

MAX REINERT

**Une méthode de classification descendante
hiérarchique : application à l'analyse
lexicale par contexte**

Les cahiers de l'analyse des données, tome 8, n° 2 (1983),
p. 187-198

http://www.numdam.org/item?id=CAD_1983__8_2_187_0

© Les cahiers de l'analyse des données, Dunod, 1983, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

UNE MÉTHODE DE CLASSIFICATION DESCENDANTE HIÉRARCHIQUE : APPLICATION A L'ANALYSE LEXICALE PAR CONTEXTE

[C.D.H. LEXICALE]

par A. Reinert (1)

1 Introduction

Si la procédure proposée se situe au carrefour de plusieurs techniques, comme nous le verrons, elle a été principalement élaborée en relation avec une pratique de l'analyse des données en psycho-sociologie et plus précisément pratique de l'analyse des entretiens ou des réponses libres à des questions ouvertes.

La définition de cet outil dépend donc d'une méthodologie particulière. Nous allons préciser notre problématique de ce point de vue, avant de décrire l'algorithme. Nous donnerons ensuite un exemple d'application.

Lors du dépouillement d'enquêtes nécessitant une analyse de contenu des réponses, il est bien rare d'avoir initialement un ensemble d'indicateurs fiables qui puissent être assimilés à des variables. En effet, aucune hypothèse valide ne peut être associée à un indicateur isolé, du fait qu'on ne connaît pas, *a priori*, le sens qu'il prend dans une sous-population donnée.

Par exemple, si les indicateurs sont des mots, leur sens dépend, pour une part, du contexte. Aussi, nous avons recherché une procédure qui, dans un premier temps, permette à l'utilisateur de regrouper ces indicateurs polysémiques dans des classes caractéristiques de certains contextes. Ces regroupements lui permettront alors de constituer ce que nous avons appelé des "indicateurs de contexte" dont le sens est plus précis et qui, par conséquent, sont susceptibles d'être associés à des hypothèses de contenu mieux définies.

Cette première condensation des données une fois effectuée, il est loisible à cet utilisateur d'étudier les interrelations entre indicateurs de contexte, ou leurs relations avec d'autres caractéristiques de la population.

Pratiquement, nous avons opté pour une classification hiérarchique des individus. Ce classement n'est qu'un moyen et non un but, le résultat utilisable étant la description du profil des classes d'individus à l'aide des indicateurs. Ce résultat est du reste plus stable que le classement lui-même, les individus à la frontière de deux classes pouvant permuter sans que change sensiblement le profil de ces classes.

Cette description des classes peut être effectuée après n'importe quel type de classification. Une telle orientation impose

(1) Ingénieur CNRS, Laboratoire de Psychologie, Université de Toulouse-Le Mirail.

toutefois d'avoir des classes de sujets de cardinal assez élevé, bien différenciées les unes des autres, ce qui explique en partie le choix d'une procédure descendante pour les obtenir.

2. Description de l'algorithme

Les tableaux de données que nous proposons d'analyser à l'aide de cette technique, sont des tableaux à double entrée, individus \times indicateurs de type présence-absence, le nombre total de "présences" étant relativement faible.

Appelons I , l'ensemble des individus ; J , l'ensemble des indicateurs ; $I \times J$ le tableau initial. A l'aide d'un algorithme approprié, on cherche à réordonner les lignes et les colonnes de ce tableau de manière à obtenir la décomposition suivante :

	J	
I_1	$I_1 \times J_1$	$\epsilon_1 \approx 0$
I_2	$\epsilon_2 \approx 0$	$I_2 \times J_2$

$\{I_1, I_2\}$ définissant une partie des individus en deux classes. Le partage des indicateurs aboutit à la définition de deux classes empiétantes J_1, J_2 , les indicateurs de $J_1 - J_1 \cap J_2$ étant nettement associés à I_1 et les indicateurs de $J_2 - J_1 \cap J_2$, à I_2 .

Les sous-tableaux $I_1 \times J_1$ et $J_2 \times J_2$ ainsi définis sont ensuite décomposés selon le même algorithme.

On aboutit de cette manière à la construction d'une hiérarchie de sous-tableaux emboîtés, indiquée pour la commodité des présentations par le χ^2 de ces derniers.

