REVUE FRANÇAISE D'AUTOMATIQUE, INFORMATIQUE, RECHERCHE OPÉRATIONNELLE. RECHERCHE OPÉRATIONNELLE

JEAN-MARIE BOUROCHE
BERNARD CURVALLE

La recherche documentaire par voisinage

Revue française d'automatique, informatique, recherche opérationnelle. Recherche opérationnelle, tome 8, n° V1 (1974), p. 65-96

<http://www.numdam.org/item?id=RO_1974__8_1_65_0>

R.A.I.R.O.
(8e année, V-1, 1974, p. 65-96)

LA RECHERCHE DOCUMENTAIRE PAR VOISINAGE (1)

par Jean-Marie BOUROCHE (2) et Bernard CURVALLE (3)

Résumé. — *Cet article présente les conclusions d'une étude ayant pour but d'explorer les possibilités d'un nouveau mode de recherche documentaire s'appuyant sur la notion de voisinage.*

Il s'agit de caractériser des concepts, non plus seulement par rapport à des mots du langage naturel ou d'un quelconque langage documentaire, mais par rapport à un certain nombre de points de vues ou critères dichotomiques permettant d'établir entre ces concepts des relations de proximité ou de distance.

La première partie décrit une indexation réalisée selon ces principes, et en analyse les résultats obtenus au moyen de programmes de structuration de données, sur ordinateur CDC 6600.

La deuxième partie analyse les résultats d'une recherche documentaire menée à partir d'une telle indexation.

Les auteurs tiennent à exprimer leurs remerciements à M. Jean Donio dont les idées sont à l'origine de ce travail et dont les conseils ont favorisé le déroulement de l'étude ainsi qu'à M. le Professeur Benzecri qui a bien voulu leur faire part de ses remarques; enfin à l'équipe du Service d'Information et de Documentation de la SEMA qui a assumé la tâche ingrate d'indexer des documents selon des principes tout à fait nouveaux.

I. INTRODUCTION : LES HYPOTHESES DE BASE

1. La recherche documentaire « classique »

Les recherches documentaires automatiques s'appuient, quel que soit leur degré de raffinement, sur le même processus opératoire : comparer un corpus de documents caractérisés par des mots-clés, ou des codes alpha-numériques

(1) Cette recherche a bénéficié du soutien financier du Comité de Recherche en Informatique (Contrat n° 70 050).
(2) SEMA (Metra International), Direction Scientifique, Montrouge.
(3) SEMA (Metra International), Service d'Information et de Documentation.

ayant la même valeur, à une question elle-même traduite en mots-clés ou en codes.

Lorsqu'il s'agit d'un simple lexique de mots-clés isolés, l'extraction des documents pertinents obéit à la loi du tout ou rien : si le mot-clé de la question figure dans un document, celui-ci est déclaré pertinent, sinon il est ignoré.

Pour pallier cet inconvénient, responsable de « silences », on a imaginé divers outils (langages documentaires) destinés à introduire des relations entre les différents mots-clés :

— classifications hiérarchiques permettant de chaîner les termes en allant du plus général au plus spécifique,

— schémas fléchés visualisant les interrelations de concepts à l'intérieur d'une même famille,

— thesauri favorisant par divers systèmes de renvois réciproques, l'utilisation de mots-clés représentatifs tant au niveau de l'analyse du document qu'à celui de la formulation de la question (ces mots-clés tirés d'un vocabulaire figé sont alors appelés descripteurs, mais on négligera cette nuance).

Si l'on arrive, grâce à ces divers artifices, à limiter les « silences », il n'en reste pas moins que toute interrelation, non prévue à l'avance, entre deux mots constitue en elle-même une perte potentielle au niveau de l'exhaustivité de la réponse. A cela s'ajoute, au travers des relations illicites introduites entre certains concepts qui n'ont rien de commun, une possibilité d'erreur intéressant la pertinence de la réponse : un document peut être retenu parce qu'il correspond à la question du point de vue du langage documentaire, alors qu'en réalité il n'est pas pertinent au niveau des concepts. Il constitue, dans l'ensemble des documents extraits, un « bruit ».

Le système est donc très rigide car il ne sélectionne un document qu'en fonction de la présence d'un élément ou de ses dérivés, sans possibilité de jouer sur des états intermédiaires.

2. La recherche documentaire par voisinage

L'idée de base de la *recherche par voisinage* réside, au contraire, dans la définition d'une distance entre mots-clés et documents qui s'appuie, non plus sur des composantes discrètes, mais sur des éléments d'un espace métrique.

Pour matérialiser cette notion de distance, on a imaginé une indexation reposant, non plus sur la caractérisation d'un concept par un mot du langage courant ou par un élément d'un langage documentaire, mais par un vecteur de 1 ou de 0 qui constituent autant de questions positives ou négatives posées à l'indexation à propos de ce concept. Ainsi demandait-on par exemple : le concept à caractériser est-il un objet : oui (1) ou non (0), etc... On arrivait alors à caractériser un concept par rapport à un certain nombre de critères permettant non seulement d'affirmer que ce concept était identique à un autre

(vecteurs de réponses aux critères identiques) mais aussi qu'il était voisin d'un autre (vecteurs identiques à une ou deux composantes près).

Ces critères sont voisins des « propriétés » que G. Salton ([13], pages 53-57) utilise pour distinguer des homonymes au cours des opérations de construction de thesaurus.

Le tableau ci-dessous donne un exemple d'une indexation par voisinage comparée à une indexation classique.

CONCEPT	INDEXATION CLASSIQUE (descripteurs)	INDEXATION PAR POINTS DE VUE (critères)													
		Abstrait	Intellectuel	Être vivant	Intermédiaire	Amélioration	Production	Objet	Offre	Croissance	Automatisme	Périodicité	Durée	Pluralité	Dynamique
Produits destinés à fertiliser le sol	ENGRAIS	0	0	0	1	1	1	0	1	1	0	1	1	1	0
Dérivés du pétrole	PRODUITS PÉTROLIERS	0	0	0	1	1	1	0	1	1	0	1	1	1	0
Utilisation d'un ordinateur par plusieurs personnes simultanément	TIME-SHARING	1	1	0	1	1	0	0	1	1	1	1	1	1	1

II. VERIFICATION DES HYPOTHESES DE BASE : LA PROCEDURE EXPERIMENTALE

1. Choix des domaines et du jeu de critères

Il était facile de concevoir a priori que les meilleurs résultats seraient obtenus si l'on disposait, pour chaque concept (ou mot-clé) à caractériser, d'un vecteur de critères particulièrement caractéristique. Il ne pouvait donc être question de chercher d'emblée des critères à la fois assez généraux et assez spécifiques pour s'appliquer à toutes sortes de documents.

On a donc limité l'expérience préalable à deux domaines assez éloignés l'un de l'autre pour que l'on puisse au moins retrouver, après caractérisation des mots-clés par critères et recherche de distances, deux familles de concepts bien distinctes.

L'expérience a porté sur 100 mots-clés dont une moitié était axée sur l'informatique, l'autre sur le marketing. On a caractérisé ces mots-clés au moyen d'un jeu de 59 critères choisis forcément parmi des mots du langage courant. On a donc obtenu pour chacun des 100 mots-clés un vecteur à 59 composantes de valeur 1 ou 0.

La figure 1 donne la liste des 59 critères retenus pour l'expérience, et sélectionnés en fonction de leur richesse en contenu sémantique.

Figure 1

Liste des critères

C01	Théorie	C31	Agrément
C02	Science	C32	Économie (argent)
C03	Spécificité	C33	Physique
C04	Initialisation	C34	Collectif
C05	Durée	C35	Périodique
C06	Fin	C36	Profondeur
C07	Intermédiaire (échange)	C37	Aide
C08	Relatif	C38	Liaison
C09	Production	C39	Fiction, Immatérialité
C10	Accessoire	C40	Mouvement, Dynamique
C11	Fonction primaire (agriculture, terre)	C41	Contrôle
C12	Fonction secondaire (industrie)	C42	Découpage
C13	Fonction tertiaire (services)	C43	Localisation
C14	Manuel	C44	Offre
C15	Mécanique	C45	Nouveauté
C16	Automatisme	C46	Recherche
C17	Être vivant	C47	Plusieurs choses
C18	Chose	C48	Forme
C19	Notion	C49	Gouvernement
C20	Cycle	C50	Matière première
C21	État normal	C51	Matière grise
C22	Statique	C52	Amélioration
C23	Volume	C53	État
C24	Loisirs	C54	Action
C25	Chose unique	C55	Modification
C26	Supplémentaire	C56	Direction
C27	Croissance	C57	Énergie
C28	Éternité	C58	Prévision
C29	Court Terme	C59	Experimentation
C30	Long terme		

On notera qu'ont été introduites dans ce jeu de critères un certain nombre de redondances (cycle, périodique, collectif, plusieurs choses) destinées à tester la logique des indexeurs, et par là la validité du modèle.

2. Recueil des données

On s'est efforcé, bien qu'il s'agisse d'une expérience limitée, de se placer d'emblée dans les conditions réelles, ou supposées telles, d'un travail effectif.

L'indexation a été effectuée par six personnes auxquelles on avait seulement expliqué la signification à donner aux mots-clés, à l'exception de toute indication concernant les critères.

Les indexeurs ont été priés de ne pas trop s'attarder sur un critère, ni d'opérer des retours en arrière, de peur qu'une réflexion trop prolongée leur fasse rechercher des associations mots-clés/critères au second degré.

