explorées consiste à attacher des métadonnées aux fichiers, de sorte que les systèmes d'indexation obtiennent directement ces informations. Les travaux les plus avancés sont ceux du programme *Dublin Core Metadata* et du projet *Warwick Framework* – le premier ainsi nommé d'après un atelier de travail qui eut lieu à Dublin, dans l'Ohio, le second d'après un colloque qui s'est tenu à Warwick, en Angleterre. Ces groupes de travail ont défini un ensemble de métadonnées plus simples que celles qui sont classiquement utilisées pour constituer les catalogues de bibliothèques et ont mis au point des méthodes pour intégrer les métadonnées dans les pages du réseau.

Ces métadonnées contiendraient, par exemple, le titre d'un document, son auteur ou son genre (texte, vidéo, etc.). Elles pourraient être produites manuellement ou par des logiciels d'indexation automatique. Par ses annotations précises et détaillées, la description donnée par l'être humain reste bien supérieure à celle qui est obtenue par un programme d'indexation automatique.

Lorsque les coûts sont justifiés, des indexeurs humains compilent laborieusement les bibliographies de certains sites. La base de données *Yahoo*, par exemple, répertorie les sites en fonction de rubriques très générales. Dans un projet de recherche de l'Université du Michigan, on cherche à mettre au point une description de sites contenant des informations spécifiquement destinées aux chercheurs.

Plus qu'une simple bibliothèque

Les stratégies adoptées pour la recherche et l'indexation (humaine ou automatisée) des documents dépendront des utilisateurs de l'Internet et des perspectives commerciales offertes aux éditeurs. Nombre de chercheurs souhaitent conserver une bibliothèque numériquement structurée, mais, pour d'autres, un média démocratique et libre serait le support idéal pour la diffusion de l'information. Certains utilisateurs, tels les analystes financiers et les espions, souhaitent un accès illimité aux bases de données brutes, libre de tout contrôle. Pour eux, les moteurs de recherche standards offrent de réels avantages, car ils ne filtrent pas les données.

La recherche d'images sur le Web

Depuis quelques années, le public découvre que le réseau Internet lui offre des photographies, des animations, des graphiques, des sons et des films consacrés à des sujets variés, du grand art à la franche obscénité. Malgré cette surabondance d'informations, la recherche d'un document, parmi des centaines de milliers de sites, repose essentiellement, aujourd'hui encore, sur des index de mots et de nombres.

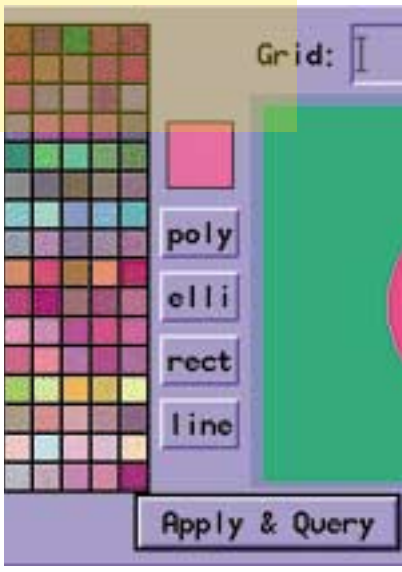
Si l'on recherche «drapeau français» sur le moteur de recherche *AltaVista*, on obtient l'image souhaitée, à condition qu'elle ait été légendée avec ces deux mots. Mais comment quelqu'un qui connaît seulement l'aspect du drapeau peut-il retrouver son pays d'origine?

Le réseau serait utile si les moteurs de recherche permettaient de dessiner ou de numériser un rectangle composé de trois bandes verticales colorées en bleu, en blanc et en rouge, et s'ils trouvaient toutes les images correspondantes stockées sur la myriade de sites. Ces dernières années, des techniques combinant indexation par mots clés et analyse d'images ont ouvert la voie aux premiers moteurs de recherche d'images.

