

les documents par la totalité des catégories qui leur correspondent, ainsi que par un autre ensemble d'attributs (par exemple, l'origine, la date, le genre et l'auteur). Les informaticiens de la bibliothèque de l'Université Stanford élaborent actuellement une interface nommée *SenseMaker*, fondée sur ce principe.

Au Centre de recherches Xerox, nous avons mis au point une autre procédure pour grouper les pages fournies par un moteur de recherche. Cette procédure nommée *Scatter/Gather* construit une table des matières qui

se modifie à mesure que l'utilisateur comprend mieux la nature des documents disponibles et découvre ceux qui sont les plus intéressants.

Imaginons qu'effectuant une recherche à l'aide du programme *Excite*, Alice reçoive les 500 premières pages proposées par le réseau. Le système *Scatter/Gather* examine alors ces pages et les répartit en trois à cinq groupes selon leurs ressemblances (voir la figure 4). Alice peut alors examiner ces groupes et choisir ceux qui lui semblent plus intéressants.

Bien que le comportement des utilisateurs soit variable, les expériences indiquent que les arborescences facilitent l'accès aux documents intéressants. Si Alice juge que le groupe des 293 textes résumés par «bulbes», «sol» et «jardinier» est prometteur, par exemple, elle peut utiliser *Scatter/Gather* pour le parcourir et le diviser en sous-groupes plus petits. Après quelques subdivisions, elle ramène son ensemble de 500 pages, pour la plupart sans intérêt, à un groupe de quelques dizaines de pages qui, elles, sont intéressantes.

4. L'INTERFACE *Scatter/Gather* groupe les pages de premier niveau identifiées par le moteur de recherche *Excite*, et il étiquette chaque groupe par des mots qui semblent caractéristiques (ces mots, puisés dans les documents réels, sont probablement différents de ceux que *Yahoo* aurait choisis). L'examen des étiquettes permet à l'utilisateur de décider que, par exemple, le groupe 2 (en vert) semble plus intéressant. Pour avoir une idée plus claire de son contenu, l'utilisateur peut restructurer les 293 documents du groupe 2 en quatre sous-groupes (au milieu). Il peut alors choisir de n'examiner que les documents du deuxième sous-groupe (dans le cadre orange).

5. LE PROGRAMME *TileBars* (à droite) aide l'utilisateur à trouver les articles les plus intéressants d'une liste. L'utilisateur peut décomposer une demande en plusieurs thèmes (en rouge), chacun contenant plusieurs mots clés. Le programme crée alors une fenêtre (en bleu) où il affiche les occurrences trouvées. La taille de la fenêtre est proportionnelle à la longueur du document. Chaque fenêtre est un quadrillage : les colonnes correspondent aux passages dans plusieurs paragraphes de texte, et les lignes aux thèmes de recherche demandés par l'utilisateur. Le programme *TileBars* affiche ainsi immédiatement, pour chaque article, les passages qui contiennent un sujet particulier, indiquant en outre le nombre d'occurrences de chaque sujet (plus les carrés sont sombres, plus les occurrences sont nombreuses). Ici, la ligne supérieure de chaque rectangle indique le nombre d'occurrences des mots «ail» ou «oignon» dans le premier thème ; la ligne du milieu correspond au nombre d'occurrences du mot «automne», etc. Cette interface guide les recherches. Les deux articles supérieurs, où les trois sujets figurent dans plusieurs pas-

sages, seront probablement plus intéressants que le document inférieur, qui ne contient aucune mention des mots «ail» ou «oignon». En cliquant sur un carreau d'une fenêtre, on affiche une fenêtre (en mauve) où s'affiche le passage correspondant à l'intérieur du texte, avec les mots clés mis en relief par des couleurs appropriées.

Cartographie thématique

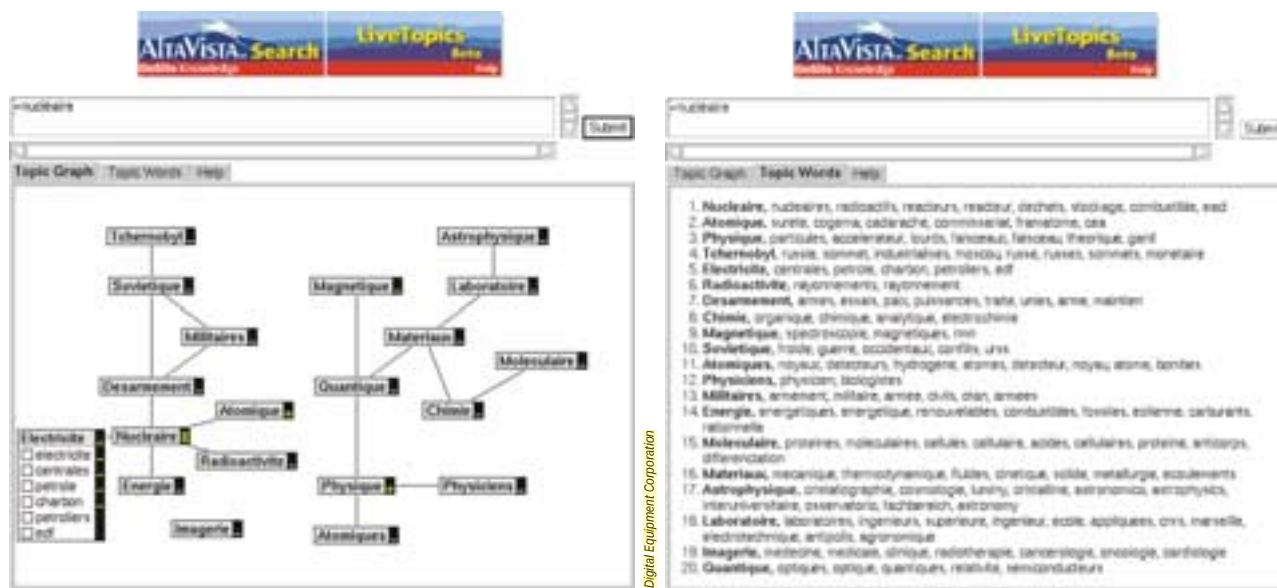
Dans une approche différente de celle des programmes qui aident à visualiser la classification *a priori* des thèmes d'une recherche, tels ceux décrits dans cet article, nous avons conçu un logiciel, nommé *LiveTopics*, qui propose à l'utilisateur, submergé par l'avalanche de documents renvoyés par sa requête, une cartographie thématique et dynamique des documents. À l'aide de cette carte, l'utilisateur appréhende rapidement le contenu des documents. En incluant ou en excluant certains des thèmes de la carte, il affine sa recherche.