L'ensemble des données correspondant aux réponses a été consigné sur des feuilles de perforation où figuraient, outre le numéro de mot-clé et les initiales de l'indexeur, les vecteurs à 59 composantes binaires.

3. Résultats

3.1 L'analyse factorielle

La prise en compte des données recueillies a permis de construire un tableau des fréquences d'association mot-clé/critère, puis de visualiser, grâce à l'analyse factorielle des correspondances (ANAFACO) la dispersion des mots-clés et des critères dans l'espace.

La méthode est la suivante :

On considère le nuage des 100 mots-clés dans l'espace $\mathbf{R}^{59}$ des critères.

Soit n_{ij}, $0 \leqslant n_{ij} \leqslant 6$ le nombre de personnes qui pensent que le critère C_j s'applique au mot clé m_i.

Soit

$$n_{i.} = \sum_j n_{ij}$$

$$n_{.j} = \sum_i n_{ij}$$

$$n_{..} = \sum_{i,j} n_{ij}$$

Le mot clé m_j est alors repéré dans $|\mathbf{R}^{59}$ par le vecteur :

$$\underline{m}_i = \begin{bmatrix} \frac{n_{i1}}{n_{i.}} \\ . \\ . \\ . \\ \frac{n_{ij}}{n_{i.}} \\ . \\ . \\ . \\ . \\ \frac{n_i 59}{n_{i.}} \end{bmatrix} \qquad \text{et est affecté du poids } \frac{n_{i.}}{n_{..}}$$

De plus on munit $|\mathbf{R}^{59}$ de la métrique du χ^2 de telle sorte que la distance entre m_i et m_k soit :

$$d^2(\underline{m}_i, \underline{m}_k) = \sum_{j=1}^{59} \frac{n..}{n.j} \left[\frac{n_{ij}}{n_{i.}} - \frac{n_{kj}}{n_{k.}} \right]^2$$

La méthode consiste alors à rechercher dans $|\mathbf{R}^{59}$ les axes principaux d'inertie du nuage des 100 mots clés, c'est-à-dire les axes d'allongements maximums du nuage.

On projette alors les mots clés dans les différents plans correspondant aux couples d'axes (1, 2), (1, 3), (2, 3), etc...

On peut répéter la même opération en plaçant les 59 critères dans l'espace $|\mathbf{R}^{100}$ muni de la métrique χ^2. On recherche de même les axes d'inertie.

Finalement on effectue une représentation simultanée des mots clés et des critères dans les plans d'inertie; cela permet de faire apparaître des proximités intéressantes.

Pour plus de détails sur la méthode, on pourra se reporter à J. P. Benzecri [2].

A l'aide de l'analyse factorielle on a extrait cinq facteurs principaux expliquant 52 % de la dispersion du phénomène. A titre d'exemple, on a représenté en figure 2 les plans correspondant aux couples de facteurs 1-3.

Une fois visualisée la dispersion, on a tenté de donner un nom synthétique aux différents axes. L'interprétation en terme de critères est donnée dans le tableau 1 et en figure 2 (où les points précédés d'un *C* correspondent aux critères de la figure 1). Il a été procédé de la même manière à une interprétation en terme de mots-clés : nous ne la donnons pas ici afin d'alléger le texte.

TABLEAU I. — *Interprétation de l'analyse factorielle en termes de critères*

	CORRÉLATIONS POSITIVES		CORRÉLATIONS NÉGATIVES		INTERPRÉTATION (recherche de noms de critères synthétiques)
Premier axe (seuil de corrélation = 0.5)	Théorie Notion Matière grise Énergie	.52 .59 .54 .65	Accessoire Manuel Mécanique Automatique Chose Statique Forme	— .66 — .55 — .77 — .62 — .80 — .66 — .66	Ce qui est intellectuel et abstrait s'oppose à ce qui est matériel et concret.
Deuxième axe (seuil de corrélation = 0.45)	Théorie Science Durée Notion Profondeur Fiction Recherche	.62 .62 .46 .49 .53 .51 .50	Intermédiaire Fonction tertiaire Être vivant Économie Offre	— .50 — .55 — .72 — .45 — .56	Ce qui est scientifique s'oppose aux notions de commerce et d'économie et à ce qui est vivant.
Troisième axe (seuil de corrélation = 0.40)	Automatique Être vivant Aide Mouvement Action	.48 .52 .43 .43 .40	Fonction primaire Volume Loisirs Croissance Agrément Économie Collectif Offre Matière première	— .51 — .52 — .52 — .48 — .45 — .51 — .49 — .42 — .47	Ce qui est travail et dynamique s'oppose à ce qui est repos et stabilité.

En analysant le tableau 1 et les graphiques correspondants, dont on n'a donné ici en figure 2 que celui qui correspond aux axes 1-3, on s'aperçoit que l'ensemble des réponses est assez cohérent et qu'il est possible d'éliminer les critères redondants puisque certains, se comportant de la même façon à l'analyse, apportent une information tautologique.

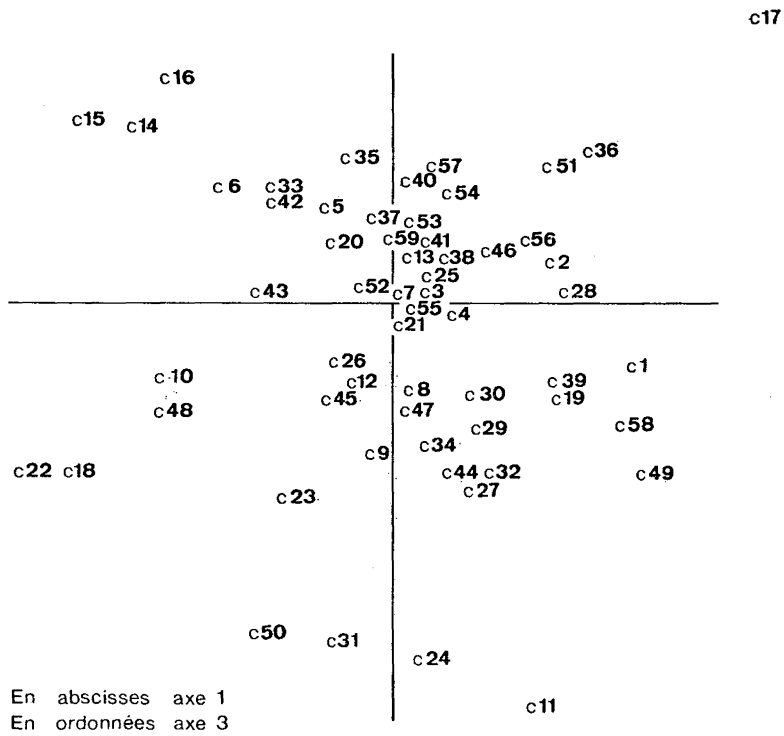

Figure 2

Résultat de l'analyse factorielle

3.2 L'analyse des proximités entre mots-clés

A la suite des résultats obtenus par l'analyse factorielle sur les critères, on a essayé d'approfondir la notion de proximité au niveau des mots-clés par application de trois méthodes permettant de réaliser :

— calcul de distances,

— partition des mots-clés en sous-groupes homogènes,

— classification hiérarchique des différents sous-groupes.

Dans cette phase de l'expérience on a calculé une distance « classique » entre mots-clés. Au niveau d'un individu on obtient :

$$d(\underline{m_j}, \underline{m_k}) = \left(\sum_{i=1}^{59} (C_{ij} - C_{ik})^2 \right)^{1/2}$$

c'est-à-dire la racine carrée du nombre de critères sur lesquels les mots-clés m_j et m_k sont jugés différents par l'individu considéré. Au niveau des trois indexeurs conservés au cours de cette phase, la distance globale calculée est identique mais la somme est étendue aux trois individus.

Sur les tableaux de distance on a appliqué la méthode du « quick clustering » basée sur le principe suivant : chaque mot-clé est associé au mot-clé dont il est le plus proche. On obtient ainsi une partition des mots-clés en classes. On a ensuite calculé des distances inter-classes (distance moyenne) et on a regroupé les classes à l'aide de deux algorithmes de classification hiérarchique (ultramétrique inférieure maximum et une solution ultramétrique supérieure minimum). Pour plus de détail sur ces méthodes on pourra se reporter à D. Borionne [3].

La classification des mots-clés a fourni en moyenne pour trois indexeurs 23 groupes différents qui peuvent se représenter sous forme de graphes. A l'intérieur de ces différents graphes, chaque mot est relié par un arc orienté au mot dont il est le plus proche.

La figure 3 reproduit à titre d'exemple le groupe n° 12 qui représente, avec leurs distances réciproques, ceux des 100 mots-clés regroupés grâce à trois indexages selon 59 critères en une « famille ».