Une mesure donnée de l'indice correspond à une partition sur les sujets. Pour ne pas alourdir cette présentation, nous ne décrivons que l'algorithme de découpage d'un sous-tableau, l'algorithme gérant les différents sous-tableaux ne constituant en définitive qu'un exercice de programmation.

Découpage d'un sous-tableau T

Il s'effectue en deux temps :

- a. / Recherche de la partition des sujets en deux classes maximisant le moment d'ordre 2,
- b. / Recherche des classes empiétantes des variables associées à cette partition.

2.1 Notation

- k_{ij} , la fréquence d'apparition de l'indicateur j dans le protocole i (L'algorithme proposé peut s'appliquer au tableau de fréquences : ce sont des raisons d'ordre méthodologique qui limitent l'étude à des tableaux à valeurs logiques).

- $k = \{k_{ij} | i \in I ; j \in J\}$; $p_{ij} = k_{ij}/k$; $p_i = \{p_{ij} | j \in J\}$;
 $n = \text{cardinal}(I)$

- on considérera principalement l'espace vectoriel R_J muni de la métrique du χ^2 .

- on notera aussi "i" pour simplifier le point de R_J associé à l'individu $i \in I$ de coordonnées $p_j^i = p_{ij}/p_i$.

- enfin $N_I = \{p_j^i, p_i | i \in I\}$ est le nuage des individus dans R_J .

2.2 Recherche de la partition $\{I_1, I_2\}$ de I maximisant le moment

d'ordre 2 : Nous n'avons pas trouvé d'algorithme qui nous permette d'affirmer que la partition obtenue maximise bien le moment d'ordre 2, mais nous avons essayé de nous approcher au mieux de la solution.

Notons d'abord que la valeur propre associée au 1er facteur de l'A.F.C. du tableau considéré, constitue une borne supérieure pour tous les moments d'ordre 2 des partitions en deux classes $\{I_1, I_2\}$ de I : sinon l'axe passant par le centre de gravité des deux classes permettrait d'extraire une part d'inertie plus forte que celle extraite par le 1er axe factoriel (théorème de Huygens appliqué à la projection du nuage sur cet axe) : cette particularité nous fait penser que, dans la plupart des cas, la ligne joignant les centres de gravité des classes de la partition recherchée ne peut être très éloignée de ce premier axe, notamment si la première valeur propre est bien distincte des autres.

(Dans les autres cas, il est probable qu'il n'y a pas de partition optimale : peu importe de prendre l'une ou l'autre, les distinctions non retenues lors de cette étape pouvant apparaître à l'étape suivante).

Aussi, avons-nous procédé de la manière suivante :

a. Recherche du premier axe factoriel dans R_J ,

b. Recherche de l'hyperplan orthogonal à cet axe découpant I en deux classes maximisant le moment d'ordre 2(*) : cet algorithme répondrait au problème si les centres de gravité des classes sont effectivement sur l'axe. Comme ce n'est pas le cas en général,

c. On améliore la partition obtenue en (b) à l'aide d'un algorithme d'échange : pour chaque individu, on observe si sa permutation de classes accroît le moment d'ordre 2 de la partition. De cette manière, on améliore progressivement la partition initiale jusqu'à un maximum local du moment d'ordre 2. Cette manière de procéder est une variante de la méthode des Nuées Dynamiques (DIDAY 1970), les noyaux des classes étant définis par la partition précédente. Toutefois le point i observé est affecté à la classe I_k qui minimise l'inertie intraclasse de $I_k \cup \{i\}$.

Reprenons plus en détail :

On suppose pour simplifier que les éléments de $I = \{i_1, i_2, \dots, i_n\}$

(*) Cet algorithme se rapproche d'une procédure utilisée dans la segmentation (BERTIER & BOUROCHE (1975) - Analyse des données multidimensionnelles. P.U.F., 175-192).

sont ordonnées selon leurs coordonnées sur le premier axe factoriel; on procède en deux temps :

A. Recherche de la partition $P^m = \{\{i_1, \dots, i_m\}, \{i_{m+1}, \dots, i_n\}\}$

$$M^2(P^m) = \|g_1 - g\|^2 \cdot P_1 + \|g_2 - g\|^2 \cdot P_2$$

avec par exemple : $g_1 = (\sum \{p_j^i \cdot p_i | i \in I_1\}) / (\sum \{p_i | i \in I_1\})$.

