

UNIVERSITE DE DROIT, D'ÉCONOMIE ET DES SCIENCES  
D'AIX-MARSEILLE

FACULTÉ DES SCIENCES ET TECHNIQUES DE SAINT-JÉRÔME

---

THESE présentée

par Charles HUOT

pour obtenir le grade de Docteur en Sciences

de l'Université de Droit, d'Économie et des Sciences d'AIX-MARSEILLE

(AIX-MARSEILLE III)

Spécialité :

Sciences de l'Information et de la Communication

ANALYSE RELATIONNELLE POUR LA VEILLE  
TECHNOLOGIQUE: VERS L'ANALYSE AUTOMATIQUE DES  
BASES DE DONNÉES

soutenue le 16 décembre 1992 devant la Commission d'examen:

**M. Dou Henri**

Professeur

**M. Marcotorchino François**

Professeur associé Paris VI

**M. Roux Michel**

Professeur Aix-Marseille II

**M. Paoli Clément**

Professeur associé Marne la Vallée

**M. Quoniam Luc**

Maître de conférence

**Melle Bédécarrax Chantal**

Docteur

*à mes enfants Mathilde et Joseph...*

### **Résumé de la thèse:**

L'émergence de la veille technologique depuis quelques années en France, sensibilise les décideurs publics ou privés à la problématique posée par l'intelligence de leur environnement. Il y a déjà une trentaine d'années que l'on a commencé à stocker dans des bases de données des références bibliographiques relatives à des documents au sens large. Aujourd'hui ces informations correspondent à 80 % de la connaissance et elles sont accessibles à tous. Ce phénomène a donné naissance il y a une quinzaine d'années à une discipline, la bibliométrie. La lourde tâche que constitue l'analyse automatique de ces gisements de données est confiée à des outils bibliométriques fiables et performants. Les recherches présentées dans la thèse ont porté sur l'application, dans le domaine de la veille technologique, d'une méthode de classification automatique appelée l'analyse relationnelle. Cette méthode d'analyse des données opérationnelle en veille technologique, permet de pallier un certain nombre de défauts que présentaient les méthodes d'analyse employées jusqu'alors. Elle présente un caractère innovateur en rendant possible l'analyse d'informations jusqu'alors inexploitables.

**MOTS CLEFS:** information stratégique, analyse de données, veille technologique, analyse relationnelle, bases de données, indice de similarité, brevets, documentation automatique, analyse factorielle relationnelle, bibliométrie.

## Remerciements

---

Avant de vous présenter ce travail, j'aimerais remercier toutes les personnes qui par leur aide ou leurs encouragements m'ont permis de réaliser cette thèse.

En tout premier lieu je tiens à remercier Chantal Bédécarrax, ingénieur de recherche au Centre Européen de Mathématiques Appliquées d'IBM. Pendant deux années Chantal a encadré cette thèse en me donnant l'occasion de collaborer à l'écriture de nombreux articles que nous avons co-signés et qui forment la majeure partie du travail présenté ici. Merci beaucoup Chantal.

Je tiens également à remercier chaleureusement François Marcotorchino directeur du Centre Scientifique d'IBM France, Henri DOU, directeur du Centre de Recherche Rétrospective de Marseille, et Parina Hassanaly responsable des enseignements du CRRM pour le sujet qu'ils ont bien voulu me confier et qui m'a passionné. pendant ces trois années.

Mes remerciements s'adressent également à Luc Quoniam, maître de conférences au CRRM de Marseille pour son soutien scientifique et moral.

Je remercie mes amis William Nivol et Hervé Rostaing, thésards en veille technologique également, avec qui j'ai eu l'occasion de collaborer à de nombreuses reprises.

Une thèse ne peut s'effectuer dans de bonnes conditions que si les équipes auxquelles on est intégré vous supportent. Ce fut le cas, et j'en remercie l'ensemble des équipes du Centre Européen de Mathématiques Appliquées d'IBM et du Centre de Recherche Rétrospective de Marseille. Merci en particulier à Pierre Michaud, Pascal Coupet, Hammou Messatfa et Albert La Teia.

Enfin, pour avoir vécu avec moi cette aventure, m'avoir soutenu de manière inconditionnelle et de plus m'avoir fait connaître les joies de la paternité pendant le temps qu'a duré cette thèse, je remercie mon épouse Anne.

Merci encore à tous.

## **Avant** propos

---

Ma thèse s'est déroulée dans le cadre d'une activité industrielle au sein du Centre Européen de Mathématiques Appliquées d'IBM France. La conjoncture professionnelle dans laquelle se sont effectués mes travaux, ainsi que la mise en application immédiate lors de projets industriels concrets, m'ont poussé à avoir une réactivité forte face au processus de recherche traditionnelle, obligeant à une publication rapide des travaux réalisés.

Dans son ouvrage consacré à l'analyse de la recherche scientifique en France, [Dehe90] le professeur Paul Deheuvels de l'université Paris VI préconise de publier beaucoup et de préférence en anglais. Pour cela il s'appuie sur un slogan bien connu dans les universités américaines "*publish or perish*" (publiez ou périssez). C'est cette voie que j'ai choisie et c'est pour cette raison que ma thèse se présente sous la forme de recueil d'articles publiés dans des revues scientifiques, assorti d'une introduction générale expliquant le lien causal entre ceux-ci.

Dans un premier temps seront présentés les articles dits applicatifs dans lesquels sont reprises des études concrètes de travaux réalisés pour des clients industriels. Ces articles reproduisent à une échelle prototypale, ou sur un plan méthodologique, des réalisations qui doivent, pour des raisons de protection de l'information, demeurer confidentielles. A la fin de la thèse sont regroupés des articles plus techniques qui touchent à des développements récents de la méthode d'analyse employée. C'est au titre de l'utilisation des résultats de ces recherches dans le domaine du traitement des informations que je les ai fait figurer en annexes techniques.

## **Table des matières**

---

---

<b>1.0</b>	<b>Introduction générale</b>	1
1.1	Introduction	2
1.2	Présentation de la thèse	3
1.3	Synthèse des articles composant la thèse	7
1.3.1	La veille technologique	7
1.3.2	Application de l'analyse relationnelle à la veille technologique: des outils d'analyse de l'information documentaire	9
1.3.3	Application d'une nouvelle méthode de classification automatique en veille technologique: l'analyse factorielle relationnelle	11
1.3.4	L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise	13
1.3.5	A new method for analysing downloaded data for strategic decision ...	18
1.3.6	Analyse relationnelle: des outils pour la documentation automatique	21
1.3.7	Développement d'indicateurs pour l'analyse factorielle-relationnelle ...	21
1.4	Conclusion	22
<b>2.0</b>	<b>La Veille Technologique</b>	<b>24</b>
2.1	Introduction	25
2.2	Pourquoi la Veille Technologique?	28
2.3	La mise en place d'une structure de veille	29
2.3.1	Une Doctrine	30
2.3.2	Une Méthodologie	31
2.3.3	Une Structure	32

2.4 Les Outils de la Veille Technologique . . . . .	35
2.4.1 Les Bases et les banques de données . . . . .	35
2.4.2 Bibliométrie, scientométrie, infométrie . . . . .	38
2.4.3 Analyse des informations . . . . .	38
2.5 Conclusion . . . . .	44

---

**3.0 Application de l'A.R. à la V.T.: des outils d'analyse de l'information documentaire . . . . . 45**

3.1 Introduction . . . . .	47
<b>3.2 Présentation générale des données . . . . .</b>	<b>48</b>
3.2.1 Extraction des références . . . . .	48
3.2.2 Présentation relationnelle . . . . .	49
3.3 Problèmes traités . . . . .	51
3.3.1 Traitement des champs Numéros de Brevets et codes CIB . . . . .	54
3.3.2 Traitement des champs Noms de Sociétés et codes CIB . . . . .	57
3.3.3 Récapitulatif . . . . .	60
3.4 Ouvertures . . . . .	61

---

**4.0 Application d'une nouvelle méthode de classification automatique en veille technologique: L'AFR . . . 64**

4.1 Introduction . . . . .	65
4.2 Stratégie d'interrogation . . . . .	66
4.3 Analyse par les codes DERWENT . . . . .	68
4.3.1 La Matrice Initiale . . . . .	68
4.3.2 L'Analyse Factorielle Relationnelle . . . . .	71
4.4 Annexes . . . . .	81

---

<b>5.0 L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise</b> . . . . .	<b>92</b>
5.1 Introduction . . . . .	93
<b>5.2</b> Constitution du corpus de références . . . . .	95
<b>5.3 Les premiers résultats</b> . . . . .	96
<b>5.4</b> Construction des tableaux à analyser . . . . .	96
5.4.1 Choix des niveaux hiérarchiques . . . . .	96
5.4.2 Construction des matrices de présence/absence . . . . .	97
5.5 La classification des brevets . . . . .	98
5.6 La classification des codes . . . . .	100
5.7 Conclusion . . . . .	101
<hr/>	
<b>6.0 A new method for analysing downloaded data for strategic decision</b> . . . . .	<b>102</b>
6.1 Introduction . . . . .	103
6.2 Basic information . . . . .	104
6.3 Different ways to analyze a specific field . . . . .	105
6.3.1 Frequency analysis . . . . .	105
6.3.2 Pairing techniques . . . . .	106
6.3.3 Matrixbuilding . . . . .	107
6.4 Classical methods of data analysis . . . . .	107
6.5 Relational Analysis . . . . .	108
6.5.1 Data representation, basic tables . . . . .	109

6.6 Relational Factorial Analysis	111
6.7 Conclusions	113
<hr/>	
7.0 Analyse Relationnelle des outils pour In documentation automatique	120
7.1 Introduction	121
7.2 Méthodologie	122
7.2.1 Classification simple	124
7.2.2 Sériation par blocs	130
7.2.3 La Quadri-Décomposition	133
7.3 Conclusion	139
<hr/>	
8.0 Développement d'indicateurs pour L'Analyse Factorielle-Relationnelle	140
8.1 Introduction	141
8.2 Développements d'indicateurs pour l'interprétation des résultats d'une Analyse Factorielle Relationnelle	141
8.2.1 Inertie totale du nuage des individus	141
8.2.2 Inertie inter-classe	142
8.2.3 Inertie between entre classes	144
8.2.4 Inertie intra classe	146
8.2.5 Développement d'indicateurs d'analyse du résultat	147
8.3 Application à la classification factorielle-relationnelle des félidés	152
8.3.1 Diagrammes factoriels et représentation des classes	154
8.3.2 Indicateurs globaux	156
8.3.3 Indicateurs par classe	157
8.3.4 Indicateurs par individu	161
8.4 Conclusion	163

**8.5 Annexes** ..... **164**

---

**9.0 Bibliographie** ..... **168**



# *Introduction générale*

*Charles Huot*

Octobre 1992

## 1.1 Introduction

L'émergence de la veille technologique depuis quelques années en France, sensibilise les décideurs publics ou privés à la problématique posée par l'intelligence de leur environnement. Ce phénomène donne lieu à plusieurs questions, parmi lesquelles la première est certainement: où trouver les informations nécessaires à l'accomplissement de cette tâche? Les sources d'informations sont multiples. Elles peuvent être informelles (réunions de travail, rencontres professionnelles, visites chez des fournisseurs, voyages à l'étranger, etc.) ou plus formelles (journaux, rapports de stages, cassettes vidéo ou audio, livres, etc.). La vigilance pour être **efficace**, doit s'inscrire dans le temps, c'est une activité de tous les instants. A l'aube du troisième millénaire, alors que l'homme a réussi à fabriquer des ordinateurs dotés d'énormes capacités de stockage, de traitement, et de transformation de données, il est légitime de penser que cet outil fantastique va pouvoir l'assister dans cette tâche de surveillance et de synthèse d'informations relatives à son environnement scientifique, technique et technologique. Il y a déjà une trentaine d'années que l'on a commencé à stocker dans des bases de données des références bibliographiques relatives à des documents au sens large (articles de revues, thèses, brevets, normes, etc.). Aujourd'hui ces informations correspondent à 80 % de la connaissance et elles sont accessibles à tous. Ce phénomène a donné naissance il y a une quinzaine d'années à une discipline, la bibliométrie. Elle a pour objet l'étude des ensembles de références bibliographiques contenues dans les bases de données.

Ceci afin de permettre initialement aux administrateurs de la recherche d'évaluer l'efficacité de leur politique. Il a fallu pour cela créer des outils d'analyse automatique afin d'assurer la systématisation des synthèses réalisées. Petit à petit, s'appuyant sur des méthodes d'analyse des données de plus en plus variées, cette discipline a connu un essor important.

Aujourd'hui, alors que les bases de données comptent parmi les plus importantes sources d'informations en veille technologique, la lourde tâche que constitue l'analyse automatique de ces gisements de données est confiée à des outils bibliométriques fiables et performants.

Parmi les méthodes employées en bibliométrie, les méthodes d'analyse des données multivariées (analyse factorielle, classification automatique, etc.) sont particulièrement bien adaptées pour l'analyse en profondeur des informations "enfouies" dans d'énormes corpus de références. Il arrive en effet que l'explication de certains phénomènes ne s'exprime pas à travers l'étude d'une seule variable, mais à travers la composition et la confrontation d'un ensemble d'entre elles.

## 1.2 Présentation de la thèse

Mes recherches ont porté sur l'application, dans le domaine de la veille technologique, d'une méthode de classification automatique appelée l'analyse relationnelle. Ces recherches ont donné lieu à plusieurs articles, publiés dans des revues scientifiques, dont certains composent cette thèse. Ils sont au nombre de sept. En voici la liste:

1. C. Huot, C. Bédécarrax: «**La veille technologique**», Publication Scientifique et Technique IBM, 4, (à paraître fin 1992).
2. C. Bédécarrax, C. Huot: «**Application de l'Analyse Relationnelle à la Veille Technologique: des outils d'analyse de l'information documentaire.**», *Les systèmes d'informations élaborées*, Congrès S.F.B.A., juin 1991.
- 3 C. Huot: «**Application d'une nouvelle méthode de classification automatique en veille technologique: l'analyse factorielle relationnelle** », S.F.B.A, n° 8, décembre 1990.

4. C. Bédécarrax, C. Huot, H. Rostaing, L. Quoniam, W. Nivol: «**L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise (I)**», les journées de l'ADEST, juin 1992.
5. C. Huot, L. Quoniam, H. Dou: «**New method concerning analysis of downloaded data for strategic decision**», *Scientometrics*, Vol 4, n° 2, pp 279-294, 1992.
6. C. Bédécarrax, C. Huot: «**La veille technologique**», sous la direction de Henri Dou et Hélène Desvals, «Analyse Relationnelle: des outils pour la documentation automatique», *Dunod, 1992*.
7. C. Bédécarrax, C. Huot: «**Développement d'indicateurs pour l'Analyse Factorielle Relationnelle**», étude MAP-005 du CEMAP IBM, mai 1992

Les cinq premiers articles sont très orientés sur les préoccupations de la veille technologique, tant sur le plan de la méthodologie que sur les types de traitements ou encore les exemples d'applications. J'ai sélectionné ces articles car ils représentent un panel significatif des travaux de recherche réalisés autour de ces thèmes.

Les deux autres articles décrivent des développements plus mathématiques relatifs à l'analyse relationnelle, c'est la raison pour laquelle j'ai choisi de les faire figurer en annexes techniques.

Enfin pour éviter les duplications et faciliter la tâche du lecteur, j'ai pris l'option de présenter une bibliographie générale à la fin de la thèse. Ainsi les références signalées dans les divers articles sont à consulter à la fin de l'ouvrage.

Dans le souci de ne présenter que les travaux les plus synthétiques et les mieux ciblés sur le sujet de la thèse, j'ai volontairement écarté un certain nombre de publications que j'ai pu réaliser, seul ou en collaboration avec d'autres chercheurs.

Voici la liste des références bibliographiques relatives à ces articles ou publications non présentés dans la thèse.

1. L. Quoniam, H. DOU, C. Huot: «**Les méthodes d'analyse des données face à l'information stratégique et l'innovation**», *Co lloque in ternational sur les*

*méthodes de sériation par blocs et applications*, ENSAIS, Strasbourg, avril 1990.

