.mc ET chronique.tc) ET etiologie.tc)] = aucun document extrait; [(prurit.mc ET chronique.tc)] = aucun document extrait; [prurit.mc] = 2 documents dont le même URL que ci-dessous.

Google : [cause prurit chronique] = 1100 documents extraits; URL pertinent en 3è position (repérable dès lecture de l'aperçu) www.john-libbey-eurotext.fr/articles/dvs/0/145/7-10

3. Quelles sont les causes des coccygodynies non traumatiques ?

CISMeF : [coccygodynie.tc] = aucun document extrait ; mot-clef MeSH : pas de MC pour coccygodynie; exploration de l'index thématique arborescence « rhumato » = pas d'URL pertinent; arborescence « médecine générale » = pas d'URL pertinent.

Google : [cause douleur coccyx] = 199 documents extraits; URL pertinent en 3è position : www.vulgaris-medical.com/textc/coccygod.html

4. Comment différencier un durillon d'une verrue plantaire ?

CISMeF : [(durillon.tc ET verrue.tc)] = aucun document extrait; exploration de l'index thématique arborescence "dermatologie" = pas d'URL pertinent.

Google : [verrue durillon] = 47 documents extraits; URL pertinent en 1è et 2è position : www.doctissimo.fr/html/sante/mauxquot/sa_41_durillons.htm et www.doctissimo.fr/html/sante/encyclopedie/sa_1181_verrues.htm

5. Quelles sont les causes de fièvres prolongées au-delà de 5 mois ?

CISMeF: [(fièvre.mc ET prolongée.tc)] = aucun document extrait. [fièvre.mc] = 12 documents extraits; http://www.john-libbey-eurotext.fr/articles/met/7/1/20-2/index.htm ("Généralités sur les fièvres durables. Logistique des examens complémentaires." : pas tellement satisfaisant); http://www.med.univ-rennes1.fr/etud/pediatrie/fievre.htm#2 (Chez l'enfant: pas complètement satisfaisant non plus)

Google: [causes fievre prolongee] = 1440 documents extraits; 1er URL pertinent: http://www.medinfos.com/principales/fichiers/pm-inf-condfievreprol2.shtml

6. Comment déterminer la cause d'un kyste situé en dessous du lobe de l'oreille chez une adulte ?

CISMeF: [(kyste.tc ET visage.tc)] = aucun document extrait; [kyste] = aucun document extrait; exploration de l'index thématique arborescence "dematologie" = pas d'URL pertinent.

Google: [diagnostic kyste oreille] = 203 documents extraits; 4è URL pertinent: http://www-sante.ujf-grenoble.fr/sante/corpmed/Corpus/corpus/question/orl054.htm

7. Qu'implique une élévation des antistreptolysines O chez un jeune homme de 18 ans ayant une hématurie ?

CISMeF: [hématurie] = 4 documents extraits dont aucun pertinent; exploration de l'index alphabétique = http://www.nephrohus.org/3_cycle_folder/sm_hematurie.html ("Diagnostic des hématuries", bien rédigé mais ne mentionne pas les ASLO); exploration de l'index thématique arborescence "urologie" = pas d'URL pertinent.

Google: [hematurie ASLO] = 31 documents extraits; 7è URL pertinent http://www.doctissimo.fr/html/sante/encyclopedie/sa_517_glomerulopathies.htm

8. Qu'implique l'absence de flore intestinale normale dans une coproculture ?

CISMeF: [coproculture] = aucun document extrait; exploration de l'index thématique arborescence "infectiologie" = pas d'URL pertinent.

Google: [coproculture absence flore normale] = 21 documents extraits; aucun pertinent; [absence "flore intestinale normale" coproculture] = 3 documents extraits, aucun pertinent.

9. Un sérodiagnostic de la maladie de Lyme est-il indiqué en présence d'un ou deux symptômes évocateurs ?

CISMeF: [Lyme] = 10 documents extraits; aucun pertinent; exploration de l'index alphabétique = http://www.snof.org/maladies/lyme.html pertinent.

Google: [serologie Lyme] = 258 documents extraits, aucun pertinent parmi les 10 premiers; [serodiagnostic Lyme] = 65 documents extraits; aucun pertinent parmi les 10 premiers.