$M^2(P^m)$ n'est autre que le χ^2/k du tableau condensé

$$\{N_{rj}^m | r = 1, 2 ; j \in J\} \text{ avec } N_{rj}^m = \sum \{k_{ij} | i \in I_r\}$$

Sur le plan informatique, c'est ce tableau qui est conservé en mémoire pour calculer successivement les $n-1$ termes $M^2(P^{m+1})$ en utilisant les relations suivantes :

$$N_{1,j}^{m+1} = N_{1,j}^m + k_{m+1,j} ; N_{2,j}^{m+1} = N_{2,j}^m - k_{m+1,j} .$$

B. "Amélioration" de la partition obtenue, à l'aide d'un algorithme d'échange.

- Définition : "sujets bien classés", "partition bien formée"

Posons $M^2(I_k) = \sum \{\|i - g_k\|^2 \cdot p_i | i \in I_k\}$, g_k étant le centre de gravité de I_k .

Posons $P = \{I_1, I_2\}$ et $N^2(P) = M^2(I_1) + M^2(I_2)$.

Appelons P^s la partition obtenue en substituant l'individu s de classe.

On dira que s est bien classé si $N^2(P^s) \geq N^2(P)$.

On dira que P est bien formée si tous les éléments de P sont bien classés.

Appelons I_1 la classe de P qui contient s , de masse p_s .

Appelons g_1 le centre de gravité de $I_1 - \{s\}$, p_1 la masse de $I_1 - \{s\}$.

Appelons g_2 le centre de gravité de I_2 , p_2 la masse de I_2 .

On a : $M^2(I_2 \cup \{s\}) = M^2(I_2) + M^2(\{s\}) + M^2(\{g_2\} \cup \{s\})$.

avec : $M^2(\{s\}) = 0$ et $M^2(\{g_2\} \cup \{s\}) = \|s - g_2\|^2 \cdot p_2 \cdot p_s / (p_2 + p_s)$

donc : $M^2(I_2 \cup \{s\}) = M^2(I_2) + \|s - g_2\|^2 \cdot p_2 \cdot p_s / (p_2 + p_s)$.

de même : $M^2(I_1) = M^2(I_1 - \{s\}) + \|s - g_1\|^2 \cdot p_1 \cdot p_s / (p_1 + p_s)$

Comme : $N^2(P^s) = M^2(I_2 \cup \{s\}) + M^2(I_1 - \{s\})$,

on a : $N^2(P^s) = N^2(P) + \|s - g_2\|^2 \cdot p_2 \cdot p_s / (p_2 + p_s)$

$$\dots - \|s - g_1\|^2 \cdot p_1 \cdot p_s / (p_1 + p_s)$$

Pour contrôler la bonne affectation d'un point s , il suffit donc de calculer :

$$W(s) = N^2(P^S) - N^2(P) = p_s \cdot [\|s - g_2\|^2 \cdot p_2 / (p_2 + p_s) - \|s - g_1\|^2 \cdot p_1 / (p_1 + p_s)].$$

Sur le plan informatique, on a en mémoire les coordonnées de s , g_1 et g_2 , de même que les masses ; il est donc aisé de calculer $W(s)$.

- Convergence et unicité

Le moment d'ordre 2 des partitions successives croissant strictement, l'algorithme ne peut que converger, puisque ces valeurs sont bornées, mais il est clair que dans la plupart des cas, il n'y a pas unicité d'une telle suite.

Dans le cas où la partition initiale n'a qu'un nombre négligeable d'éléments mal placés, on peut supposer que le centre de gravité des classes est quasiment stable lorsqu'on permute globalement ces éléments : dans ce cas la partition obtenue est bien formée, et elle est unique. Ce cas extrême permet de supposer une propriété d'unicité locale.

C. Procédure effectivement utilisée pour trouver la partition optimale.

On améliore la partition obtenue en (A), pas à pas, en réaffectant à chaque pas, l'ensemble des individus mal placés, d'une classe dans l'autre : la convergence s'avère très rapide et n'excède que très rarement 3 pas.