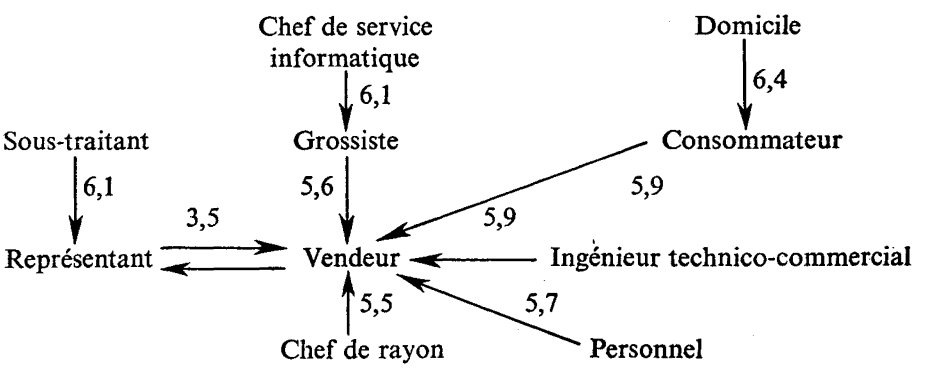

Figure 3

Groupe numéro 12

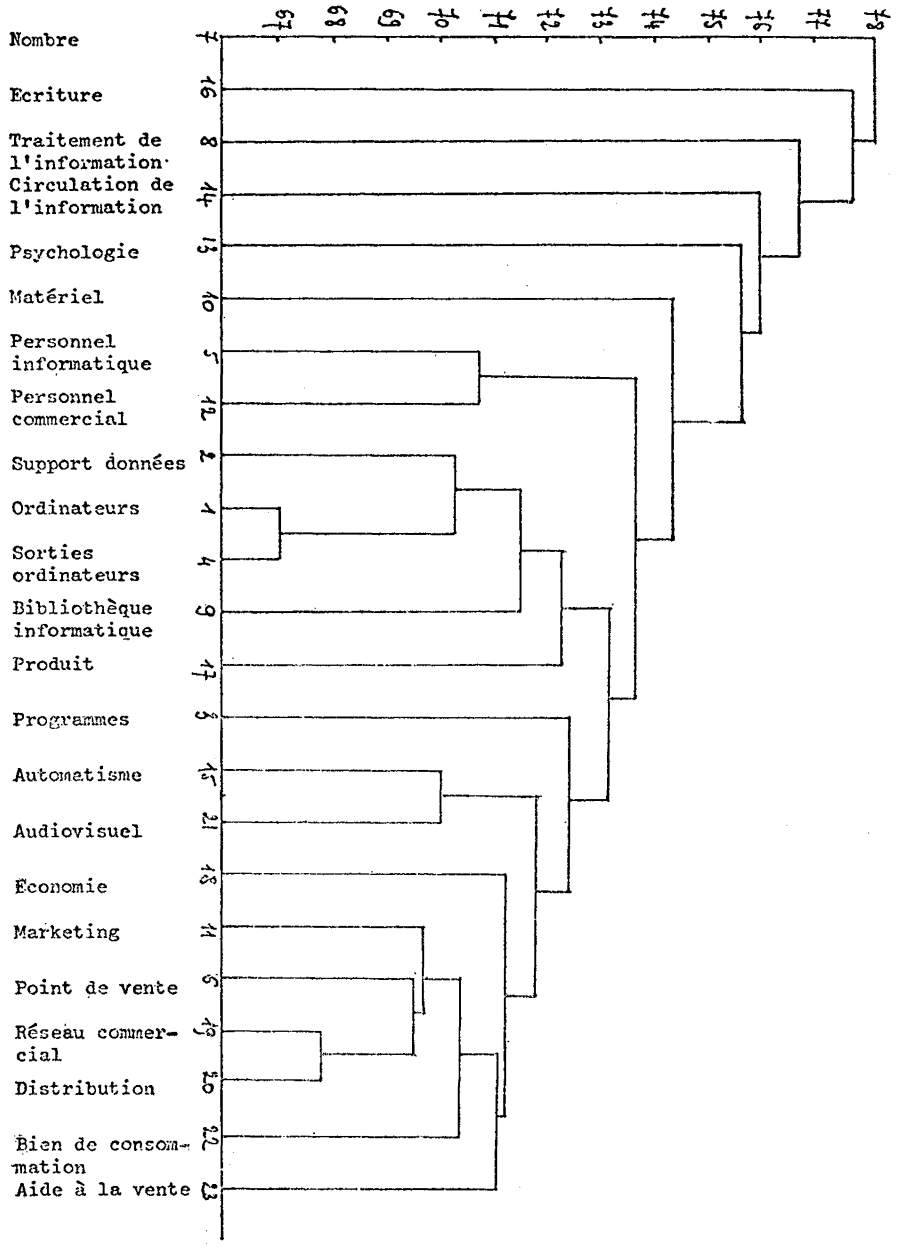

Figure 4

Classification hiérarchique des groupes de mots-clés

On a ensuite calculé un tableau de distance moyenne inter-groupes : le premier algorithme de classification hiérarchique (ultramétrique inférieure maxima) nous a donné le résultat suivant (fig. 4) où les différents groupes ont été nommés en fonction de ce que représentait leur « noyau », par exemple « personnel commercial » pour le groupe 12 déjà cité.

On constate que, à quelques perturbations près (inversion des groupes n° 17 et 3 les classes de mots-clés s'appliquant) :

— au marketing se regroupent entre elles (à droite du graphique),

— à l'informatique se regroupent entre elles (au centre),

et que les classes de mots-clés moins spécifiques de ces deux domaines sont rejetées à gauche et viennent se raccorder à l'arbre au niveau le plus haut (distance maximale ou homogénéité minimale).

Il est permis de penser qu'en utilisant les deux types de représentations précédents (graphes symbolisant les familles de mots-clés et arborescences permettant de dichotomiser les groupes de mots ou même les mots-clés eux mêmes), on ait la possibilité de dégager des voies nouvelles de recherche :

— confection automatique de schémas fléchés du type de ceux qui sont largement utilisés pour l'indexation des documents notamment par Van Dijk et Szanto [15];

— mise au point d'arborescences plus fines en essayant de « nommer » les nœuds de l'arborescence pour aboutir à une hiérarchie plus significative ou à un jeu de critères plus adapté à tel ou tel ensemble de mots-clés.

3.3 Résultats de l'analyse des liaisons entre critères : réduction du jeu de critères

Il s'agissait ici d'opérer une sélection parmi les 59 critères retenus en première analyse.

Cette sélection s'est opérée de plusieurs manières :

a) élimination des critères redondants,

b) élimination des critères peu discriminants,

c) élimination des critères trop imprécis.

C'est ainsi que le critère « cycle » a été éliminé pour la raison *a*) au profit du « périodique » avec lequel il avait de très fortes affinités (on retrouve ici une redondance introduite volontairement pour vérifier le caractère non aléatoire des réponses).

De même le critère « spécificité » a été éliminé pour la raison *b*), les indexeurs ayant jugé que chaque mot-clé avait un caractère spécifique marqué : réponse 1 pour l'ensemble des mots-clés, très peu de réponses 0. « Profondeur » a été de même éliminé : réponse 0 pour la plupart des mots-clés, très peu de réponses 1. « État normal » a été éliminé pour la raison *c*) : réponse confuse pour 59 mots-clés sur 100, et pas de mots-clés pour lesquels toutes les réponses ont été identiques.

La classification des critères obtenus à partir de la matrice de corrélation (incrementée de + 1 pour des raisons de commodité) a permis de former, de la même manière que pour les mots-clés, un certain nombre de groupes (16) dont le 7e est représenté en figure 5.

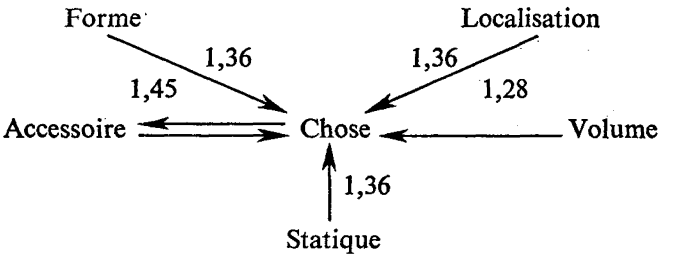

Figure 5
Groupe numéro 7

A partir de la classification des critères et de l'analyse des corrélations (tableau 1), on a isolé 14 critères dont 10 ont été désignés par un mot figurant déjà dans la première liste, et 4 ont été introduits pour tenir compte de notions générales apparues au cours de l'interprétation.

La liste finale retenue a été la suivante (on a mis en italiques les termes réellement nouveaux) :

Abstrait, *Intellectuel*, Être vivant, Intermédiaire, Amélioration, production, *objet*, offre, croissance, automatisme, périodicité, durée, pluralité, *dynamique*.

Cette réduction du jeu de critères de 59 à 14 a évidemment détruit un certain nombre de voisinages.

A titre d'exemple, la figure 6 redonne la famille de mots-clés à laquelle a été donné le nom de « personnel commercial » (fig. 2), amputée (mots encadrés) des voisinages détruits par la réduction du jeu de critères.

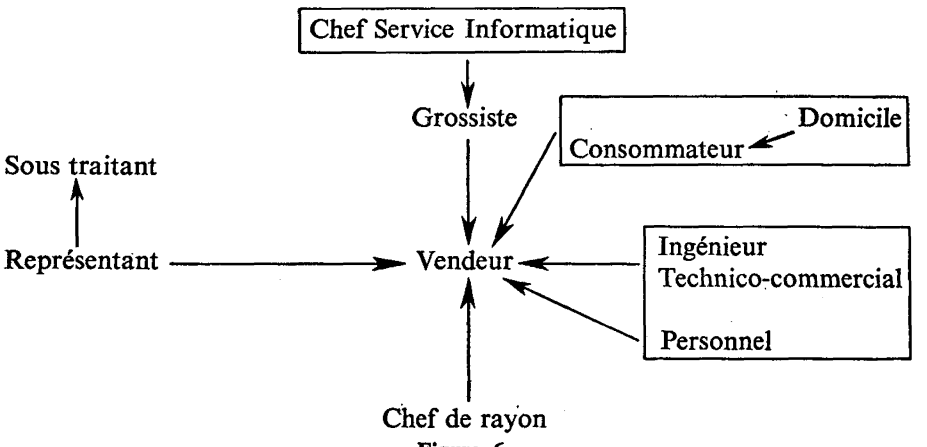

Figure 6

Pour plus de détails sur les modifications induites par la réduction du jeu de critères, on se reportera à l'article paru dans les Cahiers de l'IRIA [6] où l'on trouvera par ailleurs l'ensemble des opérations de vérification des hypothèses de base telles qu'exposées dès juillet 1971 dans le premier rapport d'études [4].