2. C. Huot, D. Gibert: «**Etude de l'évolution de la Chimie Macromoléculaire à Marseille de 1979 à 1988**», S.F.B.A, n° 7, juin 1990.
3. C. Huot: «**Etude des applications de l'Analyse Relationnelle à la veille technologique**», étude de service, septembre 1990.
4. L. Quoniam, H. DOU, C. Huot: «**Modélisation d'un outil stratégique pour la veille technologique.**», soumis au Prix d'Excellence IBM, septembre 1990.
5. L. Quoniam, C. Huot: «**Développement d'un outil graphique pour une analyse automatique des bases de données appliqué à la veille technologique.**», soumis au Prix d'Excellence IBM, septembre 1991.
6. C. Bédécarrax, C. Huot: «**Une nouvelle vision de la veille technologique.**», Actes, Journée CEMAP-CRRM, juin 1991.
7. H. Rostaing, W. Nivol, L. Quoniam, C. Bédécarrax, C. Huot: «**L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise (II)**», les cahiers de l'ADEST, (à paraître fin 1992).
8. C. Bédécarrax, C. Huot: «**A new methodology for systematic exploitation of technology databases**», (à paraître dans le journal Information Processing and Management).

L'un des objectifs d'une introduction générale, pour une thèse basée sur un recueil d'articles, est de donner l'occasion d'exprimer ce que l'on n'a pas pu dire dans ces articles, faute de place, de temps, ou simplement parce que de nouvelles idées sont apparues à la relecture. A cette opportunité, s'ajoute celle de pouvoir porter un jugement critique sur ce que l'on a écrit quelques mois, voire quelques années auparavant.

Le lecteur aura l'occasion de le constater au fil de la thèse, mes recherches ont consisté pour une grande part, et le titre est ici pour le rappeler, à étudier et développer les points de convergence entre deux mondes: celui des mathématiques

ou plus précisément celui de l'analyse des données, et celui de la veille technologique sous-partie du monde de l'information.

Le point de départ de mon travail a été fondé sur les préoccupations suivantes. La connaissance s'accroît de jour en jour. Ceci se traduit par une croissance exponentielle du nombre de publications scientifiques, techniques ou technologiques. Ces publications revêtent des formes diverses telles que: articles dans des revues, rapports de congrès, fiches techniques, brevets d'invention, normes, etc. Si l'on se place sur le plan des informations disponibles dans les entreprises, ce phénomène se voit amplifié par l'augmentation constante du nombre de serveurs télématiques, de bases de données diverses et variées et *in fine* du nombre de références bibliographiques accessibles.

Pour faire face à cette évolution considérable du nombre d'informations disponibles, il a fallu créer des outils d'analyse automatique des contenus des bases de données. Le recours à ces outils, permettant d'effectuer des synthèses automatiques de l'information issue de l'interrogation des bases, a été initialisé dans les années 80, par des équipes de recherche universitaires parmi lesquelles le CRRM (Centre de Recherche Rétrospective de Marseille) dirigé par le professeur Henri Dou, a joué un rôle de leader.

Ces outils se scindent en deux grandes catégories.

Tout d'abord les outils qui restructurent l'information sur un plan physique, par exemple en homogénéisant les formats entre des données de sources différentes, ou encore en fabriquant des tableaux de données croisant des informations de différentes natures à partir de l'analyse automatique des références. Ces outils sont appelés outils de prétraitement.

Les autres outils sont des outils mathématiques qui analysent l'information en profondeur. Ils sont utilisés à la suite des outils de prétraitement après que les données ont été correctement mises en forme. Ils font, pour une grande part, appel à la statistique et en particulier à l'analyse des données.

Mon soucis, au début de cette thèse, était de trouver une nouvelle méthode d'analyse des données, qui puisse être opérationnelle en veille technologique, en palliant un certain nombre de défauts que présentaient les méthodes d'analyse employées jusqu'alors. Cette nouvelle méthode devait également présenter un caractère innovateur en rendant possible l'analyse d'informations jusqu'alors inexploitable.

Les travaux effectués depuis plus de 15 ans par l'équipe de recherche **d'IBM**, sous la direction de François Marcotorchino, sur l'utilisation de l'**Analyse Relationnelle** à des domaines très variés comme la médecine, la linguistique ou la productique, laissaient présager de bons résultats dans le cadre d'une application en veille technologique.

C'est donc à partir d'un besoin concret, manifesté par des industriels ou des organisations étatiques dans le domaine du traitement d'informations technologiques, que je me suis intéressé à la méthodologie relationnelle et- que j'ai cherché à voir la place qu'elle pouvait prendre dans le monde du traitement des données documentaires.

### **1.3 Synthèse des articles composant la thèse**

#### **1.3.1 La veille technologique**

Fort de 3 années d'expérience dans une discipline en pleine expansion j'ai récemment publié un article intitulé **La veille technologique**. Cet article, écrit au moment où je terminais ma thèse, propose une synthèse sur le sujet et montre les diverses facettes de cette discipline, à savoir: les structures à mettre en place lorsqu'une entreprise veut organiser et structurer sa veille, l'analyse des sources d'informations nécessaires pour atteindre ce but et les outils mis à la disposition des industriels pour les opérations d'analyse de l'information.

En France, où trop souvent information est synonyme de pouvoir, il est difficile d'assurer à celle-ci la même liberté, et la même place naturelle qu'au Japon. Ceci explique à mon avis, pourquoi l'implantation d'une réelle activité de veille dans

notre pays passe d'abord par une démarche théorique, doctrinaire et dogmatique. Ce travail de pédagogue a été pris en charge par quelques personnes parmi lesquelles F.Jakobiak, veilleur technologique chez ATOCHEM, et enseignant au DEA d'information critique, veille scientifique et technique de l'Université de Marseille. A travers ses différents ouvrages, abondamment cités tout au long de cette thèse,<sup>1</sup> il met en place les fondements de cette nouvelle discipline. L'intelligence de l'information n'est pas une chose innée chez les français, comme elle peut l'être chez les japonais. Qu'à cela ne tienne, nous allons apprendre, et nous commençons très sérieusement à le faire. Le nombre croissant de journées d'informations, de formations, de colloques et autres manifestations publiques sur le sujet sont là pour en témoigner.

Au moment où j'écris ces lignes, l'**Institute** for International Research organise une conférence au titre évocateur: *Veille Stratégique, Veille concurrentielle commerciale, technologique*. Parmi les orateurs figurent les plus grands noms de la veille en France actuellement: Henri Dou (CRRM), François Jakobiak (ELF ATOCHEM), Jacques Villain (Société Européenne de Propulsion), Bruno Martinet (Ciments Français), Jean Michel Ribault (ALGOE Management), François Libmann (FLA Consultants) et d'autres encore.

En septembre 1989, le CRRM de l'Université Aix Marseille, centre de St Jérôme, a créé le premier DEA français de veille technologique, avec pour objectif de former des experts dans cette discipline qui pourront ensuite s'implanter dans l'industrie et répondre aux besoins qui s'expriment de toutes parts. Car si le recours aux experts du domaine s'impose dans la structure présentée par F.Jakobiak, la vocation du veilleur aujourd'hui est également d'être un expert dans les systèmes de traitement de l'information sous toutes ses formes. Ceci impose au veilleur technologique moderne de conserver ses compétences d'animateur de réseau, de communicateur, mais aussi d'en acquérir de nouvelles dans le domaine de l'informatique, de la bibliométrie, des systèmes d'analyse des données et d'autres encore.

---

<sup>1</sup> à l'exception de son dernier ouvrage que je n'ai découvert que récemment [Jako92]

Si la France ne possède aujourd'hui qu'un nombre relativement faible de spécialistes dans le domaine du traitement de l'information stratégique et de la veille technologique, des laboratoires de recherche privés ou publics ont en revanche beaucoup travaillé à la fabrication d'outils d'analyse automatique des informations contenues dans les bases de données. Certes ces outils sont sophistiqués et relativement complexes, mais ils ont prouvé leur efficacité dans le cadre d'applications industrielles de très grandes tailles. Leur utilisation courante se met en place dans l'industrie ainsi que dans des services de l'Etat.

Dans la présentation de la veille technologique telle que nous la décrivons aujourd'hui, il est clair que pour assurer une surveillance systématique de l'environnement de l'entreprise, le recours à des outils bibliométriques s'avère fort utile. Cette solution d'analyse automatique et de synthèse d'informations va faciliter grandement la tâche des experts des domaines scientifiques ou techniques étudiés dans la réalisation de dossiers stratégiques. Le rôle des experts en systèmes d'informations et des experts des domaines étudiés est fondamental dans le processus qui part des données brutes (omniprésentes et surabondantes) pour aboutir à une information élaborée (synthétique et validée) qui permettra à un décideur d'agir. C'est une idée qui commence à poindre, le veilleur technologique devra maîtriser les systèmes bibliométriques modernes s'il veut demeurer performant.

### **1.3.2 Application de l'analyse relationnelle à la veille technologique: des outils d'analyse de l'information documentaire**

---

Cet article, présenté au cours des journées d'étude de la SFBA en juin 1991, marque le début d'une collaboration avec Chantal Bédécarrax sur le thème de l'application de l'analyse relationnelle à la veille technologique. C. Bédécarrax, responsable au CEMAP du groupe veille technologique, est titulaire d'un doctorat en mathématiques [Bede89a] de l'université de Paris VI. Elle a collaboré au sein d'IBM au projet de lexicographie computationnelle avec I. Warnesson [Bede89b]. C'est par les débouchés industriels obtenus sur les premiers travaux réalisés en 1990 lors de ma première année de thèse qu'IBM a décidé d'élargir l'équipe de recherche. Ainsi les compétences dans le domaine des mathématiques et plus précisément dans celui de l'analyse relationnelle de Chantal Bédécarrax, alliées à

mes connaissances du domaine de la veille technologique nous ont permis d'avancer très rapidement dans nos recherches. Aujourd'hui, le fait que nous ayons réalisé plusieurs contrats industriels, portant sur des problèmes divers toujours liés à la veille technologique et nécessitant un traitement d'analyse de données va nous permettre de proposer un outil adapté aux principales préoccupations d'analyse automatique des grandes bases de données. Cet article fait état d'un certain nombre d'analyses sur différentes matrices. En effet partant de certains champs constitutifs des références bibliographiques (nom de la société, codes internationaux, numéro de brevets), nous avons imaginé de les croiser entre eux. Ceci se traduit par la génération de matrices différentes par leurs formes et leurs types (cas des matrices de présence-absence, des matrices de fréquences, etc). Pour chacune des matrices résultantes nous appliquons une technique d'analyse adéquate, en explicitant le résultat et son interprétation.

Il s'agit toujours d'effectuer une classification sur l'un des ensembles ou sur les deux simultanément pour regrouper des éléments en fonction des ressemblances qu'ils présentent.

Nous nous sommes limités, dans le présent article, à l'étude des relations entre trois champs: nom de sociétés, numéro des brevets et codes CIB. A chacune des combinaisons des champs **précités** correspond une préoccupation bien spécifique d'analyse de l'information extraite du corpus:

- positionnement relatif des brevets vis-à-vis de la stratégie d'interrogation (connaissance des liens entre différents brevets, regroupement de brevets en classes indépendantes,...)
- mise en évidence des relations entre les divers domaines de recherche ou d'applications
- recherche de liaisons spécifiques entre brevets et domaines d'applications
- identification des stratégies de recherche communes ou spécifiques des entreprises

Cet article se terminait par une ouverture sur des traitements futurs portant sur l'analyse des codifications multiples et l'analyse des pays d'extensions.

En collaboration avec le CRRM, au cours des journées d'étude de l'**ADEST**,<sup>2</sup> nous avons analysé les trois champs de codification présents dans les références WPI de la base de données **Derwent**. Cette étude a permis d'expérimenter l'idée selon laquelle il est possible d'établir une sorte de dictionnaire d'analogie pour les différentes classifications. Le fait de générer des systèmes de traduction entre diverses codifications a une application importante sur le plan industriel. Depuis que les services de propriété industrielle déposent ou stockent des brevets, chaque grosse entreprise a inventé une codification des documents scientifiques ou techniques (brevets, normes, etc) propre à son secteur. Les codifications déjà existantes ne répondant pas avec assez de finesse aux besoins. Mais aujourd'hui, alors que les traitements automatiques des bases de données se généralisent, la normalisation devient une nécessité vitale. Il est impensable de repartir de zéro en abandonnant ce qui a été fait par le passé. Il faut au contraire s'appuyer sur cet acquis formidable qui représente une partie de la mémoire de l'entreprise, pour concevoir un système moderne et compatible avec les exigences d'une compétitivité accrue.

L'une de nos études nous a conduit à nous intéresser au problème du choix des pays d'extensions dans le processus de dépôt de brevets. La matrice correspondante croisait des noms d'entreprises et des pays désignés. Ce travail a permis de faire ressortir clairement la stratégie internationale en matière de protection industrielle (donc de futurs pays d'exploitation) en comparant les divers choix entre entreprises. Ainsi, il est possible pour une société de remettre à plat sa propre politique de dépôts et de mesurer les écarts avec ses concurrents en fonction des différents domaines scientifiques couverts.

### **1.3.3 Application d'une nouvelle méthode de classification automatique en veille technologique: l'analyse factorielle relationnelle**

---

Cet article est le premier qui traite de l'utilisation d'une méthode relationnelle en bibliométrie. Il est la synthèse du travail que j'ai effectué en **DEA** sur l'application

---

<sup>2</sup> article intitulé *L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise* et présenté dans la thèse

des méthodes d'analyse relationnelle **d'IBM** dans l'élaboration de dossier de veille technologique. L'objet de cet article est de relater une expérience d'analyse bibliométrique qui s'appuyait sur une méthode d'analyse des données n'ayant jamais été expérimentée en veille technologique. Comme dans toutes les analyses bibliométriques il s'agissait d'étudier un corpus de références. Dans le cas présent, et en collaboration avec Luc Quoniam du CRRM, nous avons téléchargé 15 références bibliographiques sur le thème des brevets relatifs au traitement de surface des lentilles de contact par de la **papaïne**<sup>3</sup> (4 références sont issues de la base WPI de Derwent Incorporation et 11 de la base USPA, citant celles de WPI). Nous avons extrait de chacune des références présentes les codes de description Derwent. Ces codes sont mis par les personnes qui indexent les brevets chez Dexwent. Ils servent à décrire d'une façon synthétique le contenu technique du brevet. Concrètement ces codes sont utilisés pour interroger la base de données sur un thème de recherche référencé ou bien pour effectuer des analyses bibliométriques. Nous avons donc constitué un tableau croisant le numéro des références bibliographiques avec les codes Derwent présents dans les références. Nous avons ainsi obtenu un tableau dit de présence-absence (à valeurs 1-0 suivant que la référence possède ou non le code Derwent). Ce tableau, également appelé matrice d'incidence en théorie des graphes, est le point de départ de toutes nos analyses. Ce n'est que la retranscription sous forme relationnelle d'une information présente sous une forme linéaire dans les références bibliographiques de départ. La méthode relationnelle s'appuie sur l'analyse de ce tableau de relations brutes pour effectuer la classification des références ou des codes. La méthode expérimentée ici, l'analyse factorielle relationnelle (AFR), a pour objectif de réaliser une cartographie des références ou des codes. Les objets sont projetés dans un espace à deux dimensions • sur lequel on dessine la partition obtenue par la classification **relationnelle**<sup>5</sup> de l'espace de description représenté (références ou codes). Cette

---

<sup>3</sup> La **papaïne** est une substance qui permettant la dissolution des protéines qui se déposent sur les lentilles de contacts les rendant opaques et irritantes.

<sup>4</sup> Il s'agit de l'espace de représentation classique engendré en Analyse Factorielle par les deux premiers vecteurs propres

<sup>5</sup> selon un **critère** compatible **avec** l'analyse factorielle

combinaison résout les problèmes de l'interprétation des analyses factorielles traditionnelles [Paol87].

Même si ce résultat apporte un plus par rapport à l'analyse factorielle classique, son exploitation reste problématique dans le cadre d'une application industrielle du fait de la limite imposée par la représentation graphique. L'expérience nous a montré que les résultats des classifications sont plus facilement exploitables. A partir de ceux-ci et s'appuyant sur des indicateurs de liaisons inter-classes présentés dans l'annexe technique numéro 2<sup>6</sup>, nous pouvons fabriquer des graphes synthétiques représentant les liens entre les classes.

#### **1.3.4 L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise**

---

Cet article a été présenté aux journées de l'ADEST en juin 1992. Outre l'intérêt d'une nouvelle coopération scientifique avec le C.R.R.M, il a permis de confronter les visions respectives de 3 thésards en veille technologique sur une problématique nouvelle.

