Gouvernement du Canada
Ministère des Communications

Government of Canada
Department of Communications

Le Centre canadien de recherche sur l'informatisation du travail
Canadian Workplace Automation Research Centre

LES SYSTÈMES D'INDEXATION INTELLIGENTS [1]

par

Jennifer Farkas

Canada

Communications Canada

Centre canadien de recherche sur l'informatisation du travail

LES SYSTÈMES D'INDEXATION INTELLIGENTS [1]

par

Jennifer Farkas

Laval

juin 1991

[1]Document présenté au congrès en génie électrique et informatique, tenu à Québec, Québec, du 25 au 27 septembre 1991.

Ce document fait état de travaux de recherche réalisés dans le cadre des activités du Centre canadien de recherche sur l'informatisation du travail (CCRIT), du ministère des Communications du Canada. Les opinions exprimées dans ce document n'engagent que l'auteur.

ISBN 0-662-97851-X
No cat. Co28-1/95-1992F

Résumé

Dans cet article, nous étudions la possibilité d'appliquer les techniques de l'intelligence artificielle à l'indexation automatique de textes en langue naturelle. Nous décrivons l'utilisation de thésaurus à base sémantique propre à un domaine particulier et abordons le problème de la création de bases de connaissances appropriées pour les systèmes d'indexation intelligents. Nous examinons également la possibilité d'utiliser l'espace de Hilbert l^2 pour la représentation compacte des documents et la définition de la similitude entre des textes en langue naturelle.

1 Le problème de l'indexation

Les techniques d'indexation purement quantitatives sont inappropriées pour la création d'index fiables dans le cas d'ensembles de documents non homogènes. D'autre part, l'établissement par une personne d'un index de bonne qualité est une activité qui demande beaucoup de travail et qui nécessite un niveau de compétence qu'il est difficile d'atteindre et de maintenir dans la plupart des contextes de gestion. Au Centre canadien de recherche sur l'informatisation du travail (CCRIT), nous avons commencé à appliquer les techniques de l'intelligence artificielle pour résoudre ce problème et nous travaillons à la mise au point d'un système expert pour l'indexation de documents, nommé IndeXpert.

2 Un système d'indexation intelligent

IndeXpert est un système automatique d'indexation de documents qui fait appel aux techniques de l'intelligence artificielle. Il s'agit d'un prototype d'un système interactif bilingue assisté par ordinateur qui exploite des thésaurus propre à un domaine particulier pour obtenir une représentation d'un document par mots clés. IndeXpert, et c'est une de ses caractéristiques distinctives, modélise les connaissances d'un spécialiste humain et, par là, contribue à l'établissement d'une représentation cohérente des documents. La base de connaissances d'IndeXpert sert à la conversion de listes préliminaires de descripteurs en une représentation compacte de documents en fonction de la classification de termes dans des thésaurus particuliers.

Comme tous les systèmes d'indexation de documents ont des entrées constituées de textes en langue naturelle, toute analyse du contenu, selon Salton et Lesk [12], ≪doit intégrer des méthodes cohérentes de normalisation du langage. La construction de dictionnaires appropriés constitue l'une des façons les plus efficaces d'assurer cette normalisation≫. IndeXpert fait appel à des dictionnaires particuliers, à savoir des thésaurus, pour fournir l'interprétation sémantique d'expressions ou de termes rencontrés dans des documents. Cette technique repose sur le fait que ≪les relations entre les termes sont très souvent déterminées au moyen d'un thésaurus. Un thésaurus est une structure qui indique, pour chaque terme, un ensemble de synonymes, un ensemble de termes de sens plus restreint et un ensem-