10. Quand faut-il envisager un test de grossesse sanguin à la place d'un test urinaire ?

CISMeF: [((diagnostic.tc ET grossesse.tc) ET sanguin.tc)] = 2 documents extraits, aucun pertinent ; [(diagnostic.tc ET grossesse.tc)] = 39 documents extraits, aucun pertinent; exploration de l'index thématique arborescence "obstétrique" = pas d'URL pertinent.

Google: [test grossesse sanguin urinaire] = 593 documents extraits, 9è URL pertinent http://www.gyneweb.fr/Sources/gdpublic/debgr/diag.htm

11. Quelle est la dose de famciclovir nécessaire pour traiter une première poussée d'herpès génital ?

CISMeF: [famciclovir] = aucun document extrait; [(herpès.tc ET traitement.tc)] = 2 documents extraits, aucun pertinent; exploration de l'index thématique arborescence "dermatologie" = pas d'URL pertinent.

Google: [herpes famciclovir dose] = 79 documents extraits, 3èURL pertinent http://www.anaes.fr/ANAES/Publications.nsf/nID/LILF-553LMR?OpenDocument&Back=LILF-553KW8

12. Quelle est la dose d'aspirine prescrite en prévention de la maladie coronaire et des accidents vasculaires cérébraux ?

CISMeF: [(aspirine.tc ET antiagrégant.tc)] = aucun document extrait; [aspirine] = 6 documents extraits, aucun pertinent; exploration de l'index thématique arborescence "cardiologie" = http://cri-cirs-wnts.univ-lyon1.fr/Polycopies/Cardiologie/index.html pertinent.

Google: [aspirine antiagrégant dose] = 302 documents extraits, 1er URL pertinent http://www-sante.ujf-grenoble.fr/sante/corpmed/Corpus/corpus/question/hema172.htm

13. A-t-on la preuve que Zoloft peut être la cause d'une douleur thoracique comme cela est indiqué dans le Physician's Desk reference (équivalent du Vidal).

CISMeF: [((sertraline.tc ET douleur.tc) ET thoracique.tc)] = aucun document extrait; exploration de l'index alphabétique arborescence "effets indésirables" = pas d'URL pertinent.

Google: [zoloft douleur thoracique] = 7 documents extraits, 1er URL pertinent http://www.biam2.org/www/Sub4976.html

14. Quels sont les médicaments pouvant être à l'origine de bouffées de chaleur chez une patiente ?

CISMeF: [bouffée de chaleur.tc] = aucun document extrait; exploration de l'index alphabétique rubrique "bouffées de chaleur"= rien; exploration de l'index alphabétique rubrique "effets indésirables" = aucun document pertinent extrait.

Google: [bouffe chaleur origine medicamenteuse] = 119 documents extraits, 1er URL pertinent http://www.biam2.org/www/Spe3523.html car pointe vers http://www.biam2.org/www/SpeEIIMCBOUFFEE_DE_CHALEUR.html

15. Quel est le traitement non chirurgical du canal lombaire étroit ?

CISMeF: [((canal.tc ET lombaire.tc) ET etroit.tc) = aucun document extrait; exploration de l'index alphabétique = pas d'URL pertinent; exploration de l'index thématique, arborescence "rhumatologie": pas d'URL pertinent.

Google: [canal lombaire etroit traitement] = 249 documents extraits, 5è URL pertinent http://www.esculape.com/fmc/canallombaire.html

16. Quand faut-il hospitaliser un patient souffrant de pneumonie ?

CISMeF: [(pneumopathie.tc ET hospitalisation.tc)] = 1 document extrait, non-pertinent. Exploration de l'index thématique, arborescence "infectiologie": pas d'URL pertinent; arborescence "pneumologie" = pas d'URL pertinent.

Google: [pneumonie critere hospitalisation] = 138 documents extraits, 5è URL pertinent http://www.infectio-lille.com/diaporamas/HG/pn-com.PDF

17. Quels sont les objectifs à atteindre (en termes de contrôle de la glycémie) chez une diabétique enceinte ou désireuse de l'être ?

CISMeF: [(diabetique.tc ET enceinte.tc)] = aucun document extrait; exploration de l'index thématique arborescence "obstétrique" = pas d'URL pertinent, arborescence "endocrinologie": URL pertinent http://www.chups.jussieu.fr/polys/diabeto/POLY.Chp.4.html .

Google: [diabetique enceinte glycemie] = 350 documents extraits, 3è URL pertinent http://pro.gyneweb.fr/Sources/congres/jta/01/obs/CHRISTIN-MAITRE.HTM

18. Que faire pour un patient souffrant de cervicalgies à la suite d'un accident de la circulation, mais dont l'examen clinique est normal ?