En pratique, le nombre de sujets mal placés s'avère très faible : la partition initiale est même assez souvent une partition bien formée.

2.3 Recherche du classement des indicateurs : La nécessité de classer les indicateurs n'est pas évidente *a priori*. Elle s'est imposée peu à peu pour des raisons pratiques. Lorsqu'un thème est très associé à une classe, son résidu sur l'autre classe peut avoir un poids très faible, et occasionner, lors de l'analyse du sous-tableau associé, la présence d'un facteur-artéfact (lié à l'existence de quelques éléments de poids faible).

Nous avons choisi de ne pas étudier ces résidus. *A contrario*, un classement trop rigide peut aboutir à une instabilité des résultats, compte-tenu de l'arbitrarité du classement pour certains indicateurs : aussi nous avons opté pour la définition des classes empiriques.

Sur le plan informatique, on a conservé en mémoire le tableau $\{N_{rj} \mid r \in \{1, 2\} ; j \in J\}$ défini au § 2.2 (A).

Le calcul que nous proposons, consiste à comparer, pour chaque indicateur j , la distribution des k_j réponses dans les classes I_1 , et I_2 , à la distribution moyenne des réponses attendues dans ce cas, à l'aide du :

$$\chi^2 = \sum \{ (N_{rj} - k_j \cdot N_r / k)^2 / (k_j \cdot N_r / k) \mid r \in \{1, 2\} \}$$

avec $k_j = \sum \{k_{ij} \mid i \in I\}$ et $N_r = \sum \{N_{rj} \mid j \in J\}$

Pour rendre cette valeur indépendante du nombre k_j des réponses, on a effectué un calcul comparable à celui du coefficient de contingence :

$$C = (\chi^2 / (\chi^2 + k_j))^{1/2}$$

Lorsque ce coefficient, qui varie entre 0 et 1, est supérieur à un seuil donné, l'indicateur j n'est affecté qu'à une seule classe, et sinon au deux.

2.4 Les principaux paramètres :

CTEST : seuil du coefficient C pour le recouvrement des classes. Nous proposons une valeur par défaut égale à .3 (voir § 4).

TSJ : les indicateurs ou leur restriction à un sous-tableau, dont le poids est inférieur ou égal à cette valeur, sont éliminés de l'analyse. Sa valeur par défaut est égale à 3.

NPOS : un sous-tableau n'est plus décomposé si le nombre des individus le définissant est inférieur à cette valeur.

2.5 Le programme de diagonalisation utilisé : La matrice symétrique à diagonaliser est calculée dans la plus petite dimension du sous-tableau dont elle est issue. La méthode de diagonalisation retenue est la méthode par itération d'un vecteur (*). On en a simplifié la procédure puisqu'on ne recherche que le premier vecteur propre. Ce fait rend d'autre part inutile un calcul très précis ($\| \lambda v - A.v \| / \lambda \leq 0.001$). Sous cette forme, cette technique s'avère assez rapide. Nous recherchons toutefois une procédure plus rapide pouvant s'appliquer à des matrices de grande taille.

2.6 Remarque à propos des capacités du programme informatique :

Ce programme est actuellement implanté au C.N.U.S.C.. La place-mémoire nécessaire à son utilisation est d'environ $4.d + 100$ K octets, avec d égale à la dimension du tableau dans le programme principal, dont l'ordre de grandeur est :

$$d \approx n.p/4 + 10.(n+p) + p.(p+1)/2$$

n étant égale à la plus grande des deux dimensions du tableau de données analysé. Par exemple, un tableau de 300×5000 nécessiterait environ 2000 K octets de mémoire centrale.

Usuellement, nous analysons des tableaux comprenant entre 100 et 250 indicateurs par 100 à 700 sujets pour des coûts raisonnables: par exemple, l'analyse du tableau présenté au § 3, de dimension 46 individus \times 160 indicateurs a nécessité 5 secondes CPU de calcul, soit un coût CNRS de 1.25 Frs (coût de calcul uniquement).

3 Exemple d'application en psycho-sociologie : une étude sur l'idéologie communiste (J.M. CASSAGNE, M. DOMENC, B. GUELFUCCI, J. LARRUE, M. REINERT - L.A. 259, Université Toulouse Le-Mirail)

Dans le cadre de cet article, nous ne présenterons que la partie de cette recherche en rapport avec la méthodologie proposée.