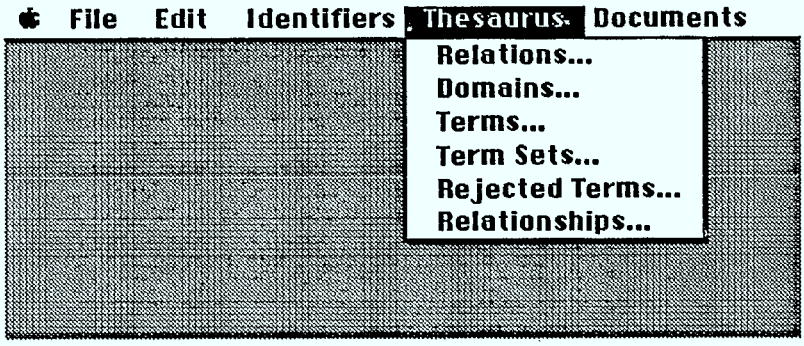

Figure 1 Le module thésaurus d'IndeXpert

ble de termes connexes. Même dans un contexte monolingue, l'emploi d'un thésaurus permet d'accroître les performances d'un système de recherche documentaire.≫ [14]. Cette caractéristique confère une puissance sémantique accrue à n'importe quel système d'indexation automatique, mais nécessite qu'on restreigne celui-ci aux domaines pour lesquels des thésaurus existent déjà ou peuvent être définis. Dans le cas d'IndeXpert, l'intérêt d'avoir recours à des thésaurus est d'autant plus évident que ce système respecte les normes internationalement acceptées de l'ISO, qui établissent des techniques normalisées pour le test de la fiabilité et de la mise au point d'un thésaurus.

3 La représentation vectorielle

Dans l'étude qui suit, nous supposons que le lecteur est familier avec le sous-espace de l^2 (voir [16]) constitué des suites de nombres réels dont seul un nombre fini de coordonnées sont non nulles. Pour la représentation vectorielle des documents, nous supposons un thésaurus T pour IndeXpert dont les termes sont bien ordonnés, soit

$$T = \{t_1, \ldots, t_n\}.$$

Nous supposons également que les documents ont été indexés et que IndeXpert a établi les listes correspondantes de mots clés pour ces documents.

Nous allons définir une fonction qui associe à chaque document un élément du sous-espace de l^2 comme suit. Soit

$$\begin{aligned} D &= \{d_{j_1}, \ldots, d_{j_p}\} \\ D' &= \{d_{k_1}, \ldots, d_{k_q}\} \end{aligned}$$

deux documents indexés, où $d_{j_r} = t_{j_r}$ dans le bon ordre de T, etc. Soit $< d(t_i) >$ la suite de 0 et de 1 obtenue à partir de D de la façon suivante :

$$d(t_i) = \begin{cases} 1 & \text{si } d(t_i) = d_i \text{ pour un } i \\ 0 & \text{autrement} \end{cases}$$

En utilisant la structure du produit scalaire de l^2, nous obtenons :

$$(D, D') = \sum d_i d'_j$$

le produit scalaire de D et D'. On définit la norme de D comme suit :

$$\| D \| = (\sum |d_i|^2)^{\frac{1}{2}}.$$

À partir de là, on définit le cosinus de l'*angle* θ entre deux documents de la façon suivante :

$$cos(\theta(D, D')) = \frac{(D, D')}{\| D \| \| D' \|}.$$

On dit que les documents D et D' sont *semblables* si l'angle θ est petit. Bien entendu, on peut préciser cette définition et l'exprimer en fonction de θ. On observera que, grâce à l'espace infini de dimensions l^2, on peut donner une définition uniforme de la similitude, indépendante de la taille des thésaurus utilisés.