***Quel apport peut-on attendre d'une étude comparative des trois classifications présentes dans les références de la base WPI de Derwent?***

Les trois classifications étudiées sont:

- les Derwents codes (DC), établis par Derwent
- les Manuels codes (MC), établis par Derwent
- les codes de la classification internationale (CIB), établis par les instituts officiels de dépôts de brevets

L'un des objectifs de l'article est d'apporter une réponse à la carence des méthodes d'analyse des relations engendrées par des combinaisons d'informations issues de

---

<sup>6</sup> Développement d'indicateur pour l'analyse factorielle relationnelle

références bibliographiques. Les classifications documentaires (DC, MC, CI B) servent à qualifier les brevets référencés dans la base de Derwent. Une **classification** documentaire découpe les domaines scientifiques en sections. Si la **classification** est construite à partir d'un principe de hiérarchie, ces sections sont alors elles-mêmes découpées en sous-sections, classes, groupes, sous-groupes... Chaque niveau dans cette hiérarchie de découpage est représenté par une codification.

Le rôle des codifications est, rappelons-le, de caractériser chaque référence en lui affectant un ou plusieurs codes pour décrire le contenu technique ou scientifique du document ou encore ses possibilités applicatives. En quoi réside donc l'intérêt de plusieurs codifications? N'ayant pas de réponse à apporter a priori à cette question, le plus intéressant était d'étudier cette diversité et de chercher quels avantages on pouvait en tirer.

Pour estimer la spécificité de chacune de ces classifications, nous les avons confrontées sur un corpus de références brevets concernant les patchs transdermiques thérapeutiques. L'exploitation par la méthode **d'Analyse Relationnelle** nous a permis d'évaluer deux caractéristiques des relations liant les divers modes de codification.

- les complémentarités de ces classifications documentaires en qualité de descripteurs de brevets: nous désirions estimer si le fait d'utiliser dans les analyses les trois classifications simultanément permettait de mieux décrire les liens réels entre les brevets.
- les correspondances ou les similarités entre les codes de ces classifications au niveau de leur sens: ceci ayant pour but d'apprécier les recouvrements de significations, les codes synonymes et les complémentarités entre les classifications.

Nous avons donc constitué un corpus de références bibliographiques portant sur les patchs. Cette interrogation a été menée de manière très minutieuse pour obtenir un ensemble de réponses le plus homogène possible. Cette stratégie d'interrogation nous a permis de dégager un corpus de brevets cohérent. Une pré-étude a été

menée pour déterminer quels niveaux hiérarchiques satisfaisaient la meilleure solution statistique. Les critères étudiés pour chaque niveau hiérarchique étaient:

- le nombre de codes restants
- le nombre de brevets
- la qualité d'agrégation des brevets par l'**Analyse** Relationnelle pour ce niveau de hiérarchie (nombre d'agrégats et nombre de codes non agrégés)

Le choix s'est porté sur les niveaux hiérarchiques dont les nombres de codes et les nombres d'agrégats pour les différentes classifications documentaires sont proches. Les niveaux hiérarchiques choisis pour l'étude ont donc été:

- les Derwent codes à 3 caractères
- les Manuel codes à 3 caractères
- les codes CIB à 7 caractères

Après ce travail préliminaire, nous avons généré quatre matrices: une pour chacune des trois codifications et une regroupant l'ensemble des codes. Le passage des données textuelles (références) aux données tabulées (matrices) est réalisé par le logiciel DataView du CRRM.

Les matrices sont construites, automatiquement, avec DataView selon la logique suivante:

- extraire tous les codes
- éliminer les codes qui ne sont pas renseignés jusqu'au niveau hiérarchique choisi
- tronquer les codes restant à ces niveaux de hiérarchie
- construire les matrices de présence-absence

Sur ces quatre matrices, naturellement très creuses nous avons effectué des classifications selon l'approche relationnelle classique. Nous avons obtenu des partitions de brevets en fonction des codes et des partitions de codes en fonction des brevets. Les résultats ont été analysés sans difficulté par des personnes qui n'étaient pas des experts de l'analyse relationnelle.

Cette méthodologie nous a permis de mettre en évidence plusieurs caractéristiques:

- grâce à la complémentarité des codes

On relève des détails techniques très précis qui sont en général noyés au travers d'aspects très généraux et par conséquent très difficiles à détecter, mais ceci sans perdre de vue les aspects généraux auxquels ils se rapportent. Donc choisir d'exploiter plusieurs champs de descriptions pour un traitement bibliométrique permet d'aboutir à une meilleure caractérisation du corpus.

- grâce à la correspondance des codes

On met en regard des ensembles de codes qui sont fortement dépendants (ou proches) les uns des autres parce que très souvent employés conjointement dans les références:

- déceler les documents non pertinents de notre corpus c'est-à-dire ceux qui représentent le bruit
- reconnaître les documents originaux (au sens innovateur du terme) qui sont basés sur l'emploi ou la description de techniques, de méthodes, de procédés qui se démarquent des autres.

- la représentation de l'information obtenue par les outils d'analyse.

L'expertise et l'interprétation résultent des traitements analytiques auxquels on a recours. Afin que l'ensemble des experts puisse interpréter de façon objective les résultats, il est nécessaire d'utiliser des méthodes de traitement qui permettent non pas de résumer l'information de départ, mais de la restructurer pour en dégager les faits marquants sans perdre le reste. Trop souvent négligé par les méthodes d'analyse traditionnelles, cet ensemble rebut constitue une part considérable de l'information initiale (lois de Zipf-Bradford). Bien que ces éléments soient présents à des fréquences très faibles, ils n'en contiennent pas moins une information intéressante. En effet, on y relève l'information marginale qui selon le cas peut se traduire en termes d'information innovante ou discordante (le bruit).

L'Analyse Relationnelle est parfaitement appropriée à l'analyse bibliométrique. Elle ne néglige aucune information même si sa faible présence lui donne a priori un caractère mineur.

- le rapport temps d'analyse - temps d'expertise.

Afin que les experts du domaine puissent se consacrer pleinement à l'étude du sujet à travers les résultats livrés par le traitement, celui-ci doit prendre le moins de temps possible. Pour minimiser le rapport temps d'analyse - temps d'expertise, on se doit d'utiliser des méthodes de traitement, de calcul, d'analyse, qui permettent d'obtenir des résultats dans des délais extrêmement brefs et qui peuvent être utilisées de façon systématique afin de réorienter les analyses selon les interprétations que l'on obtient. C'est la première caractéristique qu'un système de surveillance doit vérifier pour assurer un fonctionnement performant. L'accélération de l'obsolescence des technologies impose à la veille technologique d'être le pourvoyeur d'une information élaborée dans des temps restreints. Pourquoi dans ces conditions ne pas profiter des avantages qu'offre, l'informatique pour le traitement des données. C'est dans cet esprit que le logiciel **DataView** et la méthode d'Analyse Relationnelle ont été conçus.

- le recours permanent à des experts différents

Pour que l'analyse délivre une information fiable à vocation stratégique, il est indispensable de valider chaque étape dans l'élaboration du dossier, depuis la sélection de corpus jusqu'à l'interprétation des résultats, par des niveaux d'expertise différents et adaptés.

- Expert du domaine technique étudié
- Expert brevet
- Expert de l'information
- Expert en statistiques

Ce sont toutes ces contraintes que doit respecter un système de veille pour permettre de traiter des sujets pour lesquels la masse et la complexité des **connais-**

sances requises ne peuvent être appréhendées par de simples traitements manuels.

### **1.3.5 A new method for analysing downloaded data for strategic decision**

Cet article a été réalisé en collaboration avec Henri Dou et Luc Quoniam du CRRM. Il est écrit en anglais.

Le premier objectif est de proposer un panorama des méthodes bibliométriques depuis les dénombrements jusqu'aux analyses de données multivariées les plus complexes.

Le second objectif consiste à présenter l'analyse factorielle relationnelle comme un nouvel outil bibliométrique en entrant plus dans le détail de la méthodologie que dans l'article précédemment écrit paru dans la SFBA'.

Tout au long de la présentation de la thèse il est fait mention de diverses techniques d'analyse des informations téléchargées. Toutes ces techniques relèvent du concept général de bibliométrie:

L'organisation de l'information dans les bases de données se présente sous une forme pyramidale. A la base on trouve les références bibliographiques relatives à des articles scientifiques, techniques, économiques, juridiques et autres, des documents brevets, des normes, des rapports de congrès, etc. Ces références sont regroupées dans des bases ou des banques de données en fonction de thèmes généraux ou spécifiques (chimie, physique, brevets, médecine, etc). Au sommet de la pyramide, ces bases ou banques de données sont distribuées par des serveurs privés (**Orbit**, Dialog, Questel, ESA, etc). Ainsi, des milliards de références bibliographiques sont présentes dans quelques milliers de bases et banques de données, elles-mêmes distribuées par une centaine de centres serveurs. Cependant l'objet que nous appellerons "référence" n'est pas l'unité la plus petite de ce

---

<sup>7</sup> Application d'une nouvelle méthode de classification automatique en veille technologique: l'analyse factorielle relationnelle

schéma. En effet la référence se décompose à son tour en objets plus petits que l'on appelle "champs". Ces champs, présents en petite quantité dans une **référence**<sup>8</sup>, sont les unités minimales d'information. Ils sont caractérisés par deux lettres et servent à séparer les différents types d'informations présents dans les références. Il existe ainsi des champs contenant le titre du document, le nom du ou des auteurs, leurs adresses, des codifications diverses, un résumé du document, les pays désignés lorsqu'il s'agit de brevets et d'autres informations encore. Cette répartition en champs est spécifique à chaque base de données même s'il arrive que des serveurs fassent des efforts pour d'homogénéiser leurs différentes bases. Le travail de la bibliométrie consistera à effectuer des analyses sur l'ensemble de ces unités d'information pour les synthétiser, les condenser, en tirer une information ayant une valeur ajoutée importante vis-à-vis de l'information séquentielle présente initialement. Partant des fichiers de données téléchargés sur des ordinateurs personnels depuis les réseaux de serveurs internationaux, l'information va être exploitée au maximum de son potentiel par des méthodes d'analyse entièrement automatiques.

Il ne faut pas faire de discrimination entre les diverses méthodes d'analyse des données. La bonne méthode est celle qui apporte un résultat et une réponse à la question que l'on se pose. Certaines sont plus ou moins complexes que d'autres, ou plus ou moins adaptées au problème auquel on est confronté.

Ainsi parmi les méthodes les plus simples, on trouve la méthode des analyses de fréquences. Cette méthode fait des comptages et du dénombrement sur des listes. C'est l'étude la plus simple qui peut être effectuée sur les champs. On détermine ainsi le nombre de publications par auteur, le nombre de brevets par société, de dépôts au cours du temps, etc. Ces opérations peuvent être réalisées par un simple tableur, la difficulté provenant de la gestion et de la pré-analyse des fichiers téléchargés (nettoyage du fichier, dédoublonnage, remise au format des champs, etc).

---

<sup>8</sup> en moyenne une dizaine d'unité par référence

Une autre méthode intéressante est la méthode des paires. Il s'agit cette fois d'étudier les co-occurrences dans un champ. Car si l'on prend l'exemple du champ DC (**Derwent** codes) il est intéressant de savoir le nombre de fois où un code apparaît dans un corpus, mais il est encore plus instructif de connaître le nombre de fois où deux codes sont présents simultanément dans une référence. Cette méthode donne une vision des ponts qui existent entre différents domaines scientifiques ou techniques. Beaucoup d'études, basées sur cette technique, ont été réalisées par le CRRM.

Une autre technique donne également de bon résultat, il s'agit de la méthode de construction de matrices. Ce n'est pas une méthode d'analyse des données à proprement parler, mais plus un moyen de représentation de l'information. Ainsi l'on indique dans un tableau à double entrée le nombre de relations entre deux champs de natures différentes, par exemple le nom des sociétés et le champ code CIB. A l'intersection on fait figurer le nombre de brevets d'une société contenant un certain code CIB. Ainsi par une simple lecture du tableau résultant, qui peut être de relativement petite taille, on observe un certain nombre de faits intéressants. On détermine des matrices de co-occurrences ou des matrices de fréquences ou des matrices de présence-absence.

La lecture des matrices précédentes par un lecteur humain apporte un certain nombre d'informations, permet de faire apparaître des corrélations, mais c'est une technique relativement frustrante pour étudier en profondeur l'ensemble des relations sous-jacentes générées par les différents champs de chaque référence. Les méthodes d'analyse multivariée permettent justement d'analyser automatiquement les matrices fabriquées précédemment. Parmi les plus courantes, citons l'analyse factorielle et les classifications hiérarchiques. Un certain nombre de contraintes sont liées à l'utilisation de ces méthodes, notamment la taille des matrices utilisées, leurs types (matrices de fréquences, matrices de présence-absence, etc.), ainsi que les interprétations des résultats.

La méthode d'analyse factorielle relationnelle présentée ici, si elle permet de résoudre le problème d'interprétation des résultats sur les projections graphiques,

ne résout pas celui de la taille des corpus visualisables. Cette remarque qui n'apparaît pas dans l'article, résulte de l'application industrielle de la méthode relationnelle, qui m'a amené à revoir cette option. Il n'en reste pas moins vrai que la classification associée à l'analyse factorielle relationnelle, avec notamment le critère de Burt pondéré, reste l'élément de base de l'analyse des matrices utilisées en veille technologique. Le résultat se présente alors non plus sous forme de projections graphiques mais sous forme de classes homogènes de références bibliographiques.

### **1.3.6 Analyse relationnelle: des outils pour la documentation automataue**

Ce chapitre, publié dans le livre de Henri Dou et Hélène Desvals, *Lu veille technologique* (édition Dunod), présente les différents aspects de l'Analyse Relationnelle. Depuis les méthodes d'agrégation simple, de sériation, de quadri-décomposition jusqu'à l'analyse factorielle relationnelle, toutes les méthodes liées à l'analyse relationnelle sont détaillées. La lecture de cette partie nécessite une bonne connaissance préalable de l'analyse des données, c'est pourquoi elle est présentée en annexe technique. Il est vrai que la connaissance de l'outil n'est pas indispensable pour comprendre son utilité en bibliométrie. Ce chapitre sert de référence en matière de description de la méthodologie relationnelle dans le domaine de la veille technologique. Il joue également le rôle de pont entre le milieu de la bibliométrie et celui de l'analyse des données.

### **1.3.7 Développement d'indicateurs pour l'analyse factorielle-relationnelle**

Parallèlement aux travaux relatifs à l'application des diverses méthodes d'analyse relationnelle aux données de la veille technologique, des recherches plus théoriques ont été menées en analyse relationnelle. Cet article, présenté dans la thèse en seconde annexe technique, fait suite aux derniers développements de l'analyse factorielle-relationnelle par François Marcotorchino [Marc91 a]. Les travaux présentés portent sur la recherche d'indicateurs nouveaux pour aider à l'interprétation d'une analyse factorielle-relationnelle. Ils sont notamment très utiles lorsqu'il s'agit de réaliser des *tableaux de bord* connectés sur les sorties de

classification de corpus documentaires. De la recherche mathématique pure à l'application quasi-immédiate des résultats obtenus, le fossé n'est pas toujours aussi important qu'il y paraît. Il a été signalé dans un rapport de Christian Dutheil [Duth91], que l'équipe de François Marcotorchino était la seule équipe française à avoir appliqué sa propre méthode d'analyse des données dans le domaine de la bibliométrie. Ceci nous permet de bénéficier d'un développement et d'une avancée plus rapide des recherches.

#### 1.4 Conclusion

Pour terminer, j'aimerais rappeler quelques points sur lesquels ces travaux ont apporté des avancées dans le domaine de l'analyse automatique des bases de données.

Les exemples présentés traitent principalement d'analyses effectuées sur la base de données brevets Derwent. Ceci tient essentiellement au fait que pour la veille technologique, les experts considèrent qu'il s'agit de la base de données la plus complète sur les brevets, la mieux organisée également, et celle qui possède le meilleur niveau de maintenance. D'autre part, parmi les différentes sources de données nécessaires en veille technologique, l'information brevet occupe une place primordiale. Son analyse automatique est donc un objectif important en bibliométrie. Il est évident que la méthode de classification relationnelle permet de travailler sur tout type de base de données. Le problème qui se pose alors est celui de la constitution des matrices d'informations à traiter.