Le jeu de 14 critères que nous avons obtenu à la fin de toutes ces opérations va permettre de passer à la recherche documentaire proprement dite.

III. LE PROCESSUS DE LA RECHERCHE DOCUMENTAIRE PAR VOISINAGE

1. L'indexation préalable

Pour réaliser la recherche documentaire proprement dite, on a travaillé sur un fonds documentaire constitué d'une centaine de très anciens rapports d'études du Groupe METRA, relevant des domaines de l'informatique et du marketing.

Le choix des domaines avait l'avantage de permettre la réutilisation de mots-clés déjà indexés selon les critères au cours de la procédure expérimentale de vérification des hypothèses de base.

On n'a pas choisi systématiquement de réindexer les documents au moyen de mots déjà traduits en critères. Il était au contraire intéressant d'utiliser n'importe quel mot-support du langage commun pour pouvoir tester par la suite l'apparition de voisinages ou d'identités strictes (synonymies au sens du jeu de critères). C'est ainsi qu'ont été utilisés conjointement ou concurremment les mots essence, pétrole, produits pétroliers... sans aucun souci de normalisation ou de hiérarchisation entre eux. On a ainsi affecté en moyenne quatre mots-clés par document, mots-clés qui ont été analysés par six indexeurs selon 14 critères.

On s'est donc placé dans les conditions limites de la méthode : domaines distincts, indexeurs venant d'horizons différents, jeu de critères très restreint...

C'était choisir la difficulté, d'autant qu'avec un nombre non fixe de mots-clés par document, sous tendant eux-mêmes des concepts très différents (ex. : concept de vente associé à un concept de produit) ou très voisins (ex. : concept de produit associé à un produit semblable, comme essence et pétrole) on avait toutes les chances de provoquer des aberrations dans les rapprochements provoqués par le calcul statistique.

A partir de cette indexation des mots-clés on a pu définir des distances entre mots-clés, entre questions et documents et entre documents. On expose dans les deux paragraphes suivant quelques problèmes posés par le choix de ces distances. Utilisant ensuite ces notions on a pu mettre en œuvre une procédure de recherche par voisinage et effectuer une typologie des documents.

2. Problèmes posés par la définition des distances

La méthode de recherche par voisinage est basée sur la notion de *distance*.

— *Distance entre mots-clés :* deux mots-clés doivent avoir une signification d'autant plus proche que leur distance est faible, deux synonymes devant être à une distance nulle. C'est sur cette propriété qu'étaient fondées les classifications automatiques des mots-clés décrites dans le paragraphe II.3.2.

— *Distance entre une question et un document :* étant donné une question définie par un ensemble quelconque de mots-clés, on désire définir une distance entre cette question et chacun des documents du fichier, la distance étant d'autant plus faible que le document correspond mieux à la question. C'est sur cette propriété qu'est fondée la recherche par voisinage mise en œuvre dans cette étude.

— *Distance entre documents :* deux documents doivent être d'autant plus proches que leur contenu est voisin. C'est sur cette propriété qu'est fondée la typologie des documents réalisée afin de répartir l'ensemble documentaire sur différents fichiers homogènes (cf. § III. 5).

Il s'agissait donc de définir une (ou plusieurs) manière de calculer les différentes distances vérifiant le mieux possible les trois propriétés énoncées ci-dessus et de mettre en œuvre une méthode de typologie appropriée.

3. Définition et notations

Soit $M = \{ m_1, ..., m_i, ..., m_n \}$ l'ensemble des mots-clés utilisés.

Soit $\mathcal{C} = \{ C_1, ..., C_j ..., C_{14} \}$ les quatorze critères retenus dans la phase initiale.

Pour le mot-clé m_i le critère C_j prend la valeur 1 si C_j s'applique à m_i et 0 sinon.

Soit $\underline{m_i} = \begin{bmatrix} 0 \\ 1 \\ . \\ . \\ 1 \\ 0 \\ . \\ 0 \end{bmatrix} \in \mathbb{R}^{14}$ le vecteur des quatorze valeurs associées à m_i par les critères.

Un document D est repéré par un ensemble $M(D) = \{ m'_1, ..., m'_D \}$ de mots-clés. A D on associe donc la partie $M(D)$ de M, $M(D) \subset M$.

De même, à une question Q on associe $M(Q) \subset M$.

4. Choix des distances

Dans cette étude nous avons utilisé la distance euclidienne classique.

Si $\underline{x}_i$ et $\underline{x}'_i$ sont deux vecteurs de $|\mathbf{R}^{14}$, tels que si :

$$\underline{x}_i = \begin{bmatrix} x_{i1} \\ \cdot \\ \cdot \\ x_{ij} \\ \cdot \\ \cdot \\ x_{i14} \end{bmatrix} \quad , \quad \underline{x}_{i'} = \begin{bmatrix} x_{i'1} \\ \cdot \\ \cdot \\ x_{i'j} \\ \cdot \\ \cdot \\ x_{i'14} \end{bmatrix},$$

on a :

$$d^2(x_i\ x_{i'}) = \sum_{j=1}^{14} (x_{ij} - x_{i'j})^2$$

Si $\underline{x}_i$ et $\underline{x}_{i'}$ sont deux éléments de $|\mathbf{R}^{14}$ associés à des mots-clés, la distance est nulle lorsque les mots-clés sont codés de manière identique, et maximum ($= 14$) lorsque les mots-clés sont codés de manière différente sur tous les critères.

Pour définir la distance entre question et document trois procédures ont été utilisées : distance entre centres de gravité, distance moyenne et distance moyenne minimum. On définit le centre de gravité d'un document D par :

$$\underline{g}_D = \frac{1}{\text{Card } (M(D))} \sum_{m_i \in D} \underline{m}_i, \underline{g}_D \in |\mathbf{R}^{14}$$

et la distance entre $\underline{g}_D$, centre de gravité du document, et $\underline{g}_Q$ centre de gravité de la question est donnée par $a^2(\underline{g}_D, \underline{g}_Q)$. Cette distance sera appelée par la suite distance euclidienne.

La distance moyenne entre D et Q est définie par :

$$\frac{1}{\text{Card } (M(Q))\ \text{Card } (M(D))} \sum_{\substack{m_i \in D \\ m_j \in Q}} d^2(\underline{m}_i, \underline{m}_j)$$

La distance moyenne minimum est définie par :

$$\frac{1}{\text{Card } (M(Q))} \sum_{m_i \in Q} \min_{m_j \in D} \{ d^2(\underline{m}_i, \underline{m}_j) \}$$

Lorsque $M(Q) \subset M(D)$ la distance moyenne minimum est nulle, ce qui n'est pas forcément le cas des deux autres distances.

Pour définir les distances entre documents (problème de typologie), seule la distance entre centres de gravité a été utilisée. Il aurait été intéressant de pousser cette recherche et ces expérimentations dans plusieurs directions.

On aurait pu munir $\mathbb{R}^{14}$ de différentes métriques : métrique du χ^2 proposée par J. P. Benzecri [2], distance généralisée (D^2 de Mahalanobis) sur des données binaires [1] et [10], métriques locales (une métrique propre pour chaque groupe de documents) etc... On aurait également pu introduire un indice de distance basé sur la structure booléenne (et non métrique) des données recueillies. Malgré l'intérêt théorique que cela représentait, nous n'avons pas pu, faute de temps, pousser plus loin nos essais.

5. Recherche par voisinage et typologie des documents

Étant donné une question Q définie par un ensemble de mots-clés $M(Q)$, on calcule la distance entre Q et chacun des documents, on classe ensuite les documents dans l'ordre de leur proximité Q et on les liste dans cet ordre.

Ceci permet d'effectuer une recherche par voisinage en deux phases :

— recherche du fichier le plus proche de la question; pour ceci on calcule la distance entre le centre de gravité de la question et les différents centres de gravité des fichiers (stockés une fois pour toutes et recalculés uniquement lorsqu'on ajoute de nouveaux documents),

— à l'intérieur du fichier le plus proche, recherche normale par voisinage.

Cette procédure a l'avantage d'accélérer le temps d'accès et d'éviter la lecture systématique du fichier. Toutefois, lorsqu'une question est à la frontière de plusieurs fichiers (e.g. trois) et que l'on choisit de n'explorer que les deux fichiers les plus proches, il peut arriver que certains documents à distance faible de la question mais contenus dans le troisième fichier soient oubliés. On ne peut pas évaluer actuellement le taux d'oubli de documents pertinents dû à cette procédure en deux phases.

Pour effectuer la typologie des documents on a utilisé la méthode des nuées dynamiques de E. Diday [8 et 9], programme DYC [12]. On cherche à constituer une partition des documents de telle sorte que chaque document soit le plus proche possible des documents appartenant au même groupe et le plus différent possible des documents extérieurs. Pour plus de détails sur l'algorithme et le programme DYC on peut se reporter aux références.

Il est intéressant de signaler que G. Salton [14] D. Mc Clure Murray [11] et Cheng-Kwei Chou [7], parallèlement à cette recherche, ont également étudié les problèmes de typologie de documents et de recherches en deux phases. Pour un point sur tous les travaux effectués dans ces domaines on pourra se reporter à Wolff-Terroine et al. [16].