(*) DURAND (1961) - Solutions numériques des équations algébriques. Masson & Cie (tome 2).

Aussi, résumons rapidement ; il s'agit d'étudier les réponses de 46 militants, membres du Parti Communiste, réponses recueillies à l'occasion d'entretiens semi-directifs, à la question suivante :

"être communiste, ça veut dire quoi ?".

Les 46 militants se répartissent dans 3 sections différentes :

- a . 19 militants participent à des cellules d'entreprises (section 3: i.e. individus dont le n° d'identification commence par 3),
- b . 9 militants, à des cellules de quartier, (section 2),
- c . 18 militants, à des cellules universitaires (section 1).

Après un premier transcodage, les entretiens ont été analysés à l'aide d'un logiciel informatique que nous sommes actuellement en train de réaliser pour faciliter les opérations d'analyse de contenu.

Deux procédures d'analyses ont été utilisées :

a . Une analyse thématique classique (*) une grille d'analyse est définie *a priori*; on découpe des entretiens en "unités" de texte, puis chaque unité de texte est affectée à une catégorie particulière et une seule (ou bien reste non codée).

b . Une analyse lexicale par contexte : c'est cette dernière que nous proposons de présenter.

Des 1538 mots différents recensés, certains ayant été exclus *a priori* comme non pertinents par rapport à l'objet d'étude, les articles, les conjonctions, etc. nous avons retenu, après regroupement de certains d'entre eux, synonymie de sens, verbes conjugués à différents temps, etc. les 160 termes ayant été utilisés par au moins quatre militants différents (**).

Le codage retenu pour construire le tableau de données est de type logique : présence/absence d'un mot dans un protocole. Ce choix se justifie par le fait qu'un mot est très rarement utilisé par plus de 50% de la population et par conséquent que sa présence en elle-même est déjà très significative. Ce choix permet aussi d'éliminer un certain nombre d'artéfacts, dû par exemple aux tournures de style ou aux tics verbaux. A titre indicatif, voici les mots utilisés par au moins 10 sujets :

Faire (25), vie (22), P.C. (22), changement (21), société (20), pour (18), faut (18), tous (18), travail (17), lutter (16), essayer (15), problème (14), idées (14), autre (14), vouloir (13), actuellement (11), différence (11), justice (11), possibilité (11), salaire (11), voir (11).

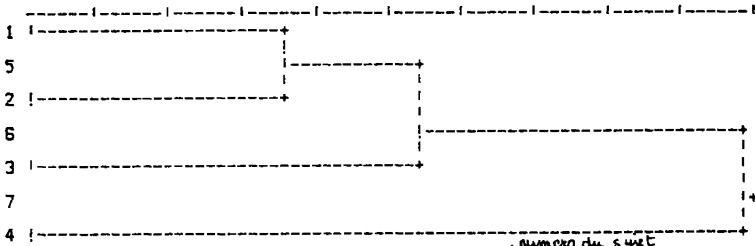
Le tableau de données binaires soumis à la classification est de dimension 46 militants x 160 mots. L'appartenance aux différentes sections est introduite en tant "qu'éléments supplémentaires".

(*) Voir BARDIN L. (1977). *L'analyse du contenu*, P.U.F. .

(**) Il semble préférable de conserver un lexique de départ assez vaste (voir § 4).

3.2 Principaux résultats obtenus (avec TSJ = 3. et CTEST = .3) :

DESCRIPTION DE LA HIERARCHIE DES CLASSES DE SUJETS (NOTATION ASCENDANTE)