L'un des résultats les plus importants de cette thèse concerne l'analyse automatique des bases d'informations. L'un des principaux problèmes des analyses multivariées était le petit nombre de références analysables simultanément, l'autre était lié à l'interprétation des résultats de l'analyse. Avec l'**Analyse** Relationnelle, non seulement il n'y a plus de limitation du nombre de références analysables, mais l'interprétation ne nécessite aucune conjecture sur le résultat. En outre il est

possible de faire porter l'analyse sur des informations présentes à des seuils extrêmement variables. La méthode d'analyse ne pré-suppose pas de suppression des informations très rares ou très fréquentes.

C'est l'ensemble des constatations faites précédemment qui a permis d'étudier des champs que les spécialistes de la bibliométrie avaient négligés jusque-là. Il en est ainsi pour le champ CIB qui contient les codes de la classification internationale et qui n'avait encore jamais été étudié.

A la question posée en début de cette introduction: l'analyse relationnelle est-elle une méthode opérationnelle pour la veille technologique?, je répondrais, pour conclure, qu'elle permet de résoudre les problèmes d'analyse automatique sur tout type de bases de données. Seuls les problèmes liés à l'étude des textes libres (cas des résumés dans les références bibliographiques) n'ont pas été abordés dans cette présentation. Des travaux ont débuté, visant à mettre en place une procédure d'analyse automatique. Cette nouvelle approche conjugue des outils de traitement du langage naturel avec les méthodes relationnelles de synthèse de l'information.

Munis de ces nouveaux outils, nous pensons que, les industriels français seront mieux à même de comprendre et d'anticiper les mutations scientifiques, technologiques, économiques et sociales du monde dans lequel ils évoluent.



# *Première Partie*

## **La veille technologique**

*Charles Huot et Chantai Bédécarrax*

Publication Scientifique et Techniques  
d'IBM France, n°4, 1992



# La veille technologique

Charles Huot et Chantal Bédécarrax

Centre Européen Scientifique de Mathématiques Appliquées

CESMAP IBM

68/76, Quai de la Rapée

75592 PARIS CEDEX 12

## 2.1 Introduction

*“On peut pardonner à **quelqu’un d’avoir**  
été battu pas **d’avoir** été surpris.”*

*Frédéric Le Grand*

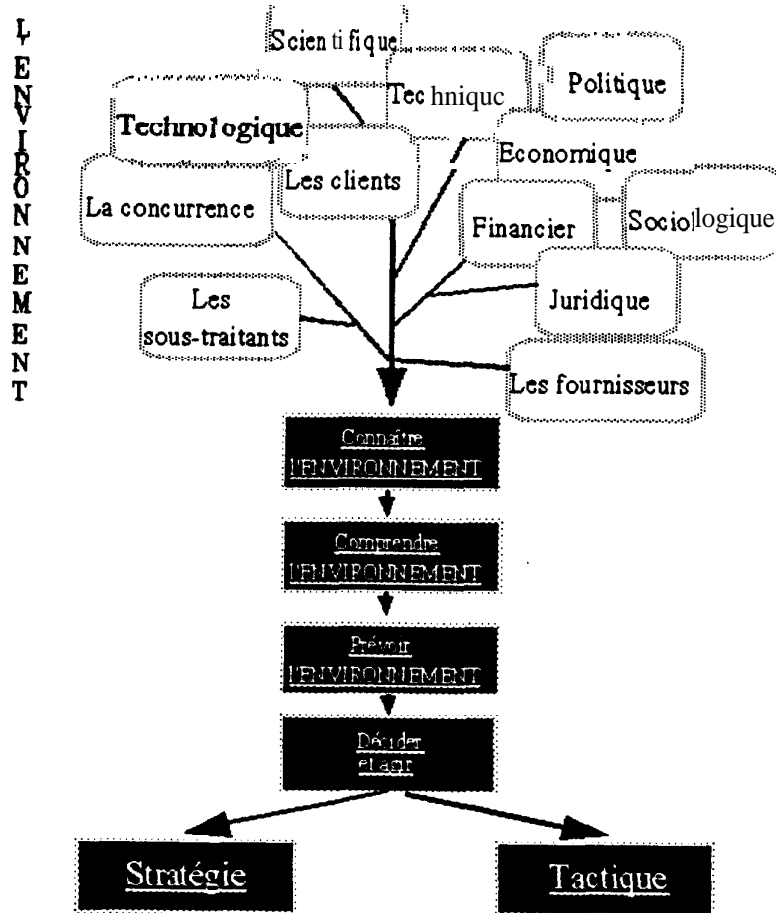
Le nom de *veille technologique*, qui depuis quelques années a fait son entrée en France dans les médias, cache une réalité complexe qui recouvre pour les entreprises *l’intelligence de leur environnement*. Cet environnement est **difficile** à appréhender dans sa globalité et ses multiples spécificités. Car si le terme de veille technologique semble renvoyer à la surveillance des technologies, il est compris en réalité le plus souvent comme synonyme d’une activité en marge de la légalité qui a pour but de tout surveiller, de tout analyser, afin d’informer les mythiques décideurs des entreprises des moindres faits et gestes de leurs concurrents. L’objectif de cet article est d’expliquer quelle est la position de la veille technologique aujourd’hui et comment l’une des composantes de la veille a pris le pas sur les autres pour la caractériser dans sa globalité.

Le monde économique et industriel a connu depuis 1945 une nette évolution. Chaque étape a été marquée par une modification de la répartition des ressources. Le tableau qui suit en récapitule les grandes lignes.

Tableau 1.		
Grandes tendances	Année	Management des ressources
Productivité	1945	De production
Contrôle de gestion	1955	Financières
Marketing	1965	Commerciales
Relations humaines	<b>1975</b>	Humaines
Technologie	1985	Technologiques

Nous voyons poindre dès 1985 un intérêt nouveau pour la technologie et depuis 1987, la veille technologique connaît un engouement extraordinaire, mais il est clair qu'un système de veille ne devient véritablement stratégique que s'il parvient à intégrer tous les types d'informations utiles pour l'entreprise. L'environnement de celle-ci comporte des aspects scientifiques, technologiques, technico-économiques, juridiques mais aussi des informations touchant à la clientèle, les fournisseurs, les concurrents, **sans** oublier le domaine social et politique, comme le synthétise **J. Villain [Vill90]** dans son image évocatrice du bonzaï de la surveillance et de l'intelligence de l'environnement.

Le "Bonzai" de la surveillance et de l' "intelligence" de l'ENVIRONNEMENT



source: "l'Entreprise aux aguets" J.Villain ed.MASSON

On rejoint la vision générale de la veille, technologique, concurrentielle, commerciale, globale, telle qu'elle est qualifiée par B. Martinet et J.M. Ribault **[Mart89]**, spécialistes de recherche industrielle et d'innovation chez les Ciments français et à la Cegos.

Dans le même esprit, H. Dou, directeur du Centre de Recherche Rétrospective de Marseille, nous rappelle que la veille n'est pas uniquement technologique, elle est aussi concurrentielle et commerciale. Le fait d'étudier et d'analyser toutes les informations qui

circulent à l'intérieur comme à l'extérieur de l'entreprise s'inscrit dans un processus général de veille stratégique [Dou92].

Ce processus de veille est naturellement indissociable des concepts d'information, de données, d'analyse et de synthèse. La chaîne qui permet de passer d'une information brute, disponible en surabondance, à une information élaborée de grande qualité, comprend le recueil des données, leur tri, leur évaluation et enfin leur synthèse et leur diffusion.

Nous essayerons de montrer, en proposant un aperçu des travaux les plus récents dans ce domaine, comment *“la véritable puissance du renseignement ne se situe pas dans la collecte de l'information mais dans les méthodes de traitement de l'information. Les résultats attendus devant aboutir au soutien et à l'élaboration d'une tactique et d'une stratégie en affaire ou en politique”*[Meye90].

## 2.2 Pourquoi la Veille Technologique?

*“La seule chose au monde qui coûte plus cher que l'information est l'ignorance des hommes.”*

*John F. Kennedy, 1962*

S'il est clair qu'aujourd'hui plus que jamais l'information est le sang de l'entreprise, d'où vient l'engouement pour la veille technologique?

Il nous arrive d'Asie et en particulier du Japon qui a su, depuis la fin de la deuxième guerre, passer de l'état de copieur à celui d'innovateur grâce au recours systématique et quasiment institutionnalisé à la veille technologique.

Les experts japonais estiment que le montant des dépenses de leurs industriels en matière d'information et de surveillance représente 1,5 % de leur chiffre d'affaires.

Selon Maurice Reyne [Reyn90] les industriels français ont depuis trop longtemps fait prévaloir une approche financière de l'entreprise. Or la vraie richesse d'un pays réside dans la transformation des matières premières par des technologies propres aux industriels. Le Japon, leader technologique mondial, en a fait l'heureuse expérience.

On doit considérer aujourd'hui que l'information est le 3<sup>ème</sup> facteur de production, après la main d'œuvre et le capital. En 1987, Jacques Delors déclarait dans le journal Le Monde

que «*Aucun retard industriel n'est jamais définitif,.... L'information est la clé de l'élaboration des stratégies*».

Ces considérations ne suffisent pas. Encore faut-il se donner les moyens d'exploiter cette information. En ce sens, la problématique de la veille stratégique consiste à répondre aux questions:

- que faut-il observer ?
- où aller chercher les informations formelles ou informelles ?
- comment les analyser ?

### 2.3 La mise en place d'une structure de veille

*" Toute décision provient de la  
conjonction d'une compétence et d'une  
information".*

**F. Bloch- Lainé**

Il est difficile de résumer en quelques **lignes** la portée d'un concept tel que celui de veille **technologique**<sup>9</sup>, nous proposerons toutefois une définition globale, que l'on rencontre assez couramment dans la littérature sur le sujet :

*L'observation et l'analyse -de révolution scientifique, technique, technologique. et des impacts économiques actuels ou potentiels correspondants pour dégager les menaces ou les opportunités de développement d'une société, soucieuse d'agir en tenant compte de son environnement.*

F. Jakobiak [Jako90], responsable de la veille technologique chez Atochem, suggère quant à lui :

*«L'observation et l'analyse de l'environnement suivies de la diffusion bien ciblée des informations sélectionnées et traitées utiles à la prise de décision stratégique.»*

Quelle que soit la formulation choisie, on comprend bien que les décideurs doivent ajouter à leur compétence personnelle l'information qui la complète. Dans ce contexte, la

---

<sup>9</sup> La veille, au sens large, porte le nom de *competitive intelligence* aux USA ou celui de *technological environment monitoring* en Grande-Bretagne.

veille technologique se présente comme un outil d'aide à la prise de décision stratégique, qui combine les fondements d'une doctrine, les ressources d'une méthodologie et la réalité d'une structure opérationnelle.

### 2.3.1 Une Doctrine

*“J’innove donc je veille”*

*François Jakobiak*

La lecture du rapport sur la veille technologique, établi par F. Jakobiak dans le cadre du **X<sup>ème</sup>** plan, apporte un éclairage précis sur le rôle de la veille dans le cadre du développement industriel et du maintien d'une bonne compétitivité [**Jako89**].

L'entreprise ne peut se contenter de produire et de vivre sur ses acquis. Elle doit innover pour éviter de disparaître ou d'être absorbée.

En permettant de surveiller les tendances, de déceler les indices de changement, de deviner les synergies possibles, d'anticiper les mutations, la veille technologique est vitale aux processus d'innovation des entreprises. Car l'innovation repose d'abord sur ce simple principe: *savoir ce que font les autres*.

Une politique de propriété industrielle active et bien adaptée est également fondamentale au maintien du dynamisme innovateur et à la pérennité de la **chaîne, Recherche-Innovation-Développement**, souvent désignée synthétiquement par R&D., qui conduit à la mise sur le marché d'un produit final, à travers **différentes** étapes :

- chercher, trouver, améliorer
- déposer des brevets ou acheter des licences si la recherche ne suffit pas à alimenter le maillon innovation
- vendre des licences si cela s'avère bénéfique pour la recherche ou si le maillon innovation ne peut pas l'utiliser, si le potentiel de recherche est, pour le moment, supérieur au potentiel de développement.
- acheter des unités de production, réaliser des *joint ventures* pour que le maillon production soit utilisé au maximum compte tenu des moyens et compétences.

### 2.3.2 Une Méthodologie

A tout seigneur tout honneur, les méthodes japonaises ont fait leurs preuves et constituent un bon modèle en matière de veille technologique. Une autre source d'inspiration peut venir de l'étude de la méthodologie du renseignement militaire, très formalisée et adaptable, dans une certaine mesure, au domaine industriel. M. Chalet, ancien directeur de la D.S.T, décrit parfaitement le système mis en place par le service de renseignement soviétique pour la récupération et l'analyse d'informations scientifiques et techniques en provenance des pays de l'ouest [Chal90].

D'une manière générale, on s'accorde à reconnaître trois étapes dans la définition des besoins en information:

#### **1. définir les facteurs critiques de succès**

Il s'agit de déterminer les objectifs poursuivis par la méthode des facteurs critiques de succès, étudiée au Massachusetts Institute of Technology, par J.F. Rockart [Rock79] et reprise par F.Jakobiak [Jako88].

Cette méthode consiste à déterminer les secteurs d'activité clé de l'entreprise, soit, selon J.F. Rockart, «*les quelques zones critiques où les choses doivent aller parfaitement pour que l'affaire soit florissante.*» Ces facteurs sont naturellement liés au domaine d'activité de l'entreprise, ainsi qu'à sa stratégie.

Souvent, en marge des buts officiels, existent des buts parallèles, non affichés, non proclamés, et recouvrant pourtant une considérable importance. C'est la connaissance des buts, dans leur intégralité, qui permet de déterminer les facteurs critiques de succès.

#### **2. élaborer un plan de recherche**

L'élaboration du plan de recherche revient à définir des orientations de recherche d'information pour savoir où trouver l'ensemble des renseignements permettant de réaliser un objectif. Un certain nombre d'outils permettent d'éclater l'objectif poursuivi en un nombre restreint d'axes de recherche de renseignements. Ces outils sont généralement issus des méthodes employées dans le domaine du marketing, ou de la prospective (check-lists, courbe de profils, courbe en S, méthodes matricielles, matrice de Delphi, etc).

#### **3. définir les indicateurs à surveiller**

Il faut ensuite **définir** un certain nombre d'indicateurs, que l'on va associer à chaque axe de recherche pour mieux le cerner, afin de suivre l'évolution des facteurs critiques de **succès**.

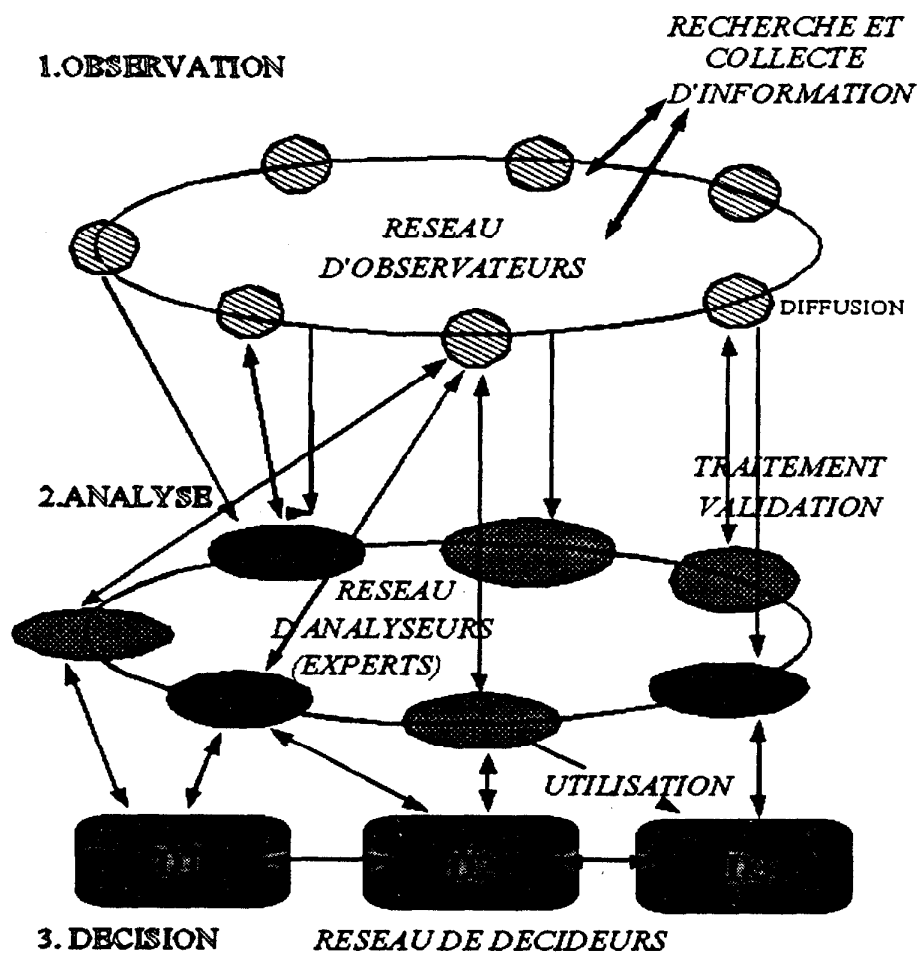
Par cette approche on **ramène** la veille technologique à une analyse spécifique d'un certain nombre de points-clés pour l'entreprise.