Dans notre contexte, on entend par *document* l'image d'IndeXpert, notée $IndeXpert(A)$, d'un texte en langue naturelle A. On dit que deux textes en langue naturelle A et B sont *congrus* lorsque

$$cos(\theta(IndeXpert(A), IndeXpert(B))) = 1.$$

On dit qu'ils sont *orthogonaux* lorsque

$$cos(\theta(IndeXpert(A), IndeXpert(B))) = 0.$$

3.1 Exemple 1

Supposons que $D = IndeXpert(A)$, l'image du système d'indexation d'un document en fonction des thésaurus de la figure 2, et que $D = \{$accounting, agreement, Aladin, Amethyst, annual report, Argument and Decidex, artificial intelligence, artificial intelligence application, authorization$\}$, et que $D' = IndeXpert(B) = D \cup \{$budget$\}$, on a alors

$$\| D \| = 3 \text{ et } \| D' \| = \sqrt{10}.$$

et

$$cos(\theta(D, D')) = \frac{3}{\sqrt{10}} \approx .95.$$

Par conséquent $\theta(D, D') \approx 18°$. Par rapport à la mesure de la similitude sémantique que nous avons définie, A et B sont plutôt congrus qu'orthogonaux.

3.2 Comptage de fréquences

La définition que nous avons donnée de $d(t_i)$ est naturelle et appropriée à la détermination de la congruence ou de l'orthogonalité des deux documents. Toutefois, les opérations dans l'espace vectoriel l^2 n'interviennent pratiquement pas dans ces calculs. En particulier, n'est pas prise en compte la fréquence d'occurrence d'un terme du thésaurus dans un document, qui est souvent considérée comme un indicateur significatif de la pertinence de ce terme dans la classification du document. Nous allons par conséquent redéfinir $d(t_i)$ comme suit :

$$d(t_i) = \begin{cases} n & \text{si } d(t_i) = d_i \text{ pour un certain } i \\ 0 & \text{autrement,} \end{cases}$$

où n désigne le nombre d'occurrences du terme t_i dans le document indexé.

Avec cette nouvelle définition, l'addition vectorielle prend un nouveau sens. En effet, si $< d(t_i) >$ et $< d'(t_i) >$ représentent deux documents A et B, alors

$$< d(t_i) + d'(t_i) >$$

représente le document C obtenu de A et B en annexant B à A.

3.3 Exemple 2

Supposons que D et D' soient les documents de l'exemple 1 et supposons que le terme ≪*accounting*≫ figure deux fois dans le document D et que le terme ≪*budget*≫ figure quatre fois dans le document D', et que tous les autres termes du thésaurus apparaissent exactement une fois dans chaque document, on a alors

$$\| D \| = \sqrt{12} \text{ et } \| D' \| = 5 \text{ et } (D, D') = 10.$$

Le cosinus de l'angle θ entre les deux documents a maintenant pour valeur

$$cos(\theta(D, D')) = \frac{10}{(\sqrt{12})(5)} = \frac{1}{\sqrt{3}} \approx .58.$$

Par conséquent

$$\theta(D, D') \approx 55°.$$

Comme on pouvait s'y attendre, la similitude entre les documents A et B est moindre par rapport à la nouvelle mesure que par rapport à la mesure originale.