PARTITION TERMINALE DES SUJETS

1101	4;	1102	2;	1103	1;	1104	4;	1105	4;	1106	4;	1201	4;	1202	1;	1203	2;	1204	1;
1205	4;	1206	4;	1207	2;	1208	2;	1209	2;	1210	2;	1212	1;	1213	3;	1214	4;	2301	2;
2302	3;	2303	1;	2304	1;	2305	3;	2306	4;	2307	1;	2308	2;	2309	3;	3501	1;	3502	1;
3503	1;	3504	2;	3505	1;	3506	1;	3507	3;	3508	3;	3509	1;	3510	1;	3513	3;	3514	1;
3515	1;	3518	1;	3611	1;	3612	1;	3616	1;	3717	1;								

numero du sujet
numero de la classe terminale

PARTITION EN 4 CLASSES : 1 2 3 4
CH12 = 0.6860E+03 INERTIE = 0.5519E+00

NUNERO FREQUENCES		1 NOMBRE DE SUJETS : 21		NUNERO FREQUENCES		2 NOMBRE DE SUJETS : 9		CH12 IDENTIFICATION	
FREQUENCES		POURC.		CH12 IDENTIFICATION		POURC.		CH12 IDENTIFICATION	
6	7	85.71	5.24	3	4	75.00	8.55		.CATHOLIQUES
9	11	81.82	7.62	3	6	50.00	4.06		.DESACCORD
9	11	81.82	7.62	3	7	42.86	2.85		.INJUSTICES
5	5	100.00	6.68	4	4	100.00	18.01	X1...	.BONHEUR
5	5	100.00	6.68	3	5	60.00	5.83	X1...	.EPANOUISSEMENT
4	4	100.00	5.22	6	17	35.29	4.24	X1...	.HOMME
5	6	83.33	3.95	4	8	50.00	5.70	X1...	.INEGALITES
8	10	80.00	6.08	3	6	50.00	4.06	X1...	.PROGRES
4	4	100.00	5.22	4	6	66.67	9.73	X1...	.SUPPRIMER
5	6	83.33	3.95	3	7	42.86	2.85	X31...	.AGIR
5	5	100.00	6.68	6	19	31.58	2.97*		*SECTION UNIVERSITAIRE
6	6	100.00	8.21						
8	9	88.89	8.43						
8	9	88.89	8.43						
7	7	85.71	5.24						
5	5	83.33	3.95						
8	10	80.00	6.08						
9	11	81.82	7.62						
9	10	90.00	10.13						
6	6	100.00	8.21						
4	4	100.00	5.22						
5	6	83.33	3.95						
4	4	100.00	5.22						
4	4	100.00	5.22						
4	4	100.00	5.22						
4	5	100.00	6.68						
9	11	81.82	7.62						
4	4	100.00	5.22						
12	17	70.59	6.76						
5	5	100.00	6.68						
7	10	70.00	3.05						
8	11	72.73	4.27						
8	10	80.00	6.08						
15	18	83.33	16.92						
8	10	80.00	6.08						
12	16	75.00	8.52						
5	5	100.00	6.68						
13	18	72.22	8.41						
12	18	66.67	5.26						
8	8	100.00	11.53						
7	10	70.00	3.05						
4	4	100.00	5.22						
5	6	83.33	3.95						
5	6	83.33	3.95						
4	4	100.00	5.22						
14	18	77.78	12.30*						*SECTION ENTREPRISE

NUNERO 3 NOMBRE DE SUJETS : 7
FREQUENCES POURC. CH12 IDENTIFICATION

NUNERO 4 NOMBRE DE SUJETS : 9
FREQUENCES POURC. CH12 IDENTIFICATION

3.3 Remarques préliminaires : Dans la présentation graphique de l'arbre, nous avons numéroté les classes de la partition terminale de 1 à 4, et les classes non-terminales de 5 à 7 (par exemple, le numéro 5 renvoie à la réunion des classes 1 et 2).

En effet, bien que nous ne présentons ici, que la description du profil des classes terminales, le programme permet aussi bien la description des autres classes (la classe 7 exceptée... puisqu'elle regroupe toute la population). Dans cette présentation, n'apparaissent que les indicateurs associés à la classe considérée avec un $\chi^2 \geq 2.7$ (seuil modifiable), et dont la fréquence des "présences" est supérieure à la fréquence théorique d'indépendance.

Ce type de description peut bien sûr être effectué après n'importe quelle classification hiérarchique sur les individus. Toutefois la technique que nous proposons, qui permet la définition de classes très contrastées, est plus particulièrement adaptée à ce mode de représentation.

3.4 Commentaires : Certains mots sont précédés d'une clé. Ces clés permettent de présenter ensemble les mots que nous avons associés à un même contexte interprétable. Elles nous serviront à définir les indicateurs de contexte.