CISMeF: [(douleur.tc ET cervicale.tc)] = 1 document extrait, non-pertinent; exploration de l'index thématique arborescence "rhumatologie" = pas d'URL pertinent retrouvé.

Google: [cervicalgie post traumatique] = 56 documents extraits, aucun pertinent parmi les 10 premiers; [cervicalgie accident circulation] = 24 résultats, 3è URL pertinent http://cri-cirs-wnts.univ-lyon1.fr/Polycopies/Rhumatologie/Rhumatologie-2.html

19. Conduite àtenir devant une masse sensible d'un sein chez une femme allaitante ?

CISMeF: [((masse.tc ET sein.tc) ET allaitante.tc)] = aucun document extrait; [(masse.tc ET sein.tc)] = 1 document extrait, non-pertinent; exploration de l'index thématique arborescence "gynécologie" = pas d'URL pertinent retrouvé; arborescence "obstétrique" = pas d'URL pertinent retrouvé.

Google: [masse sensible sein femme allaitante] = 18 documents extraits, aucun pertinent; [masse sein femme allaitante] = 72 documents extraits, aucun pertinent parmi les 10 premiers; [masse sein femme] = 19000 documents extraits, aucun pertinent parmi les 10 premiers.

20. Conduite pratique chez une femme de 70 ans asymptomatique (sur le plan cardiovasculaire) chez laquelle un examen sanguin découvre un cholestérol à 2,34 g/L et un LDL-cholestérol à 1,75 g/L ?

CISMeF: [(cholesterol.tc ET LDL.tc)] = aucun document extrait; exploration de l'index alphabétique arborescence "cholestérol" = http://www.anaes.fr/ANAES/Publications.nsf/nID/LILF-48MET2?OpenDocument&Back=APEH-3YJB66 pertinent.

Google: [cholesterol LDL] = 3490 documents extraits, 10è URL pertinent http://www.esculape.com/fmc/cholesterol.html

2. Comparaison CISMeF vs Google

Dans 17 cas sur 20, Google a permis de trouver un document pertinent.
Dans 3 cas sur 20, CISMeF a permis de trouver un document pertinent.
La question n° 9 est la seule pour laquelle CISMeF a permis de trouver un document pertinent alors que Google n'en a pas retrouvé.

Au cours des recherches menées, certains éléments sont ressortis:

1. L'interrogation de Google a donné des résultats satisfaisants avec des mots-clefs simples et intuitivement évidents: ceux dont on s'attend à ce qu'ils se trouvent dans les documents qui parlent du sujet étudié. Google est donc devenu accessible à un utilisateur sans grande expérience de la recherche d'informations pourvu qu'il ait à l'esprit, en menant sa recherche, de donner comme mot-clef ceux qu'il s'attend à trouver dans le document répondant à la question qu'il se pose.

2. A chaque fois que Google a pu extraire un document pertinent, celui-ci a été identifié en moins de 5 minutes.

3. Pour les questions qui ont trouvé une réponse à la fois sur CISMeF et sur Google, le délai nécessaire a été plus court avec Google.

4. La recherche sur CISMeF a parfois été bloquée par un incident technique (erreur interne sur le serveur, serveur saturé). A l'opposé Google est resté disponible et réactif pour toutes les recherches menées.

La grande surprise révélée à l'issue de ce travail fut non pas la prééminence de Google pour répondre aux questions ponctuelles (puisqu'il est admis qu'un annuaire est surtout l'outil pour faire le point des références sur une question plus générale), mais le fait que Google a su classer un URL pertinent parmi les 10 premiers documents extraits 17 fois sur 20, montrant ainsi le progrès réalisé depuis le test opéré sur AltaVista et Yahoo en avril 1999 (16).

Google a été capable de donner un document pertinent pour la question posée dans 17 cas sur 20, ce qui représente un taux de succès de 85%, à rapprocher de celui de 85% obtenu en croisant les 2 meilleures bases de données médicales nord-américaines (12). Sauf à supposer que le Web francophone contient significativement plus de documents médicaux pertinents que le Web anglophone (ce qui semble peu probable), ce résultat tend à signifier que Google fait aussi bien –en termes de pertinence- que la réunion des 2 meilleures bases de données médicales nord-américaines.