Bien que ces prototypes soient des pionniers de l'indexation graphique, ils montrent à quel point les outils actuels sont rudimentaires et fondés sur l'usage du texte. *WebSEEk*, un logiciel expérimental conçu à l'Université Columbia, illustre le fonctionnement d'un moteur de recherche d'images. Il commence par charger les fichiers trouvés en visitant les sites, puis tente de localiser les noms de fichiers contenant des acronymes, tels que GIF ou MPEG, désignant un contenu graphique ou vidéo. Il cherche également, dans ces noms de fichiers, des mots qui pourraient identifier le sujet des fichiers. Lorsque *WebSEEk* trouve une image, il détermine ses couleurs prépondérantes et leurs positions. À l'aide de ces informations, le programme distingue les photographies, les graphiques, les images en noir et blanc ou en gris. Le logiciel comprime également chaque image et la transforme en une icône miniature qu'il pourra afficher. Pour une vidéo, le programme *WebSEEk* extrait des images de plusieurs scènes.

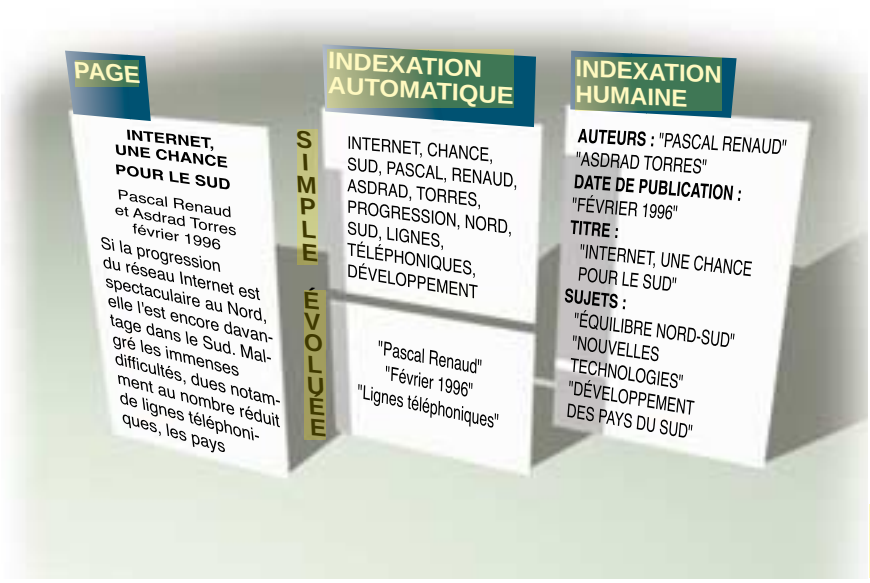
Un utilisateur commence sa recherche en choisissant une catégorie dans un menu – par exemple, «chats». *WebSEEk* affiche alors un échantillon d'icônes associées à la catégorie «chats». Pour limiter sa recherche, l'utilisateur peut cliquer sur toute icône montrant des chats noirs. Grâce à l'analyse de couleurs précédemment effectuée, le moteur de recherche cherche des correspondances avec des images ayant un profil de couleur similaire. Il présente alors un nouveau lot d'images qui représentent des chats noirs – mais aussi, par exemple, quelques chats jaune et orange assis sur des coussins noirs. L'utilisateur peut alors affiner sa recherche en ajoutant ou en excluant certaines couleurs d'une image avant de relancer la recherche. En excluant le jaune et l'orange, il fait disparaître les chats colorés. Plus simplement, l'utilisateur peut aussi désigner les images qui ne contiennent pas de chats noirs, indiquant par là au programme de ne pas rechercher ce type d'image. À ce jour, *WebSEEk* a transféré et indexé plus de 650 000 images provenant de dizaines de milliers de sites.

Plusieurs centres informatiques et quelques sociétés, telles qu'*IBM* et *Virage*, mettent au point des programmes de recherche d'images sur des réseaux privés ou dans des bases de données. Deux autres sociétés, *Excalibur Technologies* et *Interpix Software*, se sont associées pour fournir des logiciels aux moteurs de recherche *Yahoo* et *Infoseek*.