6. Le traitement informatique

On n'a pas cherché à mettre en œuvre un système opérationnel mais plutôt un « modèle » à partir duquel on désirait tester en réduction un certain nombre d'hypothèses. Toutefois les programmes ont été écrits de manière à pouvoir faire varier facilement certains paramètres essentiels :

— nombre et codification des critères,

— nombre de fichiers,

— choix des distances,

— etc...

Le système est articulé autour de trois fichiers principaux :

— *fichier « mots-clés/critères »* : il contient tous les mots-clés utilisés dans les documents et, pour chaque mot-clé, sa codification sur les critères. Ce fichier constitue une sorte de table mise en mémoire, véritable cœur du système de recherche par voisinage; c'est pourquoi il a été compacté au maximum (deux mots-machine par mot-clé : un pour l'alphanumérique, un pour le vecteur de critères),

— *fichier « documents/mots-clés »* : il s'agit en fait d'un fichier décomposé en plusieurs sous fichiers, chaque sous-fichier pouvant être appelé séparément. Un sous-fichier correspond à un groupe homogène de documents. Chaque document est repéré par son numéro, son titre, son centre de gravité et ses mots-clés,

— *fichier « centres de gravité des groupes »* : il s'agit du centre de gravité de chacun des sous-fichiers du fichier précédent.

L'organigramme général du système, qui tourne sur ordinateur CDC 6600, est donné en figure 7. Les programmes BLØC 1, DYC, PDG et MICRØ sont uniquement utilisés pour effectuer la typologie des documents, définir les documents qui doivent être dans le même sous-fichier et calculer les centres de gravité de ces sous-fichiers (constitution du fichier « centre de gravité des groupes »).

Le programme MØCRI crée ou met à jour une table contenant, pour chaque mot-clé, son indexation sur les critères.

Le programme BLØC 2 crée ou met à jour les sous-fichiers « documents/mots-clés ». On peut ajouter de nouveaux documents sans refaire la typologie à l'aide de BLØC 2. En toute rigueur il faudrait alors recalculer les centres de gravité des sous-fichiers. On a préféré ne pas modifier ces centres de gravité mais plutôt refaire la typologie (séquence BLØC 1 à MICRØ) lorsqu'il y a trop de mises à jour.

Le programme BLØC 3 contient les algorithmes de recherche par voisinage :

— Lecture d'une question,

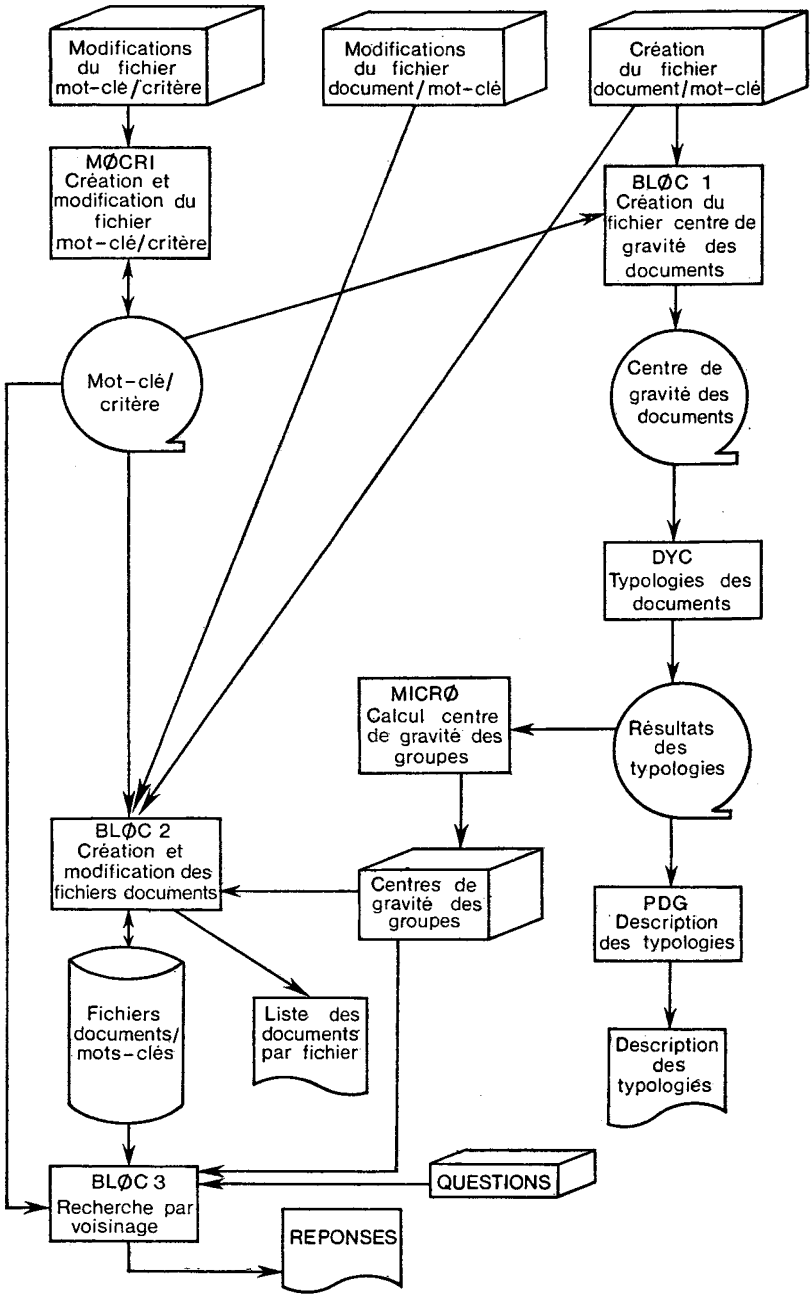

Figure 7

Organigramme de traitement

— Calcul du centre de gravité de la question à l'aide du fichier « mots-clés/ critères »,

— Recherche du sous-fichier le plus proche à l'aide du fichier « centre de gravité du groupe »,

— Calcul de la distance entre la question et les documents du sous-fichier le plus proche (trois possibilités de distance),

— Tri des documents du sous-fichier dans l'ordre de leur proximité à la question,

— Éventuellement exploration des sous-fichiers suivants,

— Impression de la réponse.

Certaines parties de ce système seraient à revoir si l'on avait une grande masse de documents, les programmes de typologie tels que DYC ne permettant de traiter qu'un nombre relativement réduit d'unités à classer (10 000 au plus). Ceci n'est pas réellement un problème car la typologie des documents pourrait être réalisée sur une partie du fichier global, les autres documents étant ensuite réaffectés par BLØC 2. L'analyse des résultats obtenus va maintenant permettre de montrer les avantages et les dangers d'un tel système.

IV. INTERPRETATION DES RESULTATS

1. Typologie des documents

L'ensemble du fonds documentaire d'essai a été éclaté en cinq sous-fichiers dont l'analyse permet de déceler entre leurs éléments de fortes affinités logiques.

Les figures 8 et 9 donnent à titre d'exemple le début des sous-fichiers 1 et 2 dans lesquels s'opposent plus ou moins les documents de type informatique ou marketing.

L'exemple montre qu'il faut tenir compte des recommandations énoncées en § III.5 : extraire un seul fichier (le plus proche de la question) est dangereux, comme en témoigne la présence des 3e et 4e documents (nº 1693 et 1775) dans le fichier 1, documents qui ne sont pas consacrés à l'informatique mais dont le profil est voisin du profil moyen du groupe.

En tout état de cause, à la lumière des résultats obtenus, qu'on pourra analyser plus en détail dans le rapport final de l'étude [5], on peut penser qu'il y a là une méthode permettant d'éviter le balayage complet du fichier documentaire; d'autant que la cohérence du découpage du fichier documentaire en sous-fichiers peut être optimisée de plusieurs façons :

— en choisissant pour réaliser la typologie des documents une distance autre que celle des centres de gravité qui s'est révélée la moins performante au stade de l'analyse des réponses aux questions (voir en III.4 et plus loin en IV.2.1),

Un document est repéré par son numéro d'identification, son titre et les mots-clés (abrégés) qui l'indexent.

Les noms de marque ont été volontairement supprimés sans que cela puisse nuire à la compréhension du sujet, les mots-clés imprimés donnant suffisamment d'indications.

```
FICHIER NO 1
************

0000
      ,LANGAGE DE DEPOUILLEMENT ET D ANALYSE DE FICHIERS,COURS DE
PROGRAMMATION
ENSEIG INF GEST FICH. LANGAGE

1739
ETUDE DU MARCHE FRANCAIS DE LA LECTURE MAGNETIQUE

LECT MAGN. BANQUE     PERIPH ORD

1693
ETUDE PSYCHOSOCIOLOGIQUE ET STATISTIQUE DU MARCHE DE LA RADIO,ANNEXES

PSYCHO SOC HAB ECOUTE TELEVISION RADIO      LOISIRS

1775
METHODE DE PREVISION A COURT TERME,METHODOLOGIE

PLAN C.T.  PREV VTE   STATISTIQ  MODELE     SAISONNAL.

0196
GENERATEUR D ORGANIGRAMMES A PARTIR DE PROGRAMMES

ORGANIGRAM C.A.O.     MISES PAGE

0197
GENATEUR DE FICHIERS D ESSAIS

GEST FICH. ENTR. DONE SYNTAXE    DESSIN CAR

0198
ETUDE DE LA GESTION ON-LINE

GEST AUTOM TEMPS REEL TIME SHAR. REMOTE BAT BANQ DONEE GEST FICH. MOD PREVIS
PREVISION

0199
LES MESSAGES D ENTREE DU      ,FASCICULE 1

LANG PROGR ENTR. DONE SYNTAXE    TRAFIC AER FRET
```