- La première classe regroupe 21 sujets provenant principalement de la section d'entreprise (14 des 18 militants de cette section).

Le contexte général se rapporte assez clairement à un engagement dans la lutte sociale, revendicative, de type syndical, principalement associée au milieu du travail.

A l'aide des clés X41 et X42, nous avons distingué deux indicateurs de contexte que nous associons aux "hypothèses" de contenu suivantes :

- X41 ~ vocabulaire de la "lutte sociale",
- X42 ~ vocabulaire des "impératifs d'action".

- La seconde classe plus réduite regroupe 9 sujets dont 7 proviennent de la section universitaire (section 1).

Pour les sujets de cette classe, l'engagement idéologique semble avoir une connotation essentiellement humaniste, teinté de valeurs morales et idéalistes :

- X1 ~ vocabulaire des "valeurs humanistes"

- La troisième classe ne comprend que 7 sujets. Elle se particularise par l'utilisation de 8 mots traduisant assez bien une idéologie de "défense de classe".

- X2 ~ vocabulaire de la "lutte des classes"

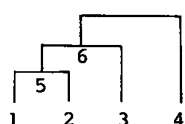
- La quatrième classe, enfin, comprend 9 sujets dont 8 proviennent de la section universitaire. Le contexte général semble faire référence à un engagement dans le combat idéologique en fonction d'une analyse de la société. Le vocabulaire utilisé est plus "théorique" et d'une certaine manière plus institutionnalisé. Nous avons distingué deux indicateurs de contexte.

- X31 ~ vocabulaire de "l'engagement idéologique",
- X32 ~ vocabulaire de "l'analyse théorique".

Remarque : sur les 30 thèmes recensés pour l'analyse du contenu, 21 sont associés à au moins un indicateur de contexte de manière significative ($p \leq 0.05$).

3.5 Comparaison des résultats obtenus avec ceux issus de la C.A.H.

3.5.1 Rappel des résultats obtenus avec la C.D.H. :



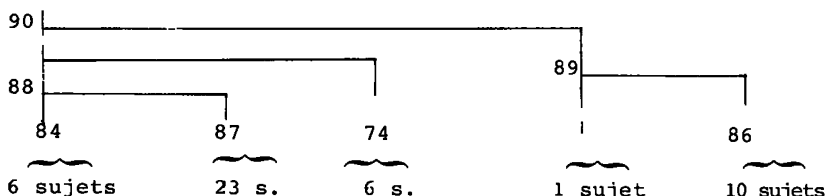
Inertie interclasse de {1,2,3,4} : 0.5519

Inertie interclasse de {5, 3, 4} : 0.3855

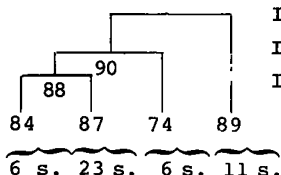
Inertie interclasse de {6, 4} : 0.2054

3.5.2 Résultats obtenus avec la C.A.H. :

a) Présentation de l'arbre (au-dessus du noeud 87) :



b) Pour comparer avec la C.D.H., nous avons utilisé le même critère de définition des classes : en descendant l'arbre, seules les classes de cardinal supérieur à 26 sont analysées. De cette manière, on obtient le dendrogramme suivant :



Inertie interclasse de {84,87,74,89} : 0.5224

Inertie interclasse de {88, 74, 89} : 0.3735

Inertie interclasse de {90, 89} : 0.1966

3.5.3 Comparaison des partitions terminales

Pour chaque triplet, on indique

- le numéro du sujet
- le numéro de la classe ter. CDH
- le numéro de la classe ter. CAH, avec le recodage suivant :
84 = 2 ; 87 = 1 ; 74 = 3 ; 89 = 4.