C'est pourtant l'un des plus anciens programmes de recherche d'images, QBIC d'*IBM*, qui fournit les meilleures correspondances entre les motifs des images. Le logiciel QBIC repère les couleurs dans une image, et il mesure plusieurs paramètres pour en évaluer la texture : le contraste (le blanc et le noir des rayures de zèbre), le grain (la différence entre les galets et les rochers) et l'orientation (la différence entre les poteaux parallèles d'une clôture et l'isotropie des pétales de fleur). Ce programme

L'information circulant sur l'Internet est bien plus variée que celle des bibliothèques classiques. Une bibliothèque ne classe pas, par ordre de qualité, les ouvrages qu'elle contient. Toutefois, comme l'Internet comprend un volume d'information supérieur, ses utilisateurs veulent utiliser au mieux le temps limité dont ils disposent pour effectuer une recherche. Ils veulent connaître, par exemple, les trois «meilleurs» documents traitant d'un



Bryan Christie

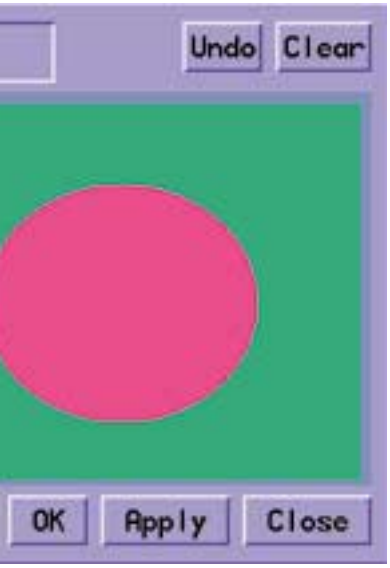
3. L'INDEXATION AUTOMATIQUE effectuée par un moteur de recherche *(panneau de gauche)* conduit, à partir d'une page d'un article du *Monde diplomatique*, à une liste de mots *(panneau central, en haut)* ou de groupes de mots simples *(panneau central, en bas)*. L'indexation humaine *(panneau de droite)* donne des indications plus structurées.

peut également rechercher des formes simples dans les images. Si on lui demande de rechercher des images similaires à un point rose sur un fond vert, il renvoie des fleurs et d'autres photographies contenant des formes et des couleurs analogues *(voir ci-dessous)*. Ce logiciel possède plusieurs applications, allant du choix des motifs de papiers peints à l'identification policière de criminels par le type de vêtements.