Figure 8

Sous-fichier à dominante automatique/intellectuel

— en recherchant le nombre optimum de mots-clés à utiliser pour « traduire » un document ou une question et en affinant le jeu de critères,

— en essayant de déterminer le nombre optimum de sous-fichiers à extraire pour répondre à une question ce qui permet de retrouver des documents mal classés au niveau de l'opération d'éclatement (voir en III.5).

```
FICHIER NO 2
**************

1811
ETUDE STATISTIQUE DU MARCHE FRANCAIS DU BARDAGE INDUSTRIEL

MAT. CONST FACADES    PROSPECTIV IMAGE MARQ USINES

1883
L INDUSTRIE AUTOMOBILE EN FRANCE PERSPECTIVES 1975 ET 1980

PROSPECTIV AUTOMOBILE PARC       AUTOBUS    CAMIONS

1572
ETUDE DU MARCHE DES DUMPERS,RAPPORT N 20,LE MARCHE FRANCAIS DES DUMPERS

BULLDOZER  IMAGE MARQ DISTRIBUT  PRIX       PARC       MATER. TP  TECHNOLOGI

1805
ETUDE DU MARCHE DE LA CONSTRUCTION DANS LA REGION PARISIENNE

FACADES    MAT. CONST CONSOMMAT. PROSPECTIV HOPITAUX   ECOLES     USINES

1533
ETUDE SUR LES MOISSONNEUSES-BATTEUSES ET LES RAMASSEUSES-PRESSES

MATER AGR. IMAGE MARQ COMP.ACHAT TECHNOLOGI

1572
ETUDE DU MARCHE DES DUMPERS,RAPPORT N 20,LE MARCHE FRANCAIS DES DUMPERS

IMAGE MARQ BULLDOZER  PARC       DISTRIBUT  MATER. TP  TECHNOLOGI

0193
SYSTEME DE GESTION DU PARC D ENGINS MOTEURS DE LA REGION MEDITERRANEE,
PROGRAMMATION DE LA METHODE GENERALE D AFFECTATION,RAPPORT FINAL

MODELE     AFECTATION GRAPHES    LOCOMOTIVE

0179
DEFINITION D UN SYSTEME AUTOMATIQUE D ACQUISITION ET DE TRAITEMENT DES
MESURES POUR LE CENTRE DE RECHERCHES

LABORATOIR DEPOUIL EX CONT NUMER
```

Figure 9

Sous-fichier à dominante offre/production

On peut aboutir par cette typologie à une méthode de préclassement des documents à l'intérieur d'un fichier documentaire, permettant une économie de temps au niveau de la recherche. A la limite, elle peut être utilisée pour constituer des sous-fichiers dans le cadre d'une recherche documentaire, classique. Mais elle trouve sa principale justification dans une utilisation conjointe des techniques de recherche documentaire par voisinage dont nous allons décrire les résultats.

2. Essai de recherche documentaire par voisinage

L'expérience a consisté à interroger le fonds de 100 documents préalablement éclatés en cinq sous-fichiers. On a préparé un certain nombre de questions, représentées par une suite de mots-clés, que l'on a comparés aux documents. Les résultats commentés en IV.2.2, ont été obtenus en utilisant globalement les indexations de six personnes.

Les questions ont été conçues de telle sorte que l'on puisse faire apparaître les avantages et les inconvénients de la méthode :

— compte tenu du faible nombre de documents on connaissait à l'avance les documents qui auraient été jugés pertinents par une recherche classique. Ce qui nous intéressait, c'était le classement des réponses par ordre de distance croissante. On pouvait de la sorte espérer la mise en évidence, à la suite des documents pertinents à 100 %, de documents voisins de la question qui auraient été négligés au cours d'une recherche classique,

— certaines questions ont été posées à partir de termes voisins traduits en critères. On attendait des réponses qu'elles fassent apparaître le peu d'importance de la dispersion des termes utilisés, les différences de vocabulaire devant être atténuées par le voisinage des vecteurs de critères.

— il était intéressant de rechercher quelle pouvait être l'influence sur la réponse du nombre de mots-clés dans la question et dans le document, ou l'effet de juxtaposition à l'intérieur d'une même question de notions différentes.

Nos questions étaient donc des « questions-piège » et nous en avons tiré un certain nombre d'enseignements.

2.1. Influence du choix de la distance

Vingt-deux questions ont été traitées en utilisant trois types de métriques :

— distance moyenne minimum (nous présentons en IV.2.2 l'analyse de deux questions à titre d'exemple),

— distance euclidienne, qui est en fait la distance entre centres de gravité telle que définie en III.4,

— distance moyenne.

D'emblée la *distance euclidienne* s'est révélée peu performante. Cela tient à la dispersion des notions à l'intérieur d'une indexation. Deux nuages de points peuvent avoir des centres de gravités voisins alors que tous les points de l'un sont très éloignés de ceux de l'autre.

Un exemple significatif est donné ci-dessous : un nuage de points constitué de deux pôles opposés (informatique et téléphone) peut avoir un centre de gravité sans rapport avec eux (camions).

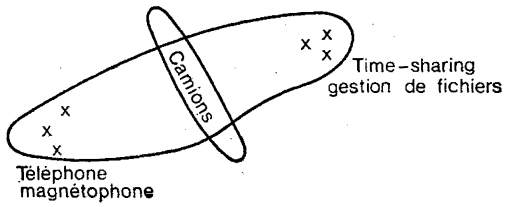

Figure 9 bis

Ce type de distance devra donc être réservé aux recherches portant sur une seule idée représentée par un ou plusieurs mots-clés de la même famille. Elle n'est donc adaptée qu'à un type particulier d'indexage et doit en conséquence être bannie en règle générale. Quant à savoir si elle est adaptée au niveau de la partition du fichier documentaire en sous-fichiers, il faudrait avoir testé toutes les autres distances pour pouvoir tirer des conclusions.

Il semble cependant qu'elle donne de bons résultats lorsque le fichier initial comporte des documents de deux ou trois types bien caractérisés, ce qui n'est que rarement le cas; au demeurant, dans une telle éventualité, la recherche par voisinage n'aurait aucun avantage méthodologique par rapport aux recherches classiques.

La distance moyenne est fortement influencée par le nombre de mots-clés utilisés pour l'indexation. A la limite, si tous les mots de la question existent dans le document, mais juxtaposés à d'autres n'ayant aucun rapport, la distance moyenne apparaîtra forte alors que le document aurait été pertinent.

EXEMPLE :

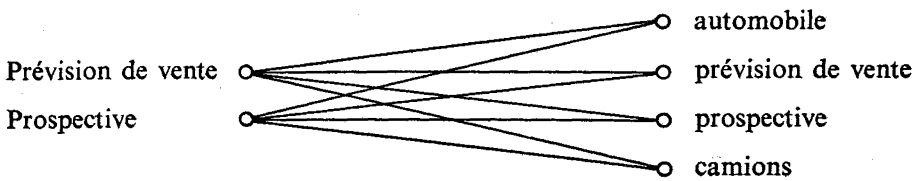

Figure 9 ter

On comprend grâce à cet exemple comment la présence de deux mots-clés tout à fait différents fausse le calcul.

Cette métrique serait donc utilisable dans le cas où le nombre de mots-clés de la question serait voisin de celui des mots-clés du document. Mais, même dans ce cas, la distance moyenne a un inconvénient : elle donne des valeurs très fortes à la distance, atténuant par là-même les différences remarquées dans le voisinage entre documents.

Question	*Documents les plus proches caractérisés par leurs mots-clés*	
Gestion de fichiers *Banque de données* *Accès direct*	D_1 o Gestion automatique o Temps réel o Time sharing o Remote Batch o *Banque de données* o *Gestion fichiers* o Modèle de prévision	2,91
	D_2 o *Gestion de fichiers* o Chaînage o *Accès direct*	3,00
	D_3 o Edition o Composition automatique o Mise en page	3,166

Le document D_2, qui répond à la question aussi bien que le document D_1, est à 0,09 de celui-ci alors que le document D_3 qui n'est pas pertinent est seulement à 0,25.

Ainsi, qu'il s'agisse de distance des centres de gravité (distance euclidienne) ou de distance moyenne, la juxtaposition de plusieurs idées et la quantité des mots-clés ont des répercussions défavorables à la recherche par voisinage.

Nous avons tout de même analysé ces métriques, car il est possible d'envisager une certaine amélioration si l'on peut utiliser des opérateurs ET, OU, SAUF, comme cela se fait pour les recherches classiques, avec pour objectif d'éviter l'influence de la juxtaposition des idées et du nombre des mots sur le calcul de la moyenne.

La *distance moyenne minimum* n'a pas ces inconvénients : elle indique la moyenne des distances de chaque mot de la question au *seul* mot qui lui est le plus proche dans le document.

Ainsi, elle est censée représenter les voisinages entre des concepts strictement comparables.

Un risque cependant : lorsque le nombre de mots est très différent dans la question et dans le document, on multiplie les chances de trouver des distances faibles; et ceci d'autant plus que le jeu de critères a moins de composantes, le nombre de combinaisons possibles dans la construction du vecteur étant moins important.

A la limite, si l'on pose une question représentée par un seul mot indexé selon un jeu de critères des plus réduits, on a de fortes chances de trouver en réponse des documents d'autant plus « voisins » (c'est de voisinages aberrants qu'il s'agit ici) que le nombre de concepts indexés sera plus grand. On classera les documents beaucoup plus en fonction d'un nombre décroissant de mots les indexant, qu'en fonction d'un voisinage réel.