La mise en place et l'enchaînement de ces trois étapes impose naturellement une sensibilisation, une volonté et une implication constante des responsables. A la charge du veilleur technologique de jouer le rôle d'animateur et de vecteur de communication à tous les niveaux.

### 2.3.3 Une Structure

La structure préconisée par F. Jakobiak dans son ouvrage sur l'exploitation systématique des informations industrielles [**Jako90**] porte le nom de **4S**, pour Système de Surveillance Sectorielle Systématique. **Elle** repose sur quelques principes de base qui, s'ils sont simples, n'en sont pas moins fondamentaux.

Veille technologique, acteurs, fonctions et flux d'information



source: "Pratique de la Veille Technologique" F. Jakobiak, Les éditions d'Organisation

Il apparaît clairement que la centralisation est à bannir de la structure, car elle est aussi néfaste à la vie des informations qu'à l'efficacité d'un système de veille.

Il est donc nécessaire de répartir la structure sur un ensemble de pôles d'information spécifiques reliés en réseau

La mise en place de ces réseaux informatiques et de télécommunications est complexe lorsque l'entreprise est importante. De plus, pour que l'ensemble fonctionne bien il est indispensable de "préparer le terrain". Ce point a été souligné par plusieurs auteurs, et

notamment par M. Lasfargue dans “Technojolies, technofolies” [Lasf89]. Pour cela il faut créer des réseaux de complicités, composés de personnes convaincues de la nécessité de travailler ensemble. Ce sera le troisième réseau, celui des spécialistes.

Étudions plus en détail chacun de ces réseaux.

### **2.3.3.1 Les réseaux de spécialistes**

Ils seront au nombre de trois:

#### **1. le réseau des observateurs**

C'est celui qui sera en charge de la recherche, de la collecte et de la diffusion de l'information brute. Les informations qu'il va recueillir seront de nature industrielle (scientifique, technique, technologique, technico-économique) mais également réglementaire et juridique. Ces informations seront extraites en majorité des bases de données.

Une autre source sera constituée de journaux, périodiques, livres, ouvrages et documents divers, moins facilement exploitables mais indispensables toutefois pour une bonne couverture de l'information. Il en est de même pour l'étude des informations de nature informelle qui recouvre la recherche et la collecte de résultats d'enquêtes, de prospectus dans les foires, d'actes dans les colloques, etc. Une fois l'information recueillie, elle doit être structurée en bases de données pour, d'une part, la rendre facilement accessible et, d'autre part, pouvoir l'analyser dans sa globalité.

C'est au sein du service de veille que seront menées les études et analyses sur les informations recueillies par le réseau. Des rapports thématiques seront ensuite réalisés et envoyés aux experts des domaines concernés pour validation.

#### **2. le réseau des experts**

On peut imaginer répartir ce réseau en quatre grandes familles, s'occupant respectivement des produits, procédés, applications et prospective. Ce réseau est constitué, suivant la famille à laquelle il appartient, de producteurs, de chercheurs, de marketing, de commerciaux, de décideurs plan ou de décideurs R&D. Les experts qui composent ce réseau ne sont pas investis “à plein temps” dans cette activité; ils effectuent des interventions ponctuelles qui viennent en complément de leur activité principale.

Ce réseau est le pivot de la structure, chargé de transmettre au réseau des décideurs l'information analysée et validée.

#### **3. le réseau des décideurs**

Sans véritable structure, compte tenu du peu de temps dont ils disposent, ils élaborent les nouvelles stratégies de l'entreprise au cours de réunions de direction. L'information analysée et les suggestions des experts composent une base solide de discussion.

### 2.3.3.2 Le **réseau des Télécommunications**

C'est à travers lui et ses différents supports (téléphone, télex, messagerie électronique, télécopie, minitel, etc.) que va transiter l'essentiel des informations entre les diverses entités qui constituent la structure de veille.

### 2.3.3.3 Les **réseaux Informatiques**

Il faut employer le pluriel car on recense :

- **le réseau télématique** qui est l'outil indispensable des observateurs spécialistes en information documentaire
- **le réseau sur site** qui relie entre eux les micro-ordinateurs et les différentes plate-formes
- **le réseau de type longue distance** qui nécessite le recours à un gros ordinateur

C'est à travers une connexion sur des lignes PTT à haut débit, par l'utilisation d'un logiciel sur micro-ordinateur que l'on peut accéder à un centre serveur et interroger des bases et banques de données de tous types. La majorité de l'information viendra des bases de données. L'un des avantages, outre la masse d'information disponible, est lié au fait que l'on peut recueillir sur son micro-ordinateur le résultat de sa consultation sous forme des fichiers électroniques. Ces fichiers constituent la base de travail pour des analyses de fond.

## 2.4 Les Outils de la Veille Technologique

### 2.4.1 Les Bases et les banques de données

Depuis la fin des années soixante, l'accès à l'information a beaucoup évolué avec l'avènement des bases et banques de données disponibles à partir d'un ordinateur via un **modem**<sup>10</sup> et le réseau Transpac<sup>11</sup>. Ce procédé constitue aujourd'hui un outil performant et très puissant de collecte de données et de recherche d'information.

---

<sup>10</sup> Le MODEM (**MOD**ulateur **DE**Modulateur) est un appareil qui transforme un signal transmis par le réseau du téléphone en un signal compréhensible par l'ordinateur.

<sup>11</sup> Réseau **PTT** spécialement conçu pour la transmission d'information à haut débit

Les bases de données accessibles en ligne couvrent l'essentiel des domaines susceptibles d'intéresser les entreprises. Qu'il s'agisse de documents scientifiques, techniques, technologiques, économiques ou juridiques, d'articles de presse ou encore de descriptifs d'organisations commerciales, il est possible d'obtenir des références bibliographiques sur toutes sortes de publications.

Mais pour optimiser l'utilisation de ces mines d'informations, il importe de savoir où chercher.

Le monde des bases de données en ligne est structuré sous forme pyramidale. A la base, on trouve les entrées individuelles de chaque base de données, appelées **références**. Au niveau suivant on accède aux bases elles-mêmes. Et au sommet se trouvent les organismes qui distribuent ces bases, appelés serveurs.

A titre d'information, nous mentionnerons quelques bases, particulièrement utiles dans le cadre de la veille technologique.

- WPI et WPIL (World Patent International et WPI **Latest**) distribuées par **Derwent** Incorporation, qui fournissent des références bibliographiques de brevets, riches en informations de nature technologique autant que concurrentielle.
- dans le domaine de la chimie, gros producteur et consommateur de données, la base **CHEMICAL ABSTRACTS** est très utilisée
- **INSPEC** pour la physique et la technologie
- **MEDLINE** pour tout ce qui concerne la médecine
- **ESSOR** pour des informations sur les entreprises
- **ECLATX** pour les classifications internationales

Côté serveur, nous pouvons citer Télésystèmes Questel, **Orbit**, Dialog ou encore ESA le serveur de l'Agence Spatiale Européenne.

*Les brevets constituent une source d'information tout à fait privilégiée dans le cadre de la veille technologique, technico-économique et concurrentielle. Ils offrent en effet une quantité considérable de références accessibles (pour information, la base WPIL contient entre 4 et 5 millions de références avec une croissance annuelle de l'ordre de 350.000 références), une couverture très étendue tant sectorielle que temporelle et enfin une variété d'informations qui va au-delà du seul aspect technologique.*

*A titre d'exemple, voici un extrait de référence de brevet avec la définition des champs principaux :*

<b>-1-</b>	
<b>AN -89-354640/48</b>	Numéro d'accès dans la base
<b>TI -Wall panel- has slits joining inter-pane space to interior, made in top of window frame, and slit joining it to ventilation cavity</b>	Titre du brevet
<b>IC -A61L-002/18 C11D</b>	Codes de la Classification Internationale
<b>DC -Q44 Q48</b>	Codes Derwent
<b>PA -(EVEN/) EVENTOV V S</b>	Société déposante
<b>IN -EVENTOV VS</b>	Inventeur
<b>PR -86.0508 86SU-075778</b>	Numéro de priorité (avec date et pays)
<b>DS -CH DE FR GB IT LI NLSU SE</b>	Pays d'extensions

Si les services en ligne sont parfois d'un coût assez élevé, il ne faut pas perdre de vue le fait qu'ils présentent des avantages incomparables, tels que la rapidité de recherche, la disponibilité et la fiabilité des données. Préciserons, en outre, que les connexions aux serveurs ainsi que les prix des références de base sont en fait extrêmement variables.

Le principe d'utilisation de ces bases de données est simple. L'utilisateur formule une requête composée de mots et d'opérateurs booléens et le serveur renvoie le nombre de références bibliographiques, contenues dans la base, qui répondent à son interrogation. L'utilisateur peut ensuite récupérer sur le disque dur de son ordinateur ces notices bibliographiques. Cette opération, appelée téléchargement, est le point de départ de tout processus d'analyse et de synthèse d'informations par des techniques relevant de la bibliométrie, de la scientométrie ou encore de l'infométrie, dont nous allons parler maintenant.

### 2.4.2 Bibliométrie, scientométrie, infométrie

Un certain nombre de disciplines ont vu le jour, pour développer des méthodes et des outils d'analyse des références téléchargées ou, plus généralement, des informations de nature documentaire.

W. Turner, chef du département Recherche et Produits Nouveaux de l'INIST/CNRS, propose les définitions suivantes [Turn89]:

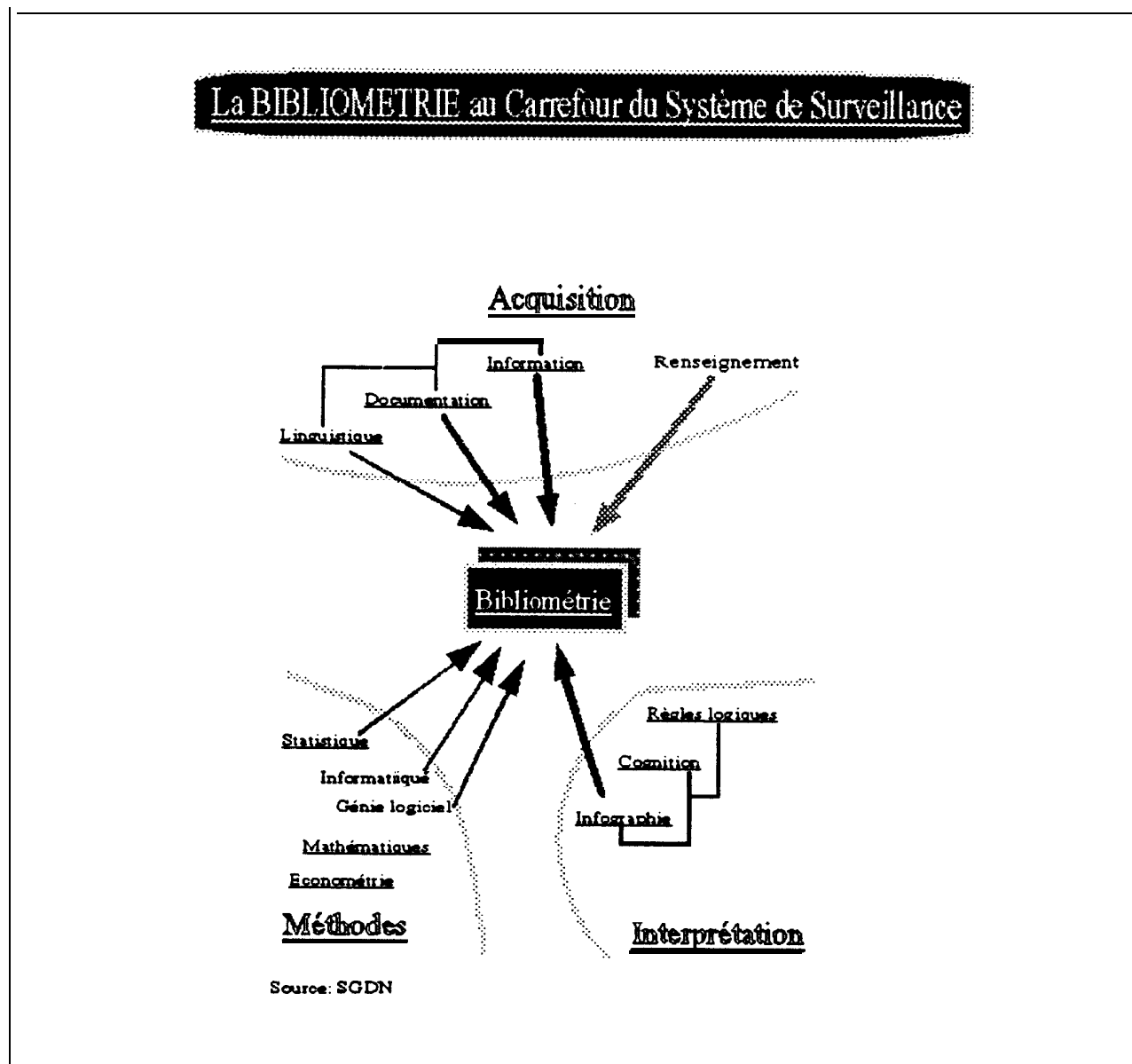
1. **La bibliométrie** a connu un essor considérable à la fin des années soixante lorsque la question de mesurer *l'output* du système de la recherche a été posée dans une conjoncture caractérisée par la convergence de deux phénomènes importants: l'informatisation des banques de données permettant le développement d'indicateurs; une demande de la part des administrateurs de la recherche d'informations chiffrées permettant d'évaluer **l'efficacité** de leurs politiques.
2. **La scientométrie vise**, quant à elle, à établir une relation entre les résultats produits et les ressources accordées à un système de recherche. L'analyse est globale, car le fait de dresser le bilan de santé d'un système de recherche sans tenir compte de ses ressources et la façon dont ces ressources sont transformées en publications marque les limites d'une démarche strictement bibliométrique. La scientométrie pose la question de la transformation des *inputs* en *outputs*. De ce point de vue, la bibliométrie peut **être** considérée comme un domaine de recherche spécifique à l'intérieur de la scientométrie.
3. **L'infométrie se** place à un degré d'agrégation encore plus élevé. Alors que la scientométrie considère le système de la recherche comme un système clos, l'infométrie pose le problème des relations que ce système entretient avec d'autres domaines d'activité économique et sociale. Parmi ses objectifs figure celui visant à cerner le rôle que peut jouer l'information dans des processus d'innovation scientifique et technique.

Si les points de **différenciation** entre ces techniques présentent un intérêt certain sur le plan épistémologique, les variantes sont vite gommées lorsqu'on se place dans le contexte industriel de la veille technologique où seuls importent les résultats et leur valeur ajoutée.

### 2.4.3 Analyse des informations

La quantité d'informations disponibles sous forme électronique ne cesse de croître et ce flot constant rend totalement impossible la consultation détaillée de tous les documents concernant un domaine de recherche ou un secteur d'activité industrielle donnés, aussi limités soient-ils.

Ces constatations justifient le développement rapide des disciplines au service de la veille technologique. Par souci de clarté, nous conviendrons de retenir le seul nom de *bibliométrie* pour désigner de façon générique l'ensemble de ces disciplines.



Comme le montre ce graphique, la **bibliométrie** constitue, en quelque sorte, le ciment des trois phases de gestion de l'information. Nous avons abordé précédemment les étapes

d'acquisition et d'interprétation, intéressons-nous maintenant aux méthodes de traitement aujourdhui disponibles.

Dans leur diversité, elles ont pour vocation commune de condenser sous forme quantitative **et/ou** qualitative une série de références bibliographiques. Ces condensés, qui peuvent prendre diverses formes, constituent un support directement exploitable ou permettent de mettre au point des grilles de lecture restreinte des documents primaires, mieux adaptées au potentiel de lecture humain.