3. Etude complémentaire de Google concernant les 3 échecs de recherche

Pour ce qui concerne les questions 8, 9 et 19 une recherche complémentaire a été menée avec Google, puis Fast et Wisenut. Les résultats ont confirmé la meilleure exhaustivité de l'index de Google sur ceux des 2 autres moteurs de recherche: les dix premiers résultats présentés par Fast et Wisenut étaient tous présents dans ceux présentés par Google.

Par ailleurs, une recherche approfondie par Google aboutit à:

- Pour la question 19:

Exploration des 73 documents extraits: aucun pertinent.

[tumeur sein femme allaitante] = 16 documents extraits, aucun pertinent.

[galactocele] = 3 documents extraits, aucun pertinent.

Preuve, s'il était nécessaire, que le Web n'a pas encore réponse à toutes les questions.

- Pour la question 8:

[flore intestinale] = 3630 documents extraits…beaucoup viennent du site de l'INRA dans les 30 premiers

[flore intestinale –inra] = 3380 documents extraits…

["absence de flore intestinale"] = 3 documents extraits, aucun pertinent.

- Pour la question 9:

[serologie Lyme] = 260 documents extraits, 10è URL presque pertinent http://www.unaformec.org/publications/bibliomed/177_Maladie_de_Lyme.pdf , 11è URL pertinent http://labo93.free.fr/lyme.html ; 21è URL est le plus clair http://perso.infonie.fr/tiquatac/troisstades.htm

En l'espace de 2 jours, 2 documents supplémentaires ont pu être extraits, dont l'un figurant dans les 10 premiers pour cette dernière recherche. Cela atteste de la progression continue de la taille de l'index de Google.

Il faut cependant remarquer que ces URL n'appartiennent pas à des sites qui peuvent être considérés comme délivrant une information fiable a priori.

Devant l'absence de document pertinent retrouvé pour les questions 8 et 19 lors de l'exploration de la totalité des documents extraits par Google, on peut remarquer que pour les 20 questions-test, à chaque fois qu'un document pertinent a pu être extrait par l'équation de recherche utilisée, il a été classé par Google dans les 10 premiers.

4. Performances de Google pour une question d'ordre plus général

Au cours des discussions qui ont été échangés sur la liste de discussion Emilie (10), une remarque a été soulevée concernant la nature très ponctuelle des questions du pannel qui a servi au test comparatif : l'hypothèse a ainsi été émise que Google aurait des performances notablement dégradées si l'on devait l'utiliser pour effectuer des recherches sur des sujets d'ordre plus général, comme des recommandations de bonne pratique ou une synthèse sur une pethologie donnée par exemple.

A titre d'essai, il a ainsi été procédé à quelques recherches afin de tester cette hypothèse. Voici ce qu'elles ont donné :

Recherche des recommandations sur le traitement de l'asthme chez l'enfant :

Google : [asthme enfant] : 4è URL pertinent http://www.esculape.com/fmc/pediasthme.html ; 5è lien pointant vers le CISMeF http://www.chu-rouen.fr/ssf/pathol/asthme.html

CISMeF : section recommandations et consensus -> lien vers le site de l'ANAES, mais pas d'autre recommandation que celle pour l'éducation du patient mise en ligne sur ce site ; section index alphabétique -> URL pertinent « asthme infantile » http://www.med.univ-rennes1.fr/etud/pediatrie/asthme.htm

Recherche des recomm sur la traitement du diabète type 2 :

Google : [traitement diabete « type 2 »] : 1er URL pertinent
http://www.chups.jussieu.fr/polys/diabeto/POLY.Chp.14.html

CISMeF : section recommandations et consensus -> lien vers le site de l'ANAES avec recommandations en fichier PDF URL pertinent
http://www.anaes.fr/anaes/Publications.nsf/wEdition/RE_LILF-4K9DKJ?OpenDocument&Retour=wSpecialites?OpenView

On peut constater que là encore, sur le plan de la pertinence, Google fait aussi bien que CISMeF. En revanche, des progrès restent encore à faire sur le plan de la validité de l'information délivrée. Il faut cependant se garder d'extrapolations hasardeuses sur la foi de 2 essais couronnés de succès.

D. Discussion

1. La distinction entre pertinence et validité

Au cours de toute l'étude, l'évaluation des documents extraits par les différentes recherches était portée uniquement sur le critère de pertinence, au sens où il a été défini dans le préambule de la deuxième partie : la capacité du contenu du document à répondre de façon satisfaisante à la question posée, sans préjuger de la validité scientifique de l'information ainsi délivrée.