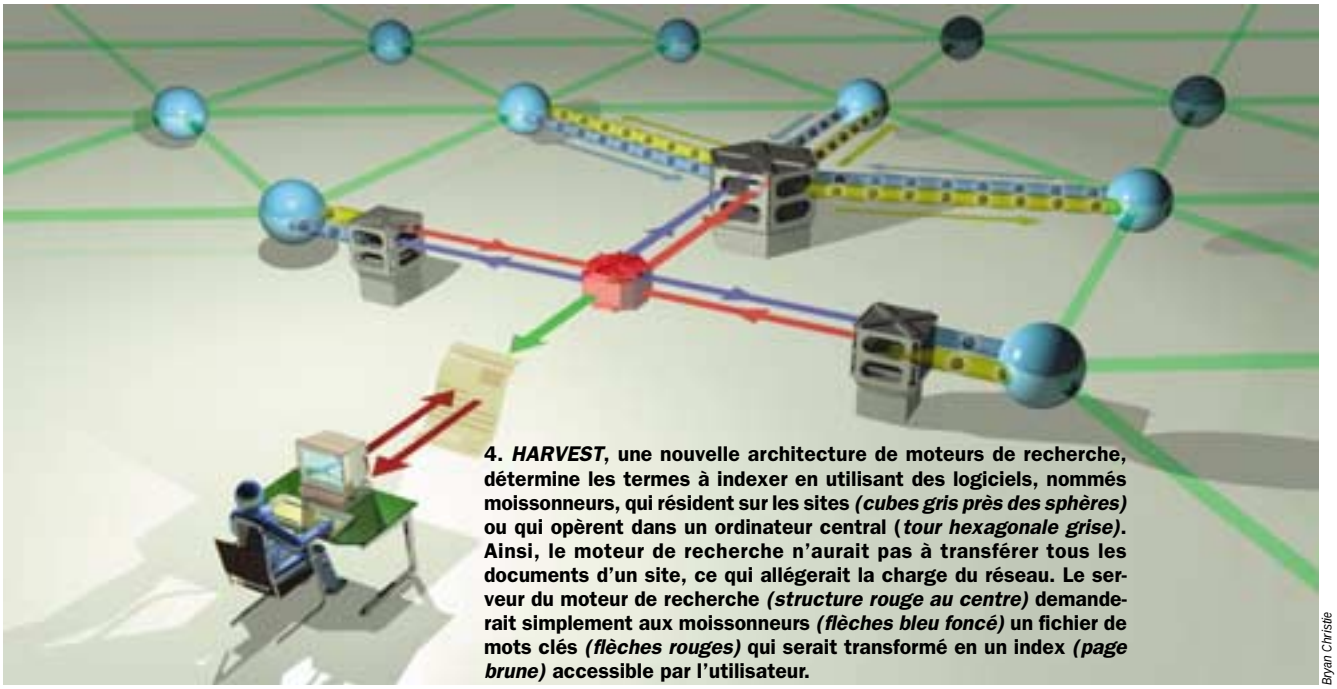
Tous ces programmes cherchent un motif analogue au motif demandé. Ils requièrent encore un observateur humain – ou un texte accompagnateur – pour confirmer si un objet est un chat ou un coussin. Depuis plus de dix ans, des chercheurs en intelligence artificielle essaient de donner aux programmes la possibilité d'établir directement l'identité d'objets présents sur une image. Ils tentent de corréler les formes d'une image avec un modèle géométrique d'objets du monde réel. Les programmes déduisent ensuite, par exemple, qu'un cylindre rose ou brun est un bras humain.

Considérons, par exemple, le logiciel mis au point par David Forsyth, de Berkeley, et par Margaret Fleck, de l'Université d'Iowa, qui recherche des photographies de personnes nues. Il commence par analyser la couleur et la texture d'une photographie. Lorsqu'il trouve des zones de couleur chair, il enclenche un algorithme qui recherche des zones cylindriques pouvant correspondre à un bras ou à une jambe. Il recherche ensuite d'autres cylindres couleur chair, disposés selon certains angles, qui confirmeraient la présence de membres. Lors d'un essai réalisé à l'automne 1996, le programme a retrouvé 43 pour cent des 565 personnes nues présentes sur un total de 4 854 photos – ce qui constitue un bon résultat pour une analyse d'images aussi complexe. En outre, le programme n'a produit que quatre pour cent de fausses réponses pour un ensemble de 4 289 images qui ne contenaient aucune personne nue.

Une dizaine d'années de recherches sera encore nécessaire avant que les utilisateurs du réseau puissent chercher des images, des corps nus, des chats ou des drapeaux aussi facilement qu'ils recherchent du texte. Les informaticiens espèrent qu'ils mettront au point, lors de cette recherche, des programmes qui collecteront l'information en comprenant ce qu'ils voient.



IBM Corporation/Rountech/Corel



4. *HARVEST*, une nouvelle architecture de moteurs de recherche, détermine les termes à indexer en utilisant des logiciels, nommés moissonneurs, qui résident sur les sites *(cubes gris près des sphères)* ou qui opèrent dans un ordinateur central (*tour hexagonale grise)*. Ainsi, le moteur de recherche n'aurait pas à transférer tous les documents d'un site, ce qui allégerait la charge du réseau. Le serveur du moteur de recherche *(structure rouge au centre)* demanderait simplement aux moissonneurs *(flèches bleu foncé)* un fichier de mots clés *(flèches rouges)* qui serait transformé en un index *(page brune)* accessible par l'utilisateur.