#### **2.4.3.1 les Méthodes de traitement existantes**

*“Les moments les plus heureux de  
l’histoire des connaissances surviennent  
lorsque des faits qui n’avaient été  
jusqu’alors que des données particulières  
sont soudain mis en rapport avec d’autres  
faits apparemment éloignés et apparaissent  
ainsi dans une nouvelle lumière.”*

*Wolfgang Kohler*

Les méthodes d'analyse de l'information, à proprement parler, sont essentiellement de nature statistique. Il importe toutefois de signaler que leur utilisation gagne à être couplée à des ressources développées dans d'autres disciplines. Ainsi, des outils de traitement du langage naturel rendent possible l'exploitation directe des références ou des textes sources, souvent porteurs de l'information dans son intégralité [Warn91]. A l'autre bout de la chaîne, des outils d'infographie ou des méthodes dérivées des sciences cognitives facilitent grandement la mise en forme et la gestion des résultats finaux. Nous ne détaillerons pas ici ces différents points. Nous renvoyons le lecteur à la littérature fort abondante sur ces sujets.

On constate, aujourd'hui, que les techniques de traitement en profondeur de l'information n'ont pas véritablement fait leur entrée dans le monde industriel, au mieux elles y sont étudiées à titre expérimental. Elles restent encore le domaine privilégié des équipes de recherche des universités ou des organismes d'état, au sein desquelles elles font toujours l'objet de recherches actives, comme le rappelle C. Dutheil [Duth91]. Le recours aux statistiques usuelles est, en revanche, beaucoup plus répandu.

#### 2.4.3.1.1 Les statistiques usuelles

Par statistiques usuelles, il faut entendre, essentiellement, opérations de dénombrement et utilisation d'indicateurs synthétiques.

Les serveurs de bases de données offrent, pour la plupart, un service de statistiques en ligne qui permet, avant même le téléchargement des données, d'effectuer des comptages et de sortir des listes d'items sélectionnés, déjà porteurs d'informations.

Des analyses statistiques effectuées sur des **références** de brevets, à l'aide de *Patstat*<sup>12</sup> de la société Derwent, sont détaillées dans les travaux d'une équipe de l'Institut Français du Pétrole [Mour90]. L'essentiel des traitements consiste à générer des listes par **fréquences** décroissantes du nombre de brevets en fonction des sociétés déposantes, ou des domaines d'activité ou encore des années de dépôts.

Le recours à des progiciels tableurs, liés aux systèmes de gestion des bases de données, offre une plus grande latitude d'analyse. Il permet d'étudier des distributions un peu plus en détail, de déterminer des variables caractéristiques de certains phénomènes et de rechercher des corrélations entre ces variables.

Ce type de démarche constitue le premier pas, dans ce que l'on peut appeler la recherche de la structure sous-jacente de l'information. Dans cette perspective, des organismes comme l'Observatoire des Sciences et des Techniques proposent des indicateurs globaux sur l'évolution des **différentes** technologies ou des recherches fondamentales [Barr91].

Le standard de ce genre d'analyse consiste à regarder l'évolution du nombre de **dépôts** de brevets au cours du temps dans **différents** pays. Ce traitement, qui croise les champs («année» et «dép& **prioritaire**»), permet de positionner un secteur d'activité dans le contexte général de la course à la protection des inventions. Il donne également un bon aperçu de la politique des grands groupes industriels des divers pays analysés,

L'étude de la courbe du nombre de **dépôts** de brevets en fonction du temps apporte, quant à elle, une indication intéressante sur l'état de développement d'une technologie. Elle permet d'évaluer la maturité d'un domaine, d'en mesurer l'expansion ou au contraire le déclin.

Ainsi, suivant les questions que se posent les dirigeants d'entreprises ou les directeurs de laboratoires de recherche, ils ont à leur disposition des outils simples qui permettent de **raffiner** des données brutes à un niveau plus ou moins grand.

---

<sup>12</sup> Patent **statistics**

#### 2.4.3.19 Les méthodes d'analyse des données

Les traitements de base que nous venons d'évoquer, s'ils permettent d'éclaircir quelque peu la "nébuleuse" des données brutes, s'avèrent rapidement assez limités dès qu'il s'agit de dégager véritablement les informations que **recèlent** les données, de mettre en évidence des interactions de phénomènes, de fournir une évaluation qualitative en complément des mesures quantitatives. Il convient alors de faire appel **à** des techniques plus sophistiquées d'analyse des données.

L'analyse des données, branche de la statistique, couvre, à travers son arsenal de méthodes, une très vaste gamme d'applications [**Marc91b**], parmi lesquelles la bibliométrie a trouvé une place de choix depuis quelques années.

Les données que l'on est amené à manipuler dans ce domaine possèdent des caractéristiques tout à fait spécifiques qui imposent d'ajuster les techniques existantes, voire de mettre au point de toutes **nouvelles** approches. L'apport de ces méthodes, en veille technologique, fait donc l'objet d'un grand nombre de travaux scientifiques tant en France qu'à l'étranger.

Aux Etats-Unis, l'**Institute** for **Scientific** Information à Philadelphie, fait figure de pionnier dans le domaine à travers l'analyse des **cocitations**, qui a pour objectif de dégager des réseaux de collaborations et d'extraire de la masse les publications les plus importantes.

En Allemagne, K. Brockhoff, directeur de l'**Institute** of Research in Innovation Management, base ses analyses brevets sur des méthodes de théorie des graphes [**Broc92**]. Que ce soit pour construire des indicateurs synthétiques d'activité dans différents domaines industriels, pour mesurer les connexions **brevets/produits finaux** ou encore pour évaluer la créativité des inventeurs, il prône l'exploitation systématique des informations contenues dans les **références** de base.

En France, le professeur H. Dou et son équipe du laboratoire de Recherche Rétrospective de Marseille [**Dou90d**] présentent les méthodes d'analyse des données comme des outils de veille technologique à la dimension des moyennes entreprises. Ils montrent comment utiliser des méthodes de classification ou des méthodes factorielles pour analyser des corpus documentaires constitués, par exemple, de publications scientifiques. Cette approche est aujourd'hui implantée dans un certain nombre d'entreprises et de centres techniques français.

Dans une démarche similaire, des scientifiques de l'**ITODYS**, **IRIT/CIT**, et de l'**IRIES** se sont associés pour une étude sur "Medline vue par l'analyse factorielle et la classification automatique". Malgré certains problèmes posés par le type de méthodes employées (taille

des données, interprétation des résultats) cette approche a permis de mettre en évidence des phénomènes impossible à détecter à la seule lecture des **références** du corpus.

Mentionnons **également** les travaux de l'équipe de W. Turner sur l'analyse des mots associés qui vise, d'une manière générale, à établir des réseaux d'associations entre objets bibliographiques, tels que les articles, les auteurs ou les sujets d'étude.

L'équipe de recherche du Centre Européen Scientifique de Mathématiques Appliquées **d'IBM** a, pour sa part, mis en place sa propre méthodologie d'analyse des informations téléchargées [**Bede91**]. Basée sur **l'Analyse** Relationnelle, méthode d'analyse des données qualitatives [**Marc78**], elle propose des outils dédiés à l'analyse des matrices creuses de très grande taille, fort courantes en bibliométrie, et conçus pour opérer sur des distributions spécifiques aux données rencontrées dans ce domaine.

L'objectif de cette méthodologie est de transformer les données élémentaires contenues dans les bases en information élaborée, concernant les points stratégiques de positionnement vis-à-vis de la concurrence, de détection de l'innovation ou dévaluation des axes de recherche. Elle vise, d'une **manière** générale, à apporter des réponses à des questions clés en veille technologique, telles que :

- la connaissance des environnements technologiques pour de nouveaux marchés
- la mise en évidence des stratégies d'extensions et de la couverture internationale de l'entreprise et de ses concurrents
- les typologies des domaines d'activité avec découpage en technologies phares, technologies innovantes, technologies chamières, pour l'Entreprise et ses concurrents
- la correspondance entre les classifications codifiées dans les bases externes et les classifications internes

Aujourd'hui, de la collaboration de différents acteurs impliqués dans ce domaine, est née la volonté d'intégrer l'essentiel de ces outils à des stations de travail spécifiquement dédiées à la veille technologique.

Ces stations rendront possible l'automatisation de la chaîne des traitements qui conduit de l'interrogation des bases de données à la génération de rapports de synthèse en passant par des analyses en profondeur de l'information documentaire.

## **2.5 Conclusion**

La veille technologique et plus généralement la veille stratégique s'intègrent tous les jours un peu plus dans les **systèmes** de prises de décisions. Les dirigeants d'entreprises sont maintenant convaincus de son utilité.

Nous avons essayé de présenter, dans cet article, le déroulement du processus qui conduit des données de base, aux véritables informations de nature stratégique.

L'un des maillons fondamentaux de cette chaîne, du moins pour ce qui concerne la collecte d'informations, est constitué par les bases de données accessibles en ligne. Avec un **chiffre** de 100 documents indexés toutes les 10 secondes, ces bases vont poursuivre leur développement et leur nombre ne va cesser d'augmenter dans les années **à** venir. Cette perspective risque de rendre encore plus crucial le recours **à** des méthodes automatiques d'analyse et de **synthèse** des informations. Aujourd'hui les industriels français prennent conscience de ce phénomène et la **bibliométrie, après** plus de 20 ans d'existence, trouve progressivement sa place dans la chaîne.

Si les maillons documentaires, **méthodologiques** et informatiques du processus de veille sont des points de passage obligés, ils n'ont véritablement d'existence **qu'à** travers des relais humains. On a coutume de dire que la veille technologique est avant tout une affaire d'état d'esprit, de motivation et d'implication de tous les collaborateurs d'une entreprise. Ce point reste incontestable. C'est au veilleur, avec le soutien de chacun de ses collègues, de jouer un **rôle** d'animateur, de **contrôler** le processus que nous avons décrit et de chercher en permanence à optimiser son fonctionnement.

# *Seconde Partie*

## **Application de l'Analyse Relationnelle à la veille technologique: des outils d'analyse pour l'information documentaire**

*Chantal Bédécarrax et Charles Huot*

Journées d'études de la SFBA, Ile Rousse,  
juin 1991



# Application de l'Analyse Relationnelle à la Veille Technologique: des outils d'analyse de l'information documentaire

Chantal BEDECARRAX , Charles HUOT  
Centre Européen de Mathématiques Appliqués IBM<sup>13</sup>

## 3.1 Introduction

**«Le TRAITEMENT *k* plus important pour l'exploitation systématique de l'information industrielle est incontestablement L'ANALYSE STATISTIQUE DES BREVETS».**

Cette phrase est extraite du dernier ouvrage de F. Jakobiak intitulé "Pratique de la Veille Technologique" [Jako90]. Elle exprime clairement l'importance des méthodes d'analyse des données dans le processus de la veille technologique. Déjà en 1988, L. Quoniam insiste sur la liaison entre l'information stratégique et les méthodes d'analyse des données [Quon88] *«Ce concept nouveau de mise à disposition de masse importante d'information va entraîner une modification radicale des méthodes de travail ouvrant des horizons nouveaux. Mais la clé de la réussite de cette mutation passera par la maîtrise de ces concepts et la faculté d'exploiter intelligemment cette masse qui ne sera rien d'autre qu'une mine de matière primaire.»*

Il ressort de ces citations que deux environnements sont en jeu; d'une part l'information et d'autre part les outils qui vont permettre de l'analyser en profondeur.

De nombreuses équipes, en France, travaillent sur l'application de méthodes d'analyse des données à la bibliométrie ou à la veille technologique. Nous pouvons citer l'équipe de H. Dou au C.R.R.M<sup>14</sup>[Dou89, Dou90a, Quon90b] , celles de W.A. Turner<sup>15</sup>[Turn89] , C. Paoli [Paol87], C. Dutheil [Duth90], ainsi que A. Girard et M. Moureau de l'IFP<sup>16</sup>[Gira88] .

---

<sup>13</sup> CEMAP IBM, 68-76 quai de la Rapée, 75592 Paris CEDEX 12, Tel: 40.01.57.11/ 40.01.53.37, Fax: 49.28.08.60

<sup>14</sup> Centre de Recherche Rétrospective de Marseille, Centre de St Jérôme, Université d'Aix-Marseille III

<sup>15</sup> INIST/CNRS

<sup>16</sup> Institut Français du Pétrole

L'équipe de recherche du CEMAP travaille depuis une quinzaine d'années sur les méthodologies classificatoires et notamment sur l'**Analyse** Relationnelle dont elle est à l'origine, par les travaux de F. Marcotorchino et P. Michaud [Marc78, Marc79]. Cette méthodologie a été exploitée avec succès par I. Wamesson [Bede89b, Warn90] dans le domaine de la lexicographie computationnelle pour la restructuration de dictionnaires, **c'est-à-dire** dans un cadre d'application en grandeur réelle.

Nous nous intéressons aujourd'hui à son utilisation dans le domaine de la veille technologique. L'objectif est de pallier les **défauts** des méthodes d'analyse plus connues, comme l'analyse factorielle des correspondances multiples ou les classifications hiérarchiques usuellement utilisées, qui s'adaptent mal au type de matrices générées par l'analyse des corpus de références. En effet, contrairement aux autres méthodes, l'analyse relationnelle permet, par ses classifications, de faire ressortir les grandes tendances autant que les particularités contenues dans l'information. **Elle** permet, qui plus est, de traiter des matrices de grande taille à faible densité informative comme celles que l'on est amené à extraire des bases de données brevets ou documentaires.

## **3.2 Présentation générale des données**

---

Cette phase préliminaire se décompose en deux étapes; d'abord l'extraction des **références** de la base de brevets, puis la génération des matrices qui seront soumises à la méthodologie classificatoire. C'est de cette deuxième partie que traite le présent article.

### **3.2.1 Extraction des références**

---

C'est la première étape à franchir avant la mise en œuvre de toute méthode d'analyse statistique des bases de données. Cette partie, très délicate, fait appel à une bonne connaissance des bases de données internationales pour isoler un corpus de références cohérent pour le problème posé. Les méthodes que nous utilisons prennent en compte des matrices qui sont fabriquées par la mise en relation des divers champs constitutifs des **références** retenues dans le corpus.

La génération des matrices a été effectuée par l'équipe du C.R.R.M de Marseille, à l'aide de logiciels prototypes spécialisés dans ce genre d'application. Ces traitements sophistiqués peuvent **naturellement** être appliqués à l'analyse de base de données internes à **l'entreprise**.

D'une manière générale, une **référence** de brevet se décompose en plusieurs champs, indexés par des codes, dédiés à une information spécifique. Citons par exemple certains champs utilisés par **Derwent**<sup>17</sup>:

- **AN**: Numéro du brevet dans la base
- **TI**: Titre du brevet
- **DC**: Codes Derwent
- **PA**: Nom de la société déposante
- **IN**: Noms des inventeurs
- **PN**: Numéro du brevet
- **DS**: Pays d'extensions
- **CT**: Brevets ou travaux scientifiques cités lors du dépôt
- **IC**: Codes **CIB**
- **MC**: Codes manuels
- **AB**: Résumé

L'ensemble de ces informations, prises séparément ou en combinaison, peut faire l'objet d'un grand nombre de traitements statistiques et générer de nombreux tableaux pour des analyses de données.

Nous nous sommes intéressés, pour notre part, à l'exploitation des relations entre les champs **nom de société**, **numéro de brevet** et codes CIB **décrivant le brevet**<sup>18</sup>.

### 3.2.2 Présentation relationnelle

A partir de ces champs, nous avons **défini** trois ensembles d'objets sur lesquels vont porter les traitements :

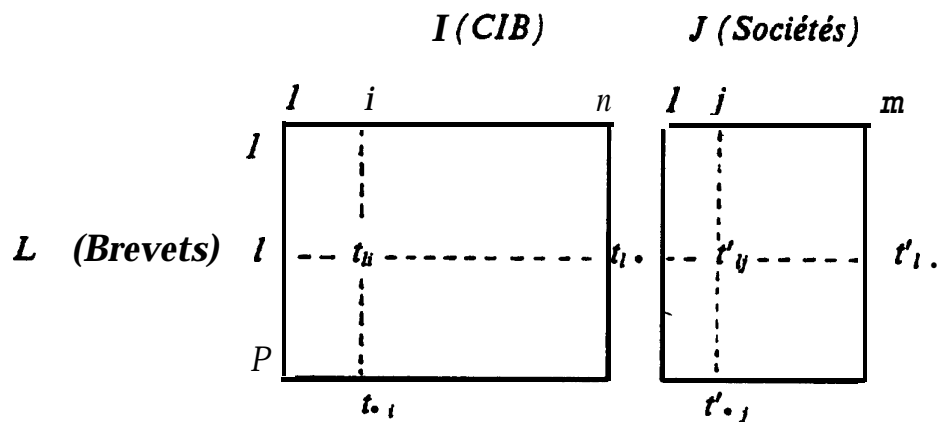
- $I$  = ensemble des codes CIB , avec  $\text{Card}(I) = n$   
La valeur de  $n$  peut varier en fonction du nombre de digits retenus **c'est-à-dire** de la précision des codes CIB.
- $J$  = ensemble des noms des sociétés déposantes , avec  $\text{Card}(J) = m$
- $L$  = ensemble des numéros des brevets, avec  $\text{Card}(L) = p$

---

<sup>17</sup> Producteur de bases de **données** (WPI, WPIL,...).