1101 4 4. 1102 2 1. 1103 1 1. 1104 4 4. 1105 4 4. 1106 4 4. 1201 4 4.
1202 1 2. 1203 2 1. 1204 1 1. 1205 4 4. 1206 4.1. 1207 2 1. 1208 2 4.
1209 1 2. 1210 2 1. 1212 1 2. 1213 3 3. 1214 4 4. 2301 2 1. 2302 3 3.
2303 1 1. 2304 1 1. 2305 3 3. 2306 4 4. 2307 1 2. 2308 2 2. 2309 3 4.
3501 1 1. 3502 1 4. 3503 1 1. 3504 2 1. 3505 1 1. 3506 1 1. 3507 3 1.
3508 3 3. 3509 1 1. 3510 1 1. 3513 3 3. 3514 1 1. 3515 1 1. 3518 1 1.
3611 1 1. 3612 1 1. 3616 1 2. 3717 1 1.

On remarque la proximité des résultats obtenus : 32 des 46 sujets sont classés dans une même classe (16 pour la classe 1 ; 2 pour la classe 2 ; 6 pour la classe 3 ; et 8 pour la classe 4).

Ce résultat indique la forte structure existant dans ces données, sans doute due à la différence existant entre les sections (université, entreprise, quartier).

On constate toutefois qu'au niveau envisagé, partitions en 4 classes et moins, les résultats obtenus avec la CDH sont sensiblement meilleurs, si l'on prend comme critère de comparaison l'inertie inter-classe.

N.B. Des analyses complémentaires portant sur les mêmes données se trouvent dans la note [COMP. LEX.] de A. Aït-Hamlat (C.A.D. Vol VIII n° 2 pp 199-204).

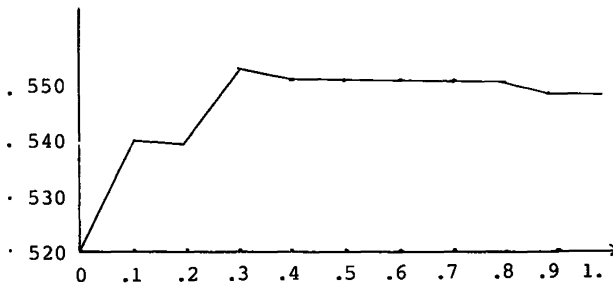
4 Comportement de l'algorithme et stabilité des résultats

Seules les valeurs des paramètres TSJ et CTEST (définis au § 2.4) peuvent influencer les résultats obtenus.

Nous comparons ci-dessous, les différentes partitions des individus en quatre classes, obtenues en faisant varier les valeurs de ces deux paramètres.

Le critère de comparaison choisi est le moment d'ordre deux de ces différentes partitions (ou inertie inter-classes appelée aussi inertie extraite par la partition, l'inertie totale du nuage pouvant être décomposée à l'aide du théorème de Huyghens).

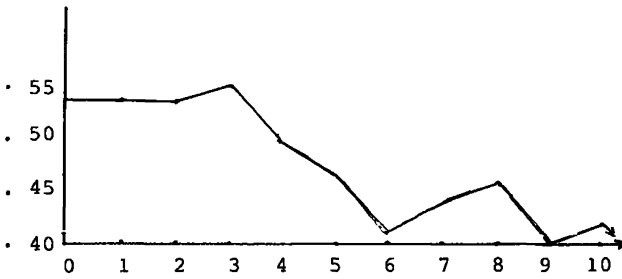
a. Variation de l'inertie extraite en fonction de la valeur de CTEST (la valeur de TSJ est fixée à 3).



La variation du taux d'inertie est faible (entre 10.9 et 11.6% de l'inertie du nuage, qui est égale à 4.757).

On remarque une légère dégradation des résultats lorsque CTEST < .3. Cet aspect semble montrer l'utilité de la définition d'un empiètement pour les classes d'indicateurs.

b. Variation de l'inertie extraite en fonction de la valeur de TSJ (la valeur de CTEST est fixée à .3).



On remarque une bonne stabilité des résultats lorsque $TSJ \leq 3$, et une dégradation rapide sinon.

Rappelons que sur les 160 mots analysés, 24 mots ont une fréquence supérieure à 10 : lorsque TSJ augmente, le nombre de mots effectivement analysés diminue rapidement.

La dégradation des résultats pour $TSJ > 3$, semble donc indiquer la nécessité de conserver dans l'analyse, les mots de faible fréquence, qui malgré leur aspect "informationnel" assez aléatoire (une réponse de plus ou de moins peut modifier sensiblement leur position dans le nuage) jouent un rôle non négligeable par leur nombre, dans l'obtention d'un classement optimal.