Bryan Christie

sujet donné, sans payer les services d'êtres humains chargés de juger la myriade de sites Web. Une solution, qui fait à nouveau appel à l'être humain, consiste à partager entre utilisateurs l'évaluation des sites. Des systèmes d'évaluation numériques récemment mis au point permettent aux utilisateurs de donner leur sentiment sur des sites particuliers (voir *Le filtrage d'information*, par Paul Resnick, page 52).

Certains moteurs explorent l'Internet et, de surcroît, séparent le bon grain de l'ivraie. Il faudrait toutefois de nouveaux programmes pour décongestionner le réseau lors de l'examen répété des sites. Certains gestionnaires de sites affirment en effet que leurs ordinateurs consacrent bien plus de temps à fournir leurs informations aux moteurs qu'aux personnes qu'ils espèrent séduire par leurs offres.

Pour résoudre ce problème, Mike Schwartz et ses collègues de l'Université de Boulder ont écrit un programme, nommé *Harvest*, qui permet à un site de compiler des données d'indexation de ses propres pages et de transmettre, à la demande, ces informations aux différents moteurs de recherche. Ainsi, grâce à *Harvest*, les moteurs n'ont plus à exporter la totalité des sites. Alors que les moteurs de recherche téléchargent un exemplaire de chaque page visitée dans leurs sites d'origine pour en extraire les termes qui constituent un index, ce qui engorge le réseau, *Harvest*, le «programme moissonneur», n'envoie qu'un fichier de termes indexés. En outre, comme il transmet uniquement l'information concernant les pages qui ont été modifiées depuis sa dernière visite, il évite d'encombrer le réseau et les ordinateurs qui lui sont connectés.

Les programmes moissonneurs pourraient également aider les éditeurs à limiter l'information qui quitte leurs sites. Ce contrôle se révèle nécessaire, car le réseau n'est désormais plus un simple support de libre circulation de l'information ; il devient progressivement un moyen d'accéder, contre paiement, à des informations protégées, dont la lecture n'est pas forcément autorisée pour les utilisateurs. Les moissonneurs, eux, distribueraient uniquement l'information que les éditeurs diffusent gratuitement, tels les sommaires et des extraits de l'information stockés sur leurs sites.

Les besoins des utilisateurs détermineront les méthodes de collecte d'information qui seront mises en œuvre. Il est encore trop tôt pour dire si l'Internet ressemblera alors à une bibliothèque, avec des collections structurées, ou s'il restera anarchique, avec un accès fourni par des systèmes automatisés.

Si les utilisateurs acceptent de payer pour protéger le travail des auteurs, alors les éditeurs, les indexeurs et les critiques privilégieront l'information structurée du type bibliothèque classique. En revanche, si l'information est fournie gratuitement ou subventionnée, l'indexation numérisée, peu coûteuse, l'emportera très probablement, et l'environnement anarchique qui caractérise aujourd'hui une grande partie de l'Internet subsistera. Ainsi, plutôt que des problèmes techniques, ce seront surtout des problèmes économiques et sociaux qui détermineront la forme que prendra, dans l'avenir, la recherche de l'information sur l'Internet.

Clifford LYNCH est informaticien à l'Université de Californie.

C.M. BOWMAN et al., *The Harvest Information Discovery and Access System*, in *Computer Network and ISDN Systems*, vol. 28, nos 1-2, pp. 119-125, décembre 1995.

Des informations sur le programme *Harvest* sont disponibles à l'adresse Web http://harvest.transarc.com

Lorcan DEMPSEY et Stuart L. WEIBEL, *The Warwick Metadata Workshop : A Framework for the Deployment of Resource Description*, in *D-lib Magazine*, juillet-août 1996. Les comptes rendus de cette conférence sont disponibles sur http://www.dlib.org/dlib/july96/07contents.html

Carl LAGOZE, *The Warwick Framework : A Container Architecture for Diverse Sets of Metadata*, ibid.