Ce logiciel est aujourd'hui proposé sur le moteur de recherche *AltaVista* de la Société *Digital Equipment*. Lorsque l'utilisateur reçoit plus d'une vingtaine de réponses, il peut demander à visualiser une synthèse thématique de l'ensemble des réponses. Par exemple, la recherche du mot «nucléaire» renvoie des thèmes variés où l'on

retrouve ce qui concerne la physique, l'énergie, l'électricité et l'astrophysique (voir la figure). L'utilisateur passe ensuite en revue chacun des thèmes et les mots qui s'y rapportent. Il peut alors modifier sa requête d'un simple clic de la souris, en incluant ou en excluant certains mots, et lancer une nouvelle recherche qui pourra de nouveau donner lieu à une analyse thématique.

Fondé sur une analyse statistique du texte intégral des documents, *LiveTopics* n'utilise aucune analyse linguistique ni sémantique, ce qui rend possible l'analyse de documents indépendamment de leur langue, pourvu que cette langue s'écrive avec des caractères latins. Une évolution vers des caractères d'autres langues est à l'étude, ainsi qu'une version de la technique adaptée au monde des réseaux d'entreprises.

François BOURDONCLE, École des mines de Paris



Dans le moteur de recherche *AltaVista* (<http://www.altavista.digital.com/>), la demande «nucléaire» renvoie 11 877 documents. Lorsque on lance *LiveTopics*, une carte des thèmes apparaît. À gauche, on trouve le thème *Nucléaire* relié à *Énergie*, *Électricité*, *Radioactivité*, mais aussi le thème du désarmement, lié à *Soviétique*, *Tchernobyl* et

Militaires. À droite de la carte apparaissent plutôt les thèmes liés à la recherche : *Quantique*, *Magnétique*, *Physique*, *Chimie*, etc. Notez le thème *Imagerie* qui est isolé du reste. Sur l'écran de droite, on trouve la liste des mots associés aux thèmes de la demande «nucléaire». On peut les exclure ou les inclure pour une nouvelle recherche.

Le groupement de documents n'est pas une panacée, car il n'explique pas comment les documents sont trouvés. En revanche, quand le formulaire d'entrée encourage les utilisateurs à décomposer leurs demandes en plusieurs groupes de mots clés apparentés, une interface graphique peut indiquer les voies de recherche obtenues et les endroits où elles sont apparues dans les documents trouvés. Si tous les sujets se retrouvent au même endroit, le document a plus de chances d'être pertinent, et le programme lui accorde une priorité supérieure. Alice peut n'avoir aucune idée précise des sujets qui doivent apparaître dans le document ou du degré de proximité qui doit les

unir, mais elle a des chances de reconnaître ce qu'elle désire lorsqu'elle le voit et peut alors affiner sa demande. Cette technique, que j'ai nommée *TileBars*, peut guider les utilisateurs lorsqu'ils choisissent les documents qu'ils souhaitent examiner et les conduit plus rapidement aux endroits les plus pertinents.

L'étude de ces nouvelles interfaces utilisateur et techniques d'analyse de texte commence seulement. D'autres techniques combinent méthodes statistiques et empirisme pour résumer automatiquement des documents et les insérer dans un ensemble de catégories déjà existant. Elles proposent des synonymes pour des mots clés et répondent à des questions simples.

Marti HEARST est informaticien au Centre de recherche *Xerox*, à Palo Alto.

Ramana RAO, Jan O. PEDERSEN, Marti A. HEARST et Jock D. MACKINLAY et al., *Rich Interaction in the Digital Library*, in *Communications of the ACM*, vol. 38, n° 4, pp. 29-39, avril 1995.

Stuart K. CARD, George G. ROBERTSON et William YORK, *The WebBook and the Web Forager: An Information Workspace for the World-Wide Web*, in *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, avril 1996. Présent sur le réseau à l'adresse : <http://www.acm.org/sigs/sigchi/ch96/proceedings/papers/Card/sk1txt.html>