Si l'on regarde de plus près les documents pertinents extraits par Google, on peut les classer en deux catégories : ceux qui appartiennent à des sites institutionnels à contenu a priori digne de confiance, et ceux qui appartiennent à des sites dont la fiabilité est plus sujette à caution.

Google a extrait des documents de la première catégorie dans 9 cas sur 17, soit une proportion de 53%. Dans les 8 cas restant, les sites auxquels appartiennent les documents extraits sont moins identifiables, donc à fiabilité plus incertaine.

L'information brute ainsi extraite d'Internet par un moteur de recherche n'est pas utilisable par n'importe qui : il est nécessaire de la soumettre à un filtre critique que suppose de solide connaissances médicales préalables.

Par ailleurs, même pour un médecin, à l'exception des sites institutionnels bien identifiés, la justesse de l'information délivrée reste parfois délicate à évaluer eu égard à la minceur de la frontière séparant une information plausible d'une information scientifiquement validée.

Ainsi resurgit la préoccupation de la validité de l'information extraite, une fois satisfaite l'exigence de pertinence.

A cette aune, CISMeF conserve naturellement sa prééminence, grâce au travail d'indexation rigoureux de son équipe de chercheurs rompus à cette tâche requérant une formation médicale significative.

Il est enfin envisageable d'ajouter un filtre supplémentaire à Google, permettant d'opérer un deuxième tri parmi les résultats qu'il fournit: cela peut se faire soit en passant par les fonctions avancées de Google (http://www.google.fr/advanced_search?hl=fr), soit de façon encore plus fine en adjoignant à Google un logiciel, qui ne retiendrait que les documents issus d'une liste

de sites considérés comme délivrant de l'information de qualité, liste qui peut faire l'objet d'une mise à jour régulière soit par un contrat de maintenance logicielle soit par enrichissement "manuel" de l'utilisateur. Un tel logiciel reste à écrire car il n'en existe pas pour l'instant dans le commerce.

2. Subjectivité de l'évaluation de la pertinence des documents extraits

A la différence de ce qui a été réalisé dans l'article comparant les bases de données médicales nord-américaines (12) où la pertinence des réponses étaient évaluées par plusieurs médecins dont les avis étaient ensuite confrontés, cette étude a été réalisée par un seul opérateur (moi-même), dont l'avis sur la pertinence du contenu des documents extraits est forcément subjectif.

Bien que, dans le cas des 20 questions du pannel ayant servi à la comparaison, chacune appelant une réponse concrète et précise, peu de place était laissé à une appréciation teintée de subjectivité, il paraissait judicieux de mentionner cette légère différence de mise en oeuvre dans la conduite des recherches afin de mieux mettre en perspective le parallèle qui peut être fait entre ces étude et celle de l'équipe d'Alper BS, Stevermer JJ, White DS, Ewigman BG (12).

3. Choix du pannel de questions ayant servi au test

En raison des contraintes de temps imparti, mais aussi afin de mieux pouvoir mettre en parallèle cette étude et celle de l'équipe d'Alper BS, Stevermer JJ, White DS, Ewigman BG (12), le pannel de questions est celui des 20 citées précédemment. Celles-ci ont été tirées au sort parmi les plus de 1100 citées dans l'article de Ely J.W., Osheroff J.A., Ebell M.H., et al. (15).

La taille modeste de l'échantillon des questions ayant servi à l'évaluation des sources d'informations médicales incite à la prudence quant à une extrapolation quant au domaine plus vaste de toutes les questions pouvant être soulevées en pratique quotidienne de la médecine générale.

Par ailleurs, un doute pouvait être soulevé quant à la nature des questions du pannel, celles-ci étant puisées dans une étude nord-américaine dont le contexte et la culture médicale ne sont sans doute pas identiques à ceux de la France: la revue générale des 20 questions permet toutefois d'estimer (avec toutes les réserves qu'on peut porter sur la subjectivité de l'auteur de ce travail) que les questions énumérées correspondent assez bien aux préoccupations qui sont celles d'un médecin généraliste français en ville.

4. Google est-il « utilisateur-dépendant » ?

Au vu des résultats assez surprenants de Google en termes de pertinence, on peut être amené à se demander si la pertinence obtenue n'est pas liée à la façon de l'interroger, c'est à dire à la formulation de l'équation de recherche.