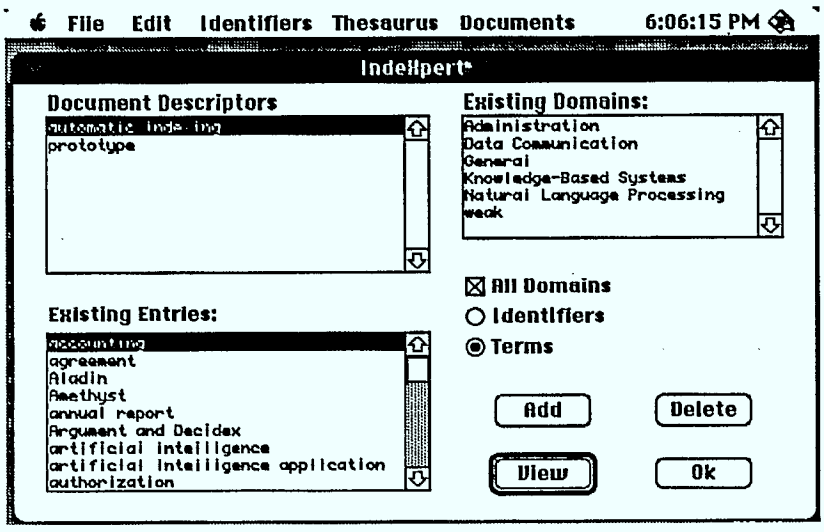

Figure 2 Le module d'indexation d'IndeXpert

4 Bases de connaissances

IndeXpert n'apporte qu'une solution partielle au problème de l'indexation sémantique automatique, puisqu'il est bien connu que des techniques purement syntaxiques ne permettent pas de saisir la signification contextuelle de tous les termes dans tous les contextes [11]. Le système est par conséquent conçu pour formaliser les méthodes d'indexation de l'homme, et ce toutefois que pour certains domaines. La composante intelligente d'IndeXpert réside dans sa formalisation de fragments significatifs gérables de la pensée cognitive. Le système offre à l'utilisateur une option de simulation qui permet l'indexation d'un document donné pour tous les domaines existants, ou pour un sous-ensemble de l'ensemble des domaines. Dans le cas de documents particuliers, on peut obtenir de meilleurs résultats en spécifiant le domaine traité. La fenêtre de la figure 2 illustre cette possibilité du système IndeXpert. La fenêtre ≪ *domaines existants* ≫ (Existing Domains) affiche les domaines existants, la fenêtre ≪*entrées existantes*≫ (Existing Entries) permet à l'utilisateur de visualiser l'ensemble T des termes du thésaurus avec lequel le document est indexé, et la fenêtre ≪ *descripteurs du document* ≫ (Docu-

ment Descriptors) affiche la liste des mots clés établie par IndeXpert pour le document selon les domaines sélectionnés. Un répertoire de désignations dans la fenêtre ≪ *identificateur* ≫ (Identifier) permet à l'utilisateur de sélectionner des options qui n'apparaissent pas dans les thésaurus.

IndeXpert comporte plusieurs améliorations par rapport aux systèmes existants.

1. Le système fait appel aux techniques de l'intelligence artificielle, lesquelles sont en train de devenir un outil indispensable pour l'indexation efficace et fiable des documents. Comme cela est mentionné dans [11], ≪la syntaxe seule ne permet pas de résoudre les nombreuses ambigüités qui compliquent l'analyse du contenu. On a fait récemment diverses tentatives d'utilisation de méthodes d'analyse syntaxique pour la génération de constructions complexes, comme les syntagmes nominaux ou prépositionnels, qui sont essentielles à la détermination du contenu pour divers systèmes automatiques d'analyse de textes≫. Un des buts visés par IndeXpert est précisément de résoudre ce problème.

2. Le fait de limiter le problème à l'indexation de domaines spécifiques et à l'emploi de thésaurus propres à un domaine particulier accroît la fiabilité, la rapidité, la précision et a mise sur pied d'une recherche documentaire.

3. Le recours à des bases de connaissances particulières pour simuler les compétences d'un spécialiste humain de l'indexation. Comme on le fait remarquer en [17], ≪les décisions prises par ces spécialistes dans le choix des descripteurs représentent, pour une seule base de données, un gros investissement intellectuel. Comme la plupart des spécialistes sont des professionnels du domaine sur lequel ils travaillent, leurs efforts pour indexer les documents constituent un ensemble de décisions d'expert sur le contenu de la littérature≫.

4. La plupart des systèmes d'indexation existants utilisent des ordinateurs centraux. Le fait que IndeXpert est conçu pour fonctionner de façon autonome sur PC étend le spectre des utilisateurs potentiels.