<sup>18</sup> La **Classification** Internationale des Brevets recouvre l'ensemble des **connaissances** que l'on peut **considérer** comme relevant du domaine des brevets d'invention. **Elle décrit** les brevets à l'aide de codes **hiérarchisés** en **sections, classes, sous-classes, groupes** et **sous-groupes**. **Cette organisation structurée** permet d'exploiter **les** codes CIB **à des** niveaux **différents** : du plus large (**les** sections : 1 digit) correspondant à 8 descripteurs, au plus **fin** (les sous-groupes : 11 digits) représentant quelque **700 000** descripteur&

Les données sont extraites des descripteurs des  $p$  brevets retenus. On peut schématiser l'aide de deux tableaux,  $T$  et  $T'$ , les différentes relations contenues dans les **références**.



### Caractéristiques des tableaux binaires $T$ et $T'$

Le tableau  $T$  croise les brevets avec les codes CIB qui les décrivent et le tableau  $T'$  croise les brevets avec les noms des sociétés qui les déposent. Ces deux tableaux sont la simple restitution, sous forme relationnelle, des informations extraites des **références** à partir des trois champs retenus pour **l'étude**.

Leurs termes **généraux** se définissent respectivement de la façon suivante :

$$t_{li} = \begin{cases} 1 & \text{si le brevet } l \text{ est décrit par le code } i \\ 0 & \text{sinon} \end{cases}$$

$$t'_{lj} = \begin{cases} 1 & \text{si le brevet } l \text{ est } \mathbf{d\acute{e}pos\acute{e}} \text{ par la soci\acute{e}t\acute{e} } j \\ 0 & \text{sinon} \end{cases}$$

- Sommes **en** lignes

$$t_{l\bullet} = \sum_i t_{li} = \text{nombre de codes décrivant le brevet } l$$

$$t'_{l\bullet} = \sum_j t'_{lj} = \text{nombre de soci\acute{e}t\acute{e}s ayant d\acute{e}pos\acute{e} le brevet } l^{19}$$

- Sommes **en** colonne

$$t_{\bullet i} = \sum_l t_{li} = \text{nombre de brevets décrits par le code } i$$

$$t'_{\bullet j} = \sum_l t'_{lj} = \text{nombre de brevets } \mathbf{d\acute{e}pos\acute{e}s} \text{ par la } \mathbf{soci\acute{e}t\acute{e} } j$$

<sup>19</sup> Il est très rare qu'un même brevet soit déposé par plus d'une société, c'est pourquoi dans notre application nous aurons systématiquement  $t'_{l\bullet} = 1$

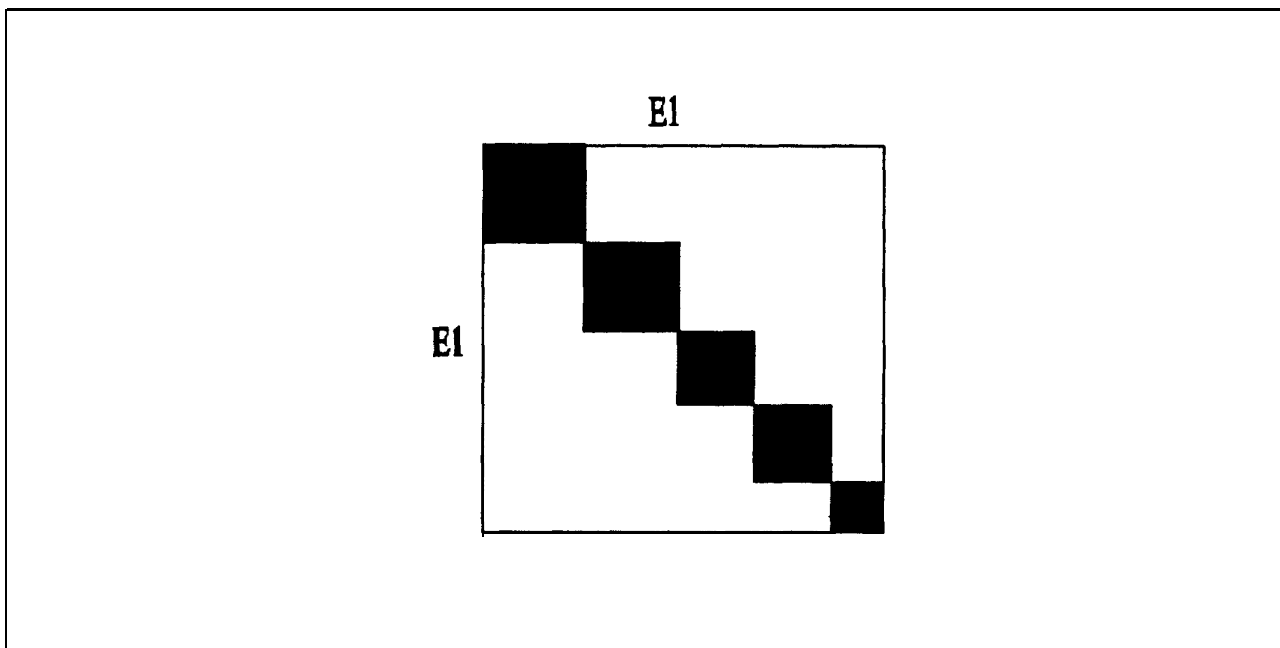
### 3.3 Problèmes traités

**Nous** avons utilisé des méthodes de classification automatique développées dans le cadre de l'**Analyse Relationnelle** pour faire émerger des informations structurées à partir des données séquentielles extraites de la base brevets.

Avant de détailler les problèmes que nous avons traités, il nous paraît important de rappeler brièvement la "philosophie" des deux procédures que nous avons mises en œuvre, classification et sériation, ainsi que le type de résultats auxquels elles conduisent.

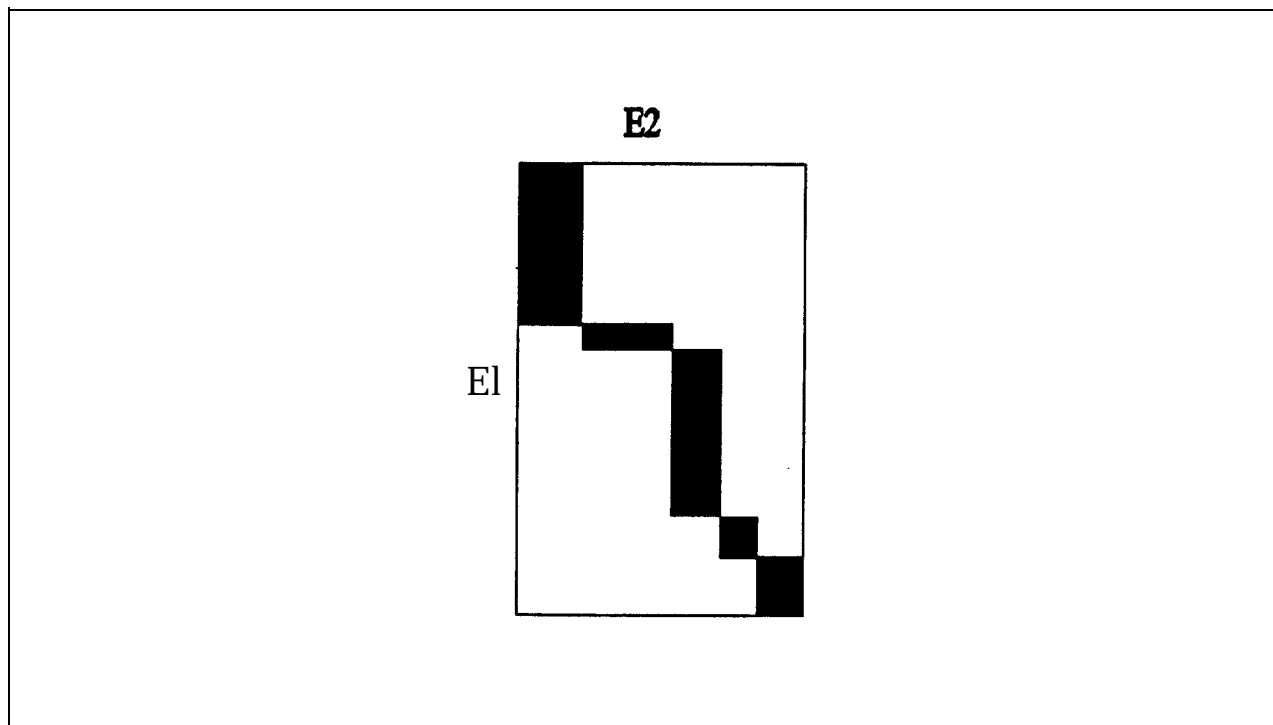
La méthode de classification travaille sur les relations qu'entretiennent les objets à l'intérieur d'un même ensemble. Elle a pour but de regrouper, dans des classes homogènes, les objets les plus ressemblants tout en séparant bien les différentes classes. Elle garantit donc de fortes densités de relations à l'intérieur des classes et de faibles densités à l'extérieur.

La relation structurée que l'on cherche à construire est une relation d'équivalence sur l'ensemble des objets, autrement dit une partition de cet ensemble, dont la forme schématique est la suivante:

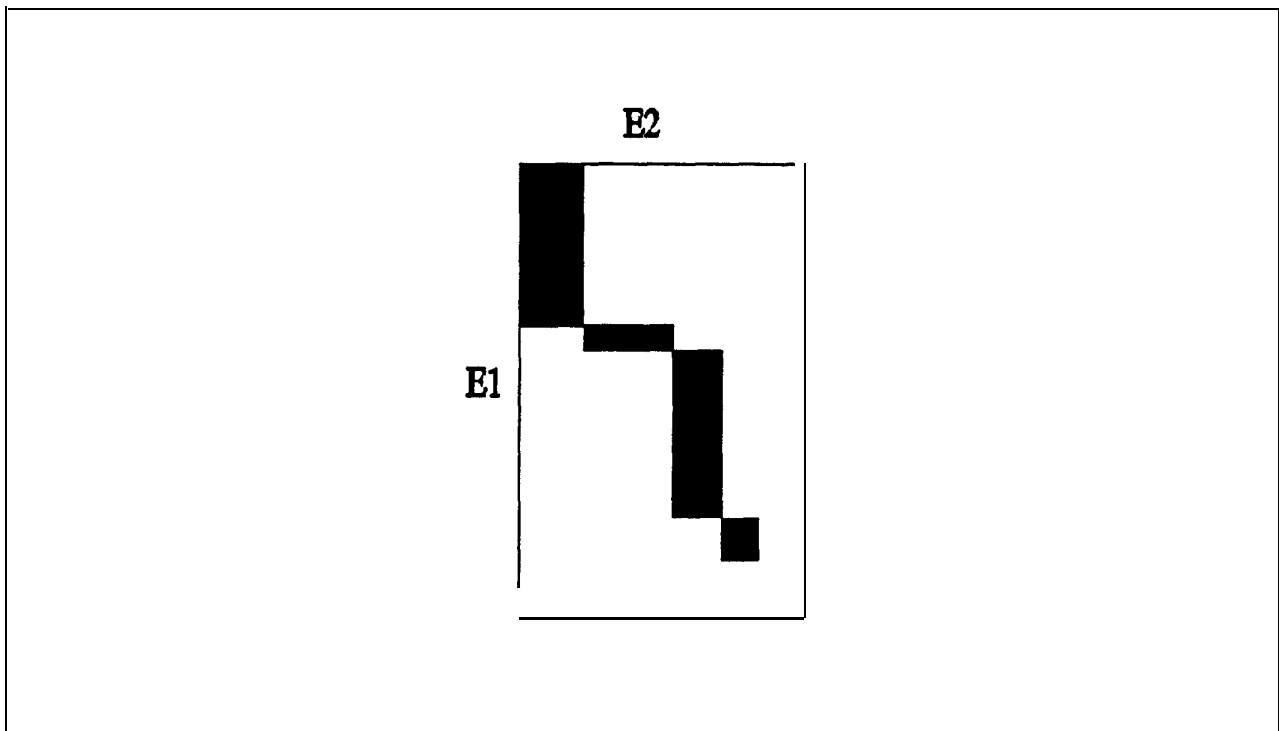


La sériation, pour sa part, opère sur deux ensembles distincts. Les relations traitées croisent des objets avec des attributs (ou descripteurs), et non plus des objets entre eux. On a cette fois affaire à deux ensembles liés par une relation de correspondance croisée. La méthode de sériation permet de trouver la correspondance optimale existant entre ces

deux ensembles, représentée par une relation, dite de correspondance par blocs, dont la structure met en relation des classes des deux ensembles de la façon suivante:



Dans cet article, nous serons également amenés à parler de quasi-sérialisation. Cette relation définit une correspondance par blocs d'un type un peu particulier: il n'y a pas d'obligation d'affectation de toute classe d'un ensemble à une classe de l'autre ensemble [Bede89a]. Ainsi certains objets lignes (ou colonnes) peuvent se trouver isolés et sans correspondance dans l'autre ensemble. La forme générale d'une telle relation est la suivante:



Les classifications et sériations (ou quasi-sériations) dont il sera question dans cet article sont effectuées dans le cadre de la méthodologie relationnelle. Autrement dit, tous les problèmes traités se modélisent sous forme de programmes linéaires avec contraintes [Bede89a, Marc78, Marc79].

D'une manière générale, la fonction économique correspond au critère d'adéquation de la solution aux données. Le choix de ce critère est un point fondamental puisque c'est lui qui induit la nature et l'intensité des ressemblances que l'on veut mettre en évidence. L'approche relationnelle permet de choisir parmi une vaste gamme de critères celui qui répond le mieux au problème posé par les données en présence. Certains critères opèrent sur des données binaires, d'autres sont plus appropriés à des données de fréquences; la plupart sont basés sur des règles de majorité qui déterminent le niveau de relation seuil au-delà duquel on considère que deux objets sont regroupables.

Rappelons qu'un des atouts majeurs de la méthodologie relationnelle réside dans le fait que l'on ne doit pas **fixer a priori** le nombre de classes de la partition ou le nombre de blocs de la correspondance cherchées. Ces paramètres, caractéristiques de la solution, sont directement issus du traitement, reflétant ainsi le potentiel **classificateur** inhérent aux données.

### 3.3.1 Traitement des champs Numéros de Brevets et codes CIB

Le tableau  $T$ , décrit plus haut, représente la relation liant, sous forme d'une matrice binaire rectangulaire de taille  $p \times n$ , les brevets et les codes CIB.

#### 3.3.1.1 Sériation Brevets x Codes CIB

Chaque brevet,  $l \in L$ , est décrit par un ensemble de codes CIB qui définissent son profil dans la matrice  $T$  et inversement, tout code CIB,  $i \in I$ , se retrouve dans les références d'un certain nombre de brevets.

L'objectif de ce premier traitement est de mettre en évidence les différentes tendances du corpus étudié, c'est-à-dire de voir se dessiner la répartition des brevets par familles de domaines et simultanément de définir les groupes de codes caractéristiques de classes de brevets.

Pour répondre à cette question, nous avons effectué une sériation sur la matrice  $T$ <sup>20</sup>. **Comme** nous l'avons précisé plus haut, l'orientation du résultat dépend du critère choisi pour sérier la matrice. En l'occurrence, nous avons opté pour le critère classique de sériation sur données binaires, à savoir :

$$T(Z) = \sum_l \sum_i (2 t_{li} - 1) z_{li}$$

où  $Z$  vérifie les contraintes d'une correspondance par blocs sur  $L \times I$ .

Ce critère est la généralisation aux structures rectangulaires du critère de Condorcet, très utilisé en Analyse Relationnelle pour ses qualités de règle d'agrégation [Mich85].