Comme on peut le constater à la lecture des résultats, qui exposent pour chaque question les équations de recherche qui ont été utilisées, la formulation de ces équations de recherche reste intuitive et assez simple : des mots-clefs reprenant les notions impliquées dans la question que se pose le praticien.

Cependant, il reste légitime de se demander ce qu'auraient été les résultats de recherche si les mêmes outils étaient utilisés par des médecins novices en termes de recherche documentaire (sur Internet ou non).

5. Synthèse

Malgré les réserves exprimés ci-dessus, les résultats de cette étude semblent montrer que Google, doté de sa technologie de classement des documents extraits et grâce à son vaste index recensant une grande partie du Web visible, a fait la preuve de son efficacité pour une utilisation quotidienne par un médecin généraliste dans le cadre de sa pratique courante en ville. Il reste à résoudre l'épineux problème de la validité scientifique des informations extraites, pour lequel CISMeF représente pour l'instant une solution incontournable de par la qualité d'indexation de son équipe de documentalistes.

Ainsi se trouve confirmée l'idée bien répandue parmi la communauté des médecins utilisateurs d'Internet et nouvelles technologies de l'information en général (liste de discussion Emilie (10)) que l'interrogation du Web par l'intermédiaire de Google permet dans la quasi-totalité des cas de trouver réponse satisfaisante à la question posée en pratique courante. Malgré ses limites qui ont été discutées ci-dessus, cette étude a permis de caractériser de façon objective cette impression, et ouvre désormais le champ à d'autres axes d'investigations sur le même thème.

6. Perspectives

Ainsi que la remarque a été mentionné ci-dessus, une comparaison entre moteur de recherche et annuaire est forcément entaché, d'un point de vue strictement scientifique, d'un biais important puisque les deux outils ont des caractéristiques très différentes, opposées et complémentaires comme il a été signalé dans l'introduction de ce travail. Le but poursuivi était cependant d'évaluer la performance de la technologie de classement des résultats extraits par Google en termes de pertinence.

Il serait intéressant de pouvoir comparer, selon un protocole similaire, l'interrogation de plusieurs moteurs de recherche, afin de savoir si d'autres technologies que celle de Google (par exemple Exalead, WiseNut ou encore Fast) possèdent les mêmes performances.

Un autre consensus dégagé dans les disucssions entre médecins utilisateurs de l'Internet et des nouvelles technologies de l'information (liste de discussion Emilie (10)) voudrait que les moteurs de recherche (tels Google) aient de piètres performances au regard de ceux des annuaires thématiques (tels CISMeF) pour ce qui est des questions d'ordre plus général comme la synthèse sur une pathologie donnée par exemple : on a pu constater au travers des deux essais réalisés avec Google que ce n'était pas forcément le cas.

Il serait intéressant de pouvoir comparer, selon un protocole similaire, l'interrogation d'un moteur de recherche et d'un annuaire thématique en utilisant un pannel de questions d'ordre plus général.

Ainsi qu'on a pu le constater dans cette étude, il y a des cas où Google n'a pas pu extraire de documents pertinent. Bien qu'un rapide tour d'horizon des autres moteurs de recherche majeurs n'ait pas permis de déceler de nouveaux documents indexés par ceux-ci, il est vraisemblable que les index des différents moteurs de recherche n'étant pas strictement inclus les uns dans les autres, leur réunion

permettrait de lancer des recherches dans un corpus de connaissances plus vaste. De telles recherches pourraient être conduite au travers de l'utilisation de méta-moteurs de recherche, avec pour inconvénient prévisible toutefois une quantité de documents extraits plus abondante et parmi lesquels figurent beaucoup de documents non-pertinents.

Il serait intéressant de pouvoir comparer un méta-moteur avec Google, tant sur le plan de la capacité à extraire un document pertinent que sur celui de la capacité à le classer parmi les dix premiers.

Enfin, on peut envisager l'écriture de programmes destinés à pallier à certaines des limites mentionnés ci-dessus : un filtrage des résultats de recherche du moteur de recherche basé sur le critère de validité scientifique, ou encore un module d'aide à la rédaction des équations de recherche à partir d'une question en langage naturel.

Il serait intéressant de connaître les performances de tels outils adjoints à l'utilisation d'un moteur de recherche ou encore d'un méta-moteur de recherche.