5 Travaux futurs

Le prototype actuel d'IndeXpert a été mis au point à partir du langage d'intelligence artificielle Prolog sur Macintosh. Il lit des documents en anglais ou en français, détermine la langue du document, choisit les thésaurus appropriés et indexe automatiquement le document. Actuellement, sa base de connaissances est constituée de toute une variété de règles d'indexation générales qui dépendent du thésaurus utilisé. Nous envisageons d'améliorer IndeXpert en ajoutant à l'ensemble des règles existantes des règles particulières, intégrant les méthodes de travail des spécialistes de l'indexation. On peut s'attendre à améliorer ainsi de façon significative la qualité de la gestion des documents dans des domaines spécifiques.

Bibliographie

[1] G. Biswas, J. C. Bezdek, M. Marques and V. Subramanian, "Knowledge-Assisted Document Retrieval: I. The Natural-Language Interface", *Journal of the American Society for Information Science,* vol. 38 (1987) 83-96.

[2] G. Biswas, J. C. Bezdek, V. Subramanian and M. Marques, "Knowledge-Assisted Document Retrieval: II. The Retrieval Process", *Journal of the American Society for Information Science,* vol. 38 (1987) 97-110.

[3] W. B. Croft and R. H. Thompson, "I^3R: A New Approach to the Design of Document Retrieval Systems", *Journal of the American Society for Information Science,* vol. 38 (1987) 389-404.

[4] J. R. Driscoll et al., "The Operation and Performance of an Artificially Intelligent Keywording System", *Information Processing and Management,* vol. 27 (1991) 43-54.

[5] D. Harman, "An Experimental Study of Factors Important in Document Ranking", in: (Fausto Rabitti, editor), 1986—ACM Conference on Research and Development in Information Retrieval, 8-10 September 1986, Pisa, Italy, 186-191.

[6] H. J. Jeffrey, "Expert Document Retrieval via Semantic Measurement", *Expert Systems with Applications,* vol. 2 (1991) 345-352.

[7] F. W. Lancaster, "Vocabulary Control for Information Retrieval", Information Resources Press, Arlington, Virginia, 1986.

[8] L. C. Malone, J. R. Driscoll and J. W. Pepe, "Modeling the Performance of an Automated Keywording System", *Information Processing and Management,* vol. 27 (1991) 145-151.

[9] "Norme internationale ISO (5964)", Documentation-Principes directeurs pour l'établissement et le développement de thésaurus multilingues, 1985.

[10] G. Salton, "Automatic Text Processing", Addison-Wesley, New York, 1989.

[11] G. Salton, C. Buckley and M. Smith, "On the Application of Syntactic Methodologies in Automatic Text Analysis", *Information Processing and Management,* vol. 26 (1990) 73-92.

[12] G. Salton and M. E. Lesk, "Information Analysis and Dictionary Construction", in: (G. Salton, editor), The SMART Retrieval System, Prentice-Hall Inc. (1971) 115-142.

[13] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval", McGraw-Hill, New York, 1983.

[14] P. Schäuble, "Improving the Effectiveness of Retrieval Systems by Information Structures", *Information Processing and Management,* vol. 25 (1989) 363-376.

[15] L. C. Smith, "Artificial Intelligence and Information Retrieval", in: (Martha E. Williams, editor), *Annual Review of Information Science and Technology,* vol. 22 (1987) 41-77.

[16] A. E. Taylor, "Introduction to Functional Analysis", John Wiley and Sons, New York, 1964.

[17] C. Todeschini and M. P. Farrell, "An Expert System for Quality Control in Bibliographic Databases", *Journal of the American Society for Information Science,* vol. 40 (1989) 1-11.

**Pour plus de détails,
veuillez communiquer avec :**

*Le Centre canadien de recherche
sur l'informatisation du travail*
1575, boulevard Chomedey
Laval (Québec)
H7V 2X2
(514) 682-3400

**For more information,
please contact:**

*Canadian Workplace
Automation Research Centre*
1575 Chomedey Blvd.
Laval, Quebec
H7V 2X2
(514) 682-3400