On pourra donc utiliser la distance moyenne minimum à condition d'utiliser dans l'indexage des questions et des documents un nombre important de concepts, ou tout au moins une quantité comparable de mots. C'est une restriction importante, mais elle n'a d'influence que pour autant que le jeu de critères est réduit.

L'ensemble des commentaires qui précèdent illustre bien les avantages et les dangers de la recherche par voisinage. Si sa souplesse et son évolutivité en font un outil intéressant, elle peut amener bien des déconvenues à qui n'aura pas toujours à l'esprit les tenants et aboutissants de la méthode.

2.2. Analyse des résultats obtenus en utilisant la distance moyenne minimum

Nous avons jusqu'ici surtout souligné les défauts de la recherche par voisinage. L'analyse des réponses à quelques questions posées pour les besoins de l'expérience va nous démontrer que son principal avantage ressort, même avec un fonds documentaire réduit et un jeu de critères à seulement quatorze composantes : il s'agit de la faculté de suggérer, au-delà de la stricte pertinence, des documents intéressant des domaines voisins.

Question n° 1 : un domaine bien précis, agriculture, engrais, exploitation agricole (fig. 10).

Cette question provoque le balayage des sous-fichiers 3 et 5. A distance très faible dans le fichier 3 (0,3333) on trouve en tête du classement les rapports 1815 et 1675. Ils correspondent parfaitement à la question et auraient été extraits de la même manière au cours d'une recherche de type classique.

A distance plus forte (0,6666) on trouve trois documents :

— rapport 1500 : il est effectivement proche des précédents, le vecteur de critères étant très semblable pour herbicide et engrais. C'est le type même de voisinage suggéré par la méthode;

— rapport 1688 : bien que portant le même titre que les documents classés en tête, ce rapport traite d'un sujet plus général, le mot engrais ne figurant dans le titre que pour des raisons de continuité bibliothéconomique.

```
QUESTION NO : 1
**************

AGRICULTURENGRAIS   EXPL.AGRIC

          DISTANCE MOYENNE MINIMUM
          ************************


          FICHIER NO  3
          *************

DISTANCE      .3333333333
1815 1
LE MARCHE DES ENGRAIS EN FRANCE DEFINITION D UNE POLITIQUE COMMERCIALE,
L ATTITUDE DES AGRICULTEURS,RESULTATS DE L ETUDE EN EXTENSION

ENQ. D OP.ENGRAIS   COMP.ACHATAGRICULTURCRIT CHOIXPRIX


DISTANCE      .3333333333
1675 1
LE MARCHE DES ENGRAIS EN FRANCE DEFINITION D UNE POLITIQUE COMMERCIALE
RAPPORT N 1 L ATTITUDE DES AGRICULTEURS HYPOTHESES DE TRAVAIL

ENGRAIS   CONSOMMAT.CRIT CHOIXCOMP.ACHATPRIX      S.A.V.    FOURNISEUR


DISTANCE      .6666666667
1500 1
DETERMINATION DU PRIX DE VENTE OPTIMUM D'UN HERBICIDE

PRIX      HERBICIDESCOURB DEMABENEFICE  ASSURANCE AGRICULTUR


DISTANCE      .6666666667
1680 1
PROMOTION DE LA MARQUE       AUPRES DES TOURISTES EUROPEENS,ETUDE
PSYCHOSOCIOLOGIQUE

PROMO VTESIMAGE MARQPROD PETR.STAT SERV.TOURISME  HOTELLERIE


DISTANCE      .6666666667
1688 1
LE MARCHE DES ENGRAIS EN FRANCE DEFINITION D UNE POLITIQUE COMMERCIALE
RAPPORT N 2 PERSPECTIVES D EVOLUTION A MOYEN TERME DES STRUCTURES AGRICOLES
PICARDIE ET CENTRE
AGRICULTURPOPULATIONEXPL.AGRICRENDEMENT


DISTANCE      .6666666667
1824 1
ETUDE ECONOMETRIQUE ET PREVISION DE LA CONSOMMATION DE CIMENT ET DE PLATRE
DE CONSTRUCTION

PROSPECTIVCONSOMMAT.CIMENT    PLATRE    MAT. CONSTPLAN C.T. MOD ECONO


DISTANCE     1.0000000000
1699 1
ETUDE PSYCHOSOCIOLOGIQUE DE LA CONSOMMATION DE BISCOTTES,ANALYSE DES
RESULTATS

COMP.ACHATCONSOMMAT.BISCOTTES PAIN      CRIT CHOIXPREV VTE
```

Figure 10

Il est normal que ce document soit considéré comme moins pertinent que les deux premiers;

— rapports 1680, 1824 : la présence ici de ces documents sans rapport avec l'agriculture s'explique par le fait que les concepts « produits pétroliers », « ciment » et « engrais » sont considérés comme synonymes par rapport au jeu de 14 critères. Le tableau du § I.2. explique cette anomalie et montre que ce n'est pas la méthode qui est en cause mais l'étroitesse du jeu de critères. Il en est de même pour le document considéré comme pertinent en tête du classement du deuxième sous-fichier extrait (fichier 5), que nous n'avons pas reproduit.

Ainsi, sommes-nous arrivés, à partir de concepts indexés selon 14 critères et en l'absence de tout langage documentaire, à définir au-delà de la stricte pertinence une parenté (réelle ou induite par la méthode) entre un certain nombre de documents de la même famille. Si l'on a évité des « silences », on a fait tout de même apparaître quelques « bruits ». Après avoir étudié une autre question nous essaierons de proposer des moyens de perfectionner le système.

Question n° 2 : plusieurs idées juxtaposées : engrais, comportement d'achat, critère de choix, prix (fig. 11).

Cette question intéresse les sous-fichiers 3 et 4, ce qui constitue déjà une première différence par rapport à la question n° 1. Dans le fichier 3, on retrouve à distance nulle les deux documents exactement pertinents. Au-delà, la distance est trop importante (1,25) pour que l'on puisse considérer un voisinage comme intéressant. A noter, par rapport à la question n° 1 que les documents 1500 et 1 688 ont disparu de la tête du classement.

Le sous-fichier 4 propose des documents voisins seulement par rapport à une partie de la question (critères de choix de produits autres que des engrais).

2.3. Commentaires

Ces exemples d'interrogation, selon les principes de la recherche par voisinage, d'un fichier documentaire réduit illustrent bien les avantages et les limites actuelles de la méthode.

L'édifice repose bien entendu sur le choix des critères à utiliser pour l'indexation. Mais un certain nombre d'autres paramètres doivent être pris en considération :

— nombre optimum de critères,

— nombre optimum de mots-clés (supports de concepts),

— type de métrique à utiliser.

L'expérience limitée de recherche documentaire à laquelle nous nous sommes livrés a apporté quelques éclaircissements sur la façon d'optimiser la procédure. Elle ne prétend pas avoir résolu tous les problèmes. Il est notamment évident que les avantages apportés par une telle méthode risquent, aux

```
QUESTION NO  2
*************

ENGRAIS   COMP.ACHATCRIT CHOIXPRIX

          DISTANCE MOYENNE MINIMUM
          ************************


          FICHIER NO  3
          *************

DISTANCE    0.0000000000
1815 1
LE MARCHE DES ENGRAIS EN FRANCE DEFINITION D UNE POLITIQUE COMMERCIALE,
L ATTITUDE DES AGRICULTEURS,RESULTATS DE L ETUDE EN EXTENSION

ENQ. D OP.ENGRAIS   COMP.ACHATAGRICULTURCRIT CHOIXPRIX


DISTANCE    0.0000000000
1675 1
LE MARCHE DES ENGRAIS EN FRANCE DEFINITION D UNE POLITIQUE COMMERCIALE
RAPPORT N 1 L ATTITUDE DES AGRICULTEURS HYPOTHESES DE TRAVAIL

ENGRAIS   CONSOMMAT.CRIT CHOIXCOMP.ACHATPRIX      S.A.V.    FOURNISEUR


DISTANCE    1.2500000000
1588 1
RECHERCHE D UNE POLITIQUE COMMERCIALE POUR LES CHAMPAGNES      L ENQUETE
AUPRES DES ACHETEURS DE CHAMPAGNE

BOISSONS  CHAMPAGNE IMAGE MARQCONSOMMAT.CRIT CHOIXPRIX      ELAST DEMA
PROD. LUXE


DISTANCE    1.5000000000
1743 1
ETUDE DE L EFFICACITE DE LA CAMPAGNE PUBLICITAIRE 1965-1966

TEST ANN. MEDIA     IMAGE MARQBONNETTERIPUBLICITE COMP.ACHAT


DISTANCE    1.7500000000
1699 1
ETUDE PSYCHOSOCIOLOGIQUE DE LA CONSOMMATION DE BISCOTTES,ANALYSE DES
RESULTATS

COMP.ACHATCONSOMMAT.BISCOTTES PAIN      CRIT CHOIXPREV VIE


DISTANCE    1.7500000000
1648 2
LE MARCHE DU PAIN ET SES PERSPECTIVES,ETUDE GRAND PUBLIC-TABLEAUX
STATISTIQUES

CONSOMMAT.PAIN      COMP.ACHATCRIT CHOIX
```