CONCLUSION

Constatant qu'Internet, ayant dépassé le phénomène de mode pour devenir une source majeure d'informations scientifiques et médicales, et prenant conscience à la fois de la complémentarité des deux types d'outils qui en permettent l'utilisation à des fins d'extraction d'informations que sont les moteurs de recherche d'une part et les annuaires thématiques d'autre part, ainsi que de la plus grande facilité d'utilisation des outils de la première catégorie (les moteurs de recherche), l'objet de ce travail était d'entreprendre une étude comparative, en se plaçant dans un cadre purement pratique, de l'utilisation d'un moteur de recherche par rapport à celle d'un annuaire thématique, en examinant les résultats d'un strict point de vue de la pertinence, cette notion étant définie comme la capacité d'un contenu à répondre de façon satisfaisante à la question posée-sans préjuger de sa validité scientifique.

Cette étude a été menée de façon à se placer dans des conditions aussi proches que possible d'un travail réalisé par une équipe américaine en 2001 (12), qui comparait les résultats de recherches effectuées sur les bases de données médicales nord-américaines en utilisant un pannel de 20 questions tirées au sort parmi plus de 1100 recensés par une étude entreprise auprès des médecins de famille américains en 1999 (15). Le même pannel de 20 questions a donc été utilisé pour évaluer les capacités d'un moteur de recherche d'une part, et d'un annuaire thématique d'autre part à extraire des documents pertinents au cours d'une interrogation de leurs interfaces d'utilisation, avec limitation du temps imparti à dix minutes maximum.

Le choix de l'annuaire thématique à tester s'est naturellement porté sur CISMeF, qui fait autorité en matière d'indexation des ressources médicales sur Internet en France, et qui s'est révélé le seul annuaire thématique à satisfaire aux critères qui ont conduit à la sélection des bases de données entrant dans le protocole d'étude de l'article cité ci-dessus (12).

Le choix du moteur de recherche a finalement été arrêté sur Google, en raison de la plus grande étendue de son index par rapport à ceux des autres moteurs de recherche, et aussi en raison de sa technologie de classement des résultats bien reconnue comme étant la plus performante selon l'avis des médecins utilisateurs d'Internet et des nouvelles technologies de l'information, consensus dégagé en consultant les archives de la liste de discussion Emilie (10).

Les résultats obtenus ont montré sans surprise la plus grande performance de Google quant à la capacité à extraire un document pertinent. Cependant, il faut souligner la très bonne performance de Google puisqu'il fait aussi bien que la réunion des deux meilleures bases de données médicales nord-américaines (réponse pertinente fournie dans 85% des cas), et qu'il réussit à classer le document pertinent dans les dix premiers à chaque fois qu'un document pertinent a pu être extrait.

Accessoirement, les résultats ont également montré la plus grande facilité d'utilisation de Google par rapport à CISMeF, tant sur le plan de la rapidité et de la

disponibilité du serveur interrogé que sur le plan du caractère intuitif de son utilisation.

La confirmation est ainsi démontrée que l'utilisation d'un moteur de recherche doté d'un index large et d'un algorithme de classement puissant tel que Google est devenu tout à fait envisageable dans le cadre de la pratique quotidienne d'un médecin généraliste au cabinet, et ce sans nécessiter de formation particulière.

Il reste qu'une bonne culture médicale est indispensable afin d'identifier les documents dont le contenu est digne de confiance. Afin de faire le point sur un sujet général avec des documents de référence, CISMeF demeure toujours la ressource prééminente grâce à la fiabilité des documents indexés par son équipe de chercheurs. L'idéal, non réalisé à ce jour, serait de pouvoir appliquer à un corpus exhaustif de connaissances validées et régulièrement mises à jour une technologie d'indexation et de classification des documents aussi performante que PageRank (propriété de Google).

Ainsi, une fois résolu l'impératif de pertinence des moteurs de recherche, on peut voir émerger à nouveau d'autres axes de recherche portant sur les notions d'exhaustivité du corpus de connaissances mis en ligne (et aussi son accessibilité au travers de la nuance entre Web visible et Web invisible) ou encore de validité scientifique de l'information mise en ligne.

Le champ des recherches sur les apports de l'Internet en matière d'informations médicales reste ainsi encore très ouvert et bien des études intéressantes restent à mener afin de mieux cerner les procédés qui permettent son utilisation optimale.