Figure 11

```
                FICHIER NO  4
                ************

DISTANCE      .2500000000
1866 1
ORIENTATION OF THE MARKETING POLICY IN THE FIELD OF SPRAYING EQUIPMENT
GENERAL CONCLUSIONS AND RECOMMENDATIONS FOR FRANCE

AIR COMPR.APAREILLAGMAT. CONSTPEINTURES TECHNOLOGICOMP.ACHATIMAGE MARQ
DISTRIBUT PRIX       S.A.V.     PROMO VTESCRIT CHOIX


DISTANCE      .2500000000
1761 1
ETUDE PSYCHOSOCIOLOGIQUE DES ATTITUDES DES PEINTRES EN BATIMENT A L EGARD
DES REVETEMENTS PLASTIQUES MURAUX

PSYCHO SOCMAT. CONSTREV. PLASTIMAGE MARQPEINT.BAT CRIT CHOIXCOMP.ACHAT
PRIX       PROMO VTESDISTRIBUT


DISTANCE      .5000000000
1802 1
LE MARCHE DU BEURRE ET DU FROMAGE,ANALYSE DES RESULTATS,SYNTHESE ET
CONCLUSIONS

BEURRE     FROMAGE    CONSOMMAT.TYPOLOGIE CRIT. CHOIXPRIX       COMP.ACHAT
MOTIVATIONIMAGE MARQGASTRONOMI


DISTANCE     1.0000000000
1567 2
LES ATTITUDES DU PUBLIC A L EGARD DE LA MONTRE ET DE SON RENOUVELLEMENT

ETUDE PSYCHOSOCIOLOGIQUE,ANNEXES

IMAGE MARQCOMP.ACHATHORLOGERIEVTE DETAILCRIT CHOIXBIJOUTERIE
```

Figure 11 *(suite)*

yeux de certains, d'être obérés par les nombreuses anomalies (bruits introduits par des voisinages aberrants) qui ont été constatées à l'analyse.

Mais il ne faut pas oublier que la principale caractéristique de la recherche par voisinage est la modularité et l'évolutivité de l'outil.

En particulier, il serait possible, à la lumière des aberrations observées, de revoir l'indexation des mots-clés sur les critères, éventuellement d'introduire de nouveaux critères ou d'en supprimer. La procédure est donc perfectible, nous allons en tracer ci-après quelques perspectives d'avenir.

V. CRITIQUES DE LA METHODE ET PERSPECTIVES DE RECHERCHE

Jusqu'à présent, un « prototype » a été réalisé. De nombreuses améliorations doivent être apportées à ce modèle avant qu'il soit opérationnel. Le Professeur J. P. Benzecri a bien voulu faire quelques remarques sur ce travail et a proposé de nouvelles voies de recherche. Tenant compte de ses remarques, on expose dans les paragraphes suivants quelques critiques et quelques nouvelles directions de travail qu'il serait intéressant de suivre maintenant.

1. Sélection des critères de référence

La méthode est basée sur le choix d'un jeu de critères sur lesquels on indexe les mots-clés. Plus il y a de critères, plus la recherche est précise et plus l'indexation est lourde. On est donc amené à chercher un compromis entre la précision et la rapidité d'indexation.

D'autre part, il n'est pas sûr qu'un seul jeu de critères permette d'indexer tous les mots-clés d'un domaine : dans une application grandeur nature il serait sans doute nécessaire de choisir un jeu de critères par domaine. Ces deux points devraient être approfondis dans des recherches ultérieures.

2. Calcul des distances

C'est certainement sur ce point que l'on pourrait améliorer le plus ce modèle. On a déjà vu (paragraphe III.4) que l'on pouvait envisager plusieurs modes de calcul. Dans une optique légèrement différente, J. P. Benzecri pense qu'« il eût été préférable de placer les mots-clés sur les axes issus d'une analyse Mots × Critère. Sur ces axes on place immédiatement toute question q et tout document d (comme barycentre d'un système de mots) il est ensuite facile d'explorer le voisinage d'un point q pour y chercher des d ».

Ce type de recherche oblige à conserver pour chaque mot-clé ses coordonnées sur les différents facteurs et donc utilise une plus grande place en mémoire. Il doit, par contre, éliminer certaines anomalies dues au mauvais choix des critères ou à une mauvaise codification des mots-clés. Une étude assez systématique tendant à montrer les avantages et les inconvénients de ces différentes formules devrait être entreprise.

3. Organisation des fichiers

Il s'agit de trouver un compromis entre la taille des fichiers (et donc leur homogénéité) et leur nombre. J. P. Benzecri suggère « d'appliquer au moyen des documents considérés dans l'espace des 5 ou 6 premiers facteurs (ceci est essentiel pour accélérer le calcul) un algorithme de classification tel que

celui de E. Diday [9] pour définir 30 à 100 îlots de documents. Eventuellement, il serait bon d'organiser ensemble des îlots par une classification ascendante hiérarchique ».

Il est évident que la solution que nous avons adoptée n'est pas vraiment satisfaisante : nos fichiers sont trop hétérogènes et contiennent des documents trop différents. Une solution consisterait peut-être à organiser les documents en fichiers, puis les fichiers en îlots, et éventuellement, de disposer d'une table de distance entre les îlots.

4. Réalisation d'un système grandeur nature

Il serait intéressant de tester nos idées sur un système grandeur nature. Ceci permettrait en effet d'évaluer le coût et les performances de notre modèle, et en particulier de tester la pertinence et l'exhaustivité des réponses.

CONCLUSIONS

Le « modèle » de recherche documentaire par voisinage que nous avons mis au point propose une alternative à la construction a priori d'outils documentaires sophistiqués.

Nous avons successivement défini la notion de distance et mis au point une méthode de recherche documentaire faisant appel à cette notion.

Elle permet de moduler les réponses aux questions de telle sorte qu'au-delà des documents pertinents le système soit capable de suggérer des documents appartenant à des domaines voisins.

C'est en quelque sorte un rôle actif qui est attribué au demandeur invité à faire le tri entre les voisinages licites et les aberrations.

Pour arriver à cela, aucun investissement de départ n'est nécessaire : l'outil se construit à mesure qu'apparaissent de nouveaux documents. L'indexation reste valable en permanence puisqu'elle peut subir sans problème des mises à jour successives.

Un certain nombre de précautions doivent cependant être prises pour atténuer les effets de la subjectivité inhérente au processus de codification des concepts selon les critères.

Nous avons vu que pour limiter les anomalies, de nombreuses recherches devaient encore être effectuées : choix des critères, des distances, nombre optimum de concepts indexant documents et question, découpage en sous-fichiers, en îlots, etc... Compte tenu des résultats obtenus au stade actuel de la recherche, on peut espérer arriver, par améliorations successives, à disposer de véritables outils documentaires adaptables à des domaines non justiciables d'un thesaurus ou d'une classification.

A ce stade de développement de la méthode, on peut affirmer que ce n'est pas une utopie.

BIBLIOGRAPHIE

[1] BALAKRISHNAN V. et SANGHVI L. D. *Distance between populations on the basis of attribute data*, Biometrics, vol. 24, December 1968, pp. 859-865.

[2] BENZECRI J. P. *Distance distributionnelle et métrique du χ^2 en analyse factorielle des correspondances* (cours polycopié), Laboratoire de Statistique Mathématique, Université Paris VI, 1970.

[3] BORIONNE D. *A propos des classifications hiérarchiques*. Note de Travail n° 124, Metra International, Direction Scientifique, septembre 1970.

[4] BOUROCHE J. M. et CURVALLE B., *La recherche documentaire par voisinage. Principe, vérification des hypothèses de base, premiers résultats* (rapport intermédiaire). Rapport de Recherche n° 56, Metra International, Direction Scientifique, juillet 1971.

[5] BOUROCHE J. M. et CURVALLE B., *La recherche documentaire par voisinage. Constitution des fichiers, traitement des questions* (rapport final), Rapport de Recherche n° 66, Metra International, Direction Scientifique, septembre 1972.

[6] BOUROCHE J. M., CURVALLE B. et DONIO J., *La recherche documentaire par voisinage*, Cahiers de l'IRIA, n° 7, novembre 1971, pp. 187-227.

[7] CHENG-KWEI CHOU, *Algorithmes for hash coding and document classification*, Ph. D. Thesis, Illinois University, Urbana, January 1972.

[8] DIDAY E., *Optimisation en classification automatique et reconnaissance des formes*, Note Scientifique n° 6, supplément au bulletin de l'IRIA n° 12, mai/juin 1972.

[9] DIDAY E., *An introduction to the dynamic clusters method*, METRA, vol. XI, n° 3, septembre 1972, 505-519.

[10] KURCZYNSKI T. W., *Generalised distance and discrete variables*, Biometrics, vol. 26, september 1970, pp. 525-534.

[11] McCLURE MURRAYD., *Document retrieval based on clustered files*, Report ISR 20, Cornell University, Department of Computer Science, June 1972.

[12] PLUCHET M., *Quelques remarques sur la méthode de classification des nuées dynamiques (programme DYC)*, Note de travail n° 138, Metra International, Direction Scientifique, août 1971.

[13] SALTON G., *Automatic information organization and retrieval*, Mc Graw Hill, 1968.

[14] SALTON G., *Dynamic document processing*, Communications of the A.C.M., vol. 15, n° 7, July 1972, pp. 658-667.

[15] VAN DIJK M. et SZANTO G., *La documentation économique dans l'Administration des Affaires*, Bureau Marcel Van Dijk, Bruxelles et INSEAD, Fontainebleau 1967.

[16] WOLFF-TERROINE M., RIMBERT D. et FLEURY D. *La classification automatique et son application aux systèmes documentaires*, Automatisme, Vol. XVII, n° 11, novembre 1972, pp. 347-361.