La maximisation du **critère**  $T(Z)$ , sous les contraintes liées à la structure de la relation cherchée  $Z$ , conduit à la mise en correspondance optimale de groupes de brevets avec des groupes de codes CIB, autrement dit on retrouve, à l'intérieur des blocs de la sériation, la réunion de brevets majoritairement décrits par les codes CIB également affectés à ce bloc. Ce résultat donne une connaissance globale des informations contenues dans les références choisies, en ce qui concerne la répartition des brevets en termes de **classification** internationale ainsi que l'appariement de domaines d'application au regard des brevets du corpus.

Par rapport à une analyse statistique descriptive classique, qui fournirait des informations "séquentielles" sur chaque brevet ou sur chaque code, nous aboutissons avec l'approche relationnelle, à une restructuration des données qui débouche sur une vision plus globale des informations de base.

---

<sup>20</sup> Sériation signifie d'une manière générale classification croisée de deux ensembles avec possibilité de classes non affectées, autrement dit le résultat pourra **être** soit une sériation, soit une quasi-sériation.

Ce mode de surveillance sectorielle permet donc d'acquérir une bonne connaissance de l'état de couverture de domaines d'applications croisés.

La méthode de sériation, à travers la correspondance optimale qu'elle fournit, génère *de facto* une partition de l'ensemble  $L$  et une partition de l'ensemble  $I$ . Mais par construction, ces deux partitions n'ont pas un caractère optimal sur leurs ensembles respectifs. Pour trouver les relations d'équivalences qui restituent au mieux les ressemblances à l'intérieur de chacun des ensembles, il nous faut effectuer des classifications séparées.

### 3.3.1.2 Classification des brevets

A partir du tableau  $T$ , on peut construire la matrice de similarités  $\hat{B}$ , de taille  $p \times p$ , croisant les brevets entre eux. Son terme général est défini comme suit:

$$\hat{b}_{ll'} = \sum_i \frac{t_{li} t_{l'i}}{t_{\bullet i}}$$

$\hat{b}_{ll'}$  est un indice de présence rareté : deux brevets sont d'autant plus ressemblants ( $\hat{b}_{ll'}$  élevé) qu'ils partagent des codes CIB ( $t_{li} = t_{l'i} = 1$ ) rares dans le corpus ( $t_{\bullet i}$  faible).

Selon la méthodologie relationnelle, on construit les similarités complémentaires (ou dissimilarités) entre objets,  $\hat{b}_{ll'}$ , de la façon suivante :

$$\bar{\hat{b}}_{ll'} = \frac{\hat{b}_{ll} + \hat{b}_{l'l'}}{2} - \hat{b}_{ll'}$$

La recherche de la partition optimale sur l'ensemble  $L$  s'effectue alors par la maximisation du critère suivant:

$$B(X) = \sum_i \sum_{l'} (\hat{b}_{ll'} - \bar{\hat{b}}_{ll'}) x_{ll'} \text{ soit encore } B(X) = \sum_j \sum_{l'} (2\hat{b}_{ll'} - \frac{\hat{b}_{ll} + \hat{b}_{l'l'}}{2}) x_{ll'}$$

où la relation  $X$  vérifie les contraintes d'une relation d'équivalence sur  $L \times L$ .

Nous avons choisi, ici, de travailler sur une similarité basée sur un indice de présence rareté.

Nous aurions pu opter pour une similarité plus couramment utilisée dans les problèmes de classification, qui se construit de la façon suivante:

$$b_{ll'} = \sum_i t_{li} t_{l'i} = \text{nombre de codes CIB communs aux brevets } l \text{ et } l'.$$

Ce traitement conduirait à regrouper les brevets ayant une majorité de codes communs, puisque les ressemblances ne sont liées, dans ce cas, qu'au nombre de descripteurs partagés par les brevets. C'est le mode de classification qu'il faut adopter si l'on cherche à

mettre en évidence des ressemblances brutes sur les profils et donc à restituer les tendances les plus fortes contenues dans les données. Mais dans ce cas, les phénomènes plus “subtils” se trouvent “écrasés”.

Or il nous paraît important, par rapport au problème que nous nous sommes posé, de ne pas négliger les configurations rares. Ceci est d’autant plus vrai que nous avons affaire, ici, à des distributions caractérisées par des lois de type **ZIPF**<sup>21</sup>. L’indice  $\hat{b}$  nous permet précisément de prendre en compte les cas de ressemblances peu fréquents et de les restituer au niveau de la solution. Ainsi, deux brevets appartiendront à une même classe non seulement parce qu’ils couvrent les mêmes domaines mais aussi parce que ces domaines sont peu souvent partagés par d’autres brevets.

On voit alors apparaître des phénomènes intéressants et difficilement décelables sur les données de base tels que l’existence de brevets **a priori** sans rapport direct combinant des technologies analogues. En outre, la classification obtenue permet à une **firme** de positionner l’ensemble de ses brevets par rapport à la concurrence et de détecter des brevets qui “se rapprochent” beaucoup des siens.

Ce type d’outil d’analyse et d’aide à la décision facilitera grandement le travail du réseau des experts, tout en diminuant les risques d’erreurs et les oublis éventuels liés à la **non prise en compte globale** de l’ensemble des liens entre les brevets.

### 3.3.1.3 Classification des codes CIB

Du tableau  $T$ , on peut également dériver la matrice de similarités  $\hat{C}$ , de taille  $n \times n$ , croisant les codes CIB entre eux. Son terme général se définit comme suit:

$$\hat{c}_{ii'} = \sum_i \frac{t_{ii'} t_{i'ii}}{t_{i \bullet}}$$

$\hat{c}_{ii'}$  est un indice de présence rareté : deux codes sont d’autant plus ressemblants ( $\hat{c}_{ii'}$  plus élevé) qu’ils décrivent simultanément des brevets ( $t_{ii'} = t_{i'ii} = 1$ ) ayant peu de codes ( $t_{i \bullet}$  faible).

Comme dans la partie précédente, la classification des codes s’effectue par maximisation de la fonction :

$$C(Y) = \sum_i \sum_{i'} (\hat{c}_{ii'} - \bar{\hat{c}}_{ii'}) y_{ii'} \quad \text{soit encore } C(Y) = \sum_{i'} \sum_i (2\hat{c}_{ii'} - \frac{\hat{c}_{ii} + \hat{c}_{i'i'}}{2}) y_{ii'}$$

sous les contraintes de relation d’équivalence sur  $I \times I$  pour  $Y$ .

<sup>21</sup> Notons, par exemple, que 20 % des sociétés déposent 80 % des brevets.

La partition obtenue regroupe les codes CIB dans des classes qui s'expliquent par les similitudes des profils des codes qu'elles contiennent mais également par le fait que ces codes apparaissent simultanément dans des brevets par ailleurs peu décrits.

A l'aide de cette classification nous obtenons une vision de la situation technique actuelle: les grosses classes sont caractéristiques d'une grande activité. Mais un autre aspect à plus long terme, tout aussi avantageux et stratégique pour l'entreprise, réside dans la recherche de l'innovation. L'innovation est une nécessité absolue pour l'entreprise. Partant du principe qu'elle se mesure aux faibles fréquences, les toutes petites classes contenant les éléments qui sont faiblement liés aux autres ou fortement liés entre eux, les parties d'innovations se trouvent dans ces classes marginales. Cette méthodologie s'avère très utile pour répondre aux besoins des entreprises dans la recherche de renseignements très ponctuels afin de réaliser une action technique ou pour connaître l'état de l'art dans un domaine précis.

Par ailleurs, la comparaison de ce type de classifications, effectuées année par année, fournirait un aperçu panoramique sur l'évolution des tendances en matière d'applications et de technologies.

### 3.3.2 Traitement des champs Noms de Sociétés et codes CIB

L'exploitation simultanée des deux tableaux  $T$  et  $T'$  permet de construire des matrices portant sur les relations croisées qu'entretiennent les ensembles  $I$  et  $J$ .

#### 3.3.2.1 Sériation Noms de sociétés x Codes CIB, en termes de poids

La première matrice que l'on pense à créer est la matrice rectangulaire, notée  $S$ , de taille  $n \times m$ , croisant les CIB et les noms de sociétés. Elle se définit, à partir des tableaux  $T$  et  $T'$ , de la façon suivante:

$$s_{ij} = \sum_l t_{il} t'_{lj} = \text{nombre de brevets décrits par le code } i \text{ et déposés par la société } j$$

Caractéristiques de la matrice  $S$  :

- Sommes en lignes

$$s_{i \cdot} = \sum_j s_{ij} = \sum_j \sum_l t_{il} t'_{lj} = \sum_l t_{il} \sum_j t'_{lj} = \sum_l t_{il} = t_{\cdot i}$$

= nombre de brevets **décrits** par le codé  $i$

- Sommes en colonnes

$$s_{\cdot j} = \sum_i s_{ij} = \sum_i \sum_l t_{il} t'_{lj} = \sum_l t'_{lj} \sum_i t_{il} = \sum_l t'_{lj} t_{l \cdot}$$

= nombre de **codes** (avec redondances) 'décrivant les brevets déposés par la société  $j$

Il est important de noter que nous travaillons, ici, en termes de *POIDS* d'une société dans un domaine car on compte les occurrences de codes et pas seulement leur présence. Expliquons cette remarque sur un petit exemple.

Soit la société  $j_0$  déposant 3 brevets  $l_1, l_2$  et  $l_3$  décrits par les 6 codes  $i_1, i_2, i_3, i_4, i_5$  et  $i_6$ , comme le montrent les sous-tableaux de  $T$  et  $T'$  suivants:

	$i_1$	$i_2$	$i_3$	$i_4$	$i_5$	$i_6$	$t_{i \cdot}$	$j_0$
$l_1$	1	1	1	0	0	1	4	1
$l_2$	1	1	0	0	1	0	3	1
$l_3$	1	1	0	1	1	1	5	1

On a  $s_{j_0} = 4 + 3 + 5 = 12$ , les codes  $l_1$  et  $l_2$  sont comptés trois fois, les codes  $l_3$  et  $l_4$  deux fois et les codes  $l_5$  et  $l_6$  une fois, **c'est-à-dire** autant de fois qu'ils apparaissent.

Mais on pourrait s'intéresser plus spécifiquement à la présence des sociétés dans les différents domaines, en pondérant les occurrences des codes CIB par leur présence globale.

Dans cette optique, on parlerait de SPECTRE d'une société dans un domaine. Nous présenterons cette approche dans la prochaine section.

Revenons au traitement de la matrice  $S$ . Sur cette matrice rectangulaire, c'est une sériation qu'il faut effectuer. Les données n'étant ni binaires, ni déduites d'une somme de relations avec une notion de majorité sous jacente, le problème du choix d'un critère de sériation se pose.

Nous avons opté pour un critère qui opère sur les profils des poids des sociétés dans les **différents** domaines en présence :

$$S_i(Z) = \sum_i \sum_j \left( \frac{s_{ij}}{s_{i \cdot}} - \frac{1}{m} \right) z_{ij}$$

**Interprétation** de ce critère:

$\frac{s_{ij}}{s_{i \cdot}}$  = part de la société  $j$  dans le domaine  $i$  (par rapport aux autres sociétés)

et  $\frac{1}{m}$  est la moyenne arithmétique des  $\frac{s_{ij}}{s_{i \cdot}}$ .

En effet, cette moyenne se calcule de la façon suivante:

$$\frac{1}{n} \sum_i \sum_j \frac{s_{ij}}{s_{i \cdot}} = \frac{1}{n} \sum_i \frac{1}{s_{i \cdot}} \sum_j s_{ij} = \frac{1}{n} \sum_i \frac{s_{i \cdot}}{s_{i \cdot}} = \frac{n}{n} = \frac{1}{m}$$

Ce critère peut s'interpréter comme un critère d'écart à la moyenne : plus  $\frac{s_{ij}}{s_{i\bullet}}$  est élevé par rapport à la moyenne  $\frac{1}{m}$ , plus la société  $j$  tient une part importante dans le domaine  $i$ .

La sériation de la matrice, avec ce critère, a pour effet de regrouper en blocs les codes CIB et les sociétés qui sont les plus "représentatifs" les uns des autres.

Ainsi, on voit apparaître des classes de sociétés ayant des poids similaires dans des groupes de domaines qui leurs sont affectés par la mise en correspondance optimale. A l'intérieur d'un bloc donné, le poids d'une société dans les domaines concernés peut aller d'une part importante jusqu'à l'exclusivité. Ce type de traitement permet de mettre en évidence le positionnement d'une société par rapport à ses concurrents à travers la présence plus ou moins accentuée des différentes firmes dans les domaines technologiques qu'elles couvrent.

Le pendant de ce critère pour l'exploitation des **profils** des poids des domaines sur les différentes **sociétés** se **construit** naturellement de la façon suivante :

$$S_2(Z) = \sum_i \sum_j \left( \frac{s_{ij}}{s_{\bullet j}} - \frac{1}{n} \right) z_{ij}$$

Interprétation de ce critère:

$\frac{s_{ij}}{s_{\bullet j}}$  = part du code  $i$  dans les dépôts de la société  $j$  (par rapport aux autres codes)  
et  $\frac{1}{n}$  est la moyenne arithmétique des  $\frac{s_{ij}}{s_{\bullet j}}$ .

En effet :

$$\frac{1}{nm} \sum_i \sum_j \frac{s_{ij}}{s_{\bullet j}} = \frac{1}{nm} \sum_j \frac{1}{s_{\bullet j}} \sum_i s_{ij} = \frac{1}{nm} \sum_j \frac{s_{\bullet j}}{s_{\bullet j}} = \frac{m}{nm} = \frac{1}{n}$$

Ce **critère** s'interprète aussi comme un critère d'écart à la moyenne :

plus  $\frac{s_{ij}}{s_{\bullet j}}$  est élevé par rapport à la moyenne  $\frac{1}{n}$ , plus le domaine  $i$  tient une part importante **dans** les dépôts de la **société**  $j$ .

Comme nous l'avons expliqué plus haut, l'indicateur  $s_{\bullet j}$  compte les occurrences brutes des codes. Le critère  $S_2(Z)$ , basé sur ces valeurs, va mettre en évidence les codes les plus utilisés par toutes les sociétés et isoler ceux qui sont moins systématiquement présents.

En l'occurrence, son utilisation ne présente donc pas un intérêt majeur. C'est pourquoi nous avons envisagé le recours à un nouvel indice de similarité  $\hat{S}$ .

### 3.3.2.2 Sériation Noms de sociétés x codes CIB, en termes de spectre

La matrice  $\hat{S}$  est construite à partir des tableaux  $T$  et  $T'$  de la façon suivante:

$$\hat{s}_{ij} = \sum_i \frac{t_{ii} t'_{ij}}{\sum_i t_{ii}} = \sum_i \frac{t_{ii} t'_{ij}}{t_{i\bullet}}$$

$\hat{s}_{ij}$  est un indice de présence rareté : un code  $i$  et une société  $j$  sont d'autant plus liés ( $\hat{s}_{ij}$  élevé) qu'ils apparaissent simultanément dans des références de brevets ( $t_{ii} = t'_{ij} = 1$ ) décrits par peu de codes ( $t_{i\bullet}$  faible).

Cette fois, on **considère** comme une information de base la fréquence d'apparition des codes CIB dans les références de brevets. Si une société couvre un domaine, même de façon peu dominante, la sériation résultat saura en tenir compte.

### Caractéristiques de $\hat{S}$

- Sommes en lignes

$$\hat{s}_{i\bullet} = \sum_j \hat{s}_{ij} = \sum_j \sum_i t_{ii} \frac{t'_{ij}}{t_{i\bullet}} = \sum_i \frac{t_{ii}}{t_{i\bullet}} \sum_j t'_{ij} = \sum_i \frac{t_{ii}}{t_{i\bullet}}$$

$\hat{s}_{i\bullet}$  est d'autant plus élevé que le code  $i$  apparaît ( $t_{ii} = 1$ ) avec peu d'autres codes dans des références de brevets ( $t_{i\bullet}$  faible).

Les codes les plus fréquents "n'absorberont" pas toute l'information, permettant ainsi de mettre en lumière des phénomènes plus rares.

- Sommes en colonnes

$$\hat{s}_{\bullet j} = \sum_i \hat{s}_{ij} = \sum_i \sum_i' \frac{t_{ii} t'_{ij}}{t_{i\bullet}} = \sum_i' \frac{t'_{ij}}{t_{i\bullet}} \sum_i t_{ii} = \sum_i' t'_{ij}$$

= nombre de brevets déposés par la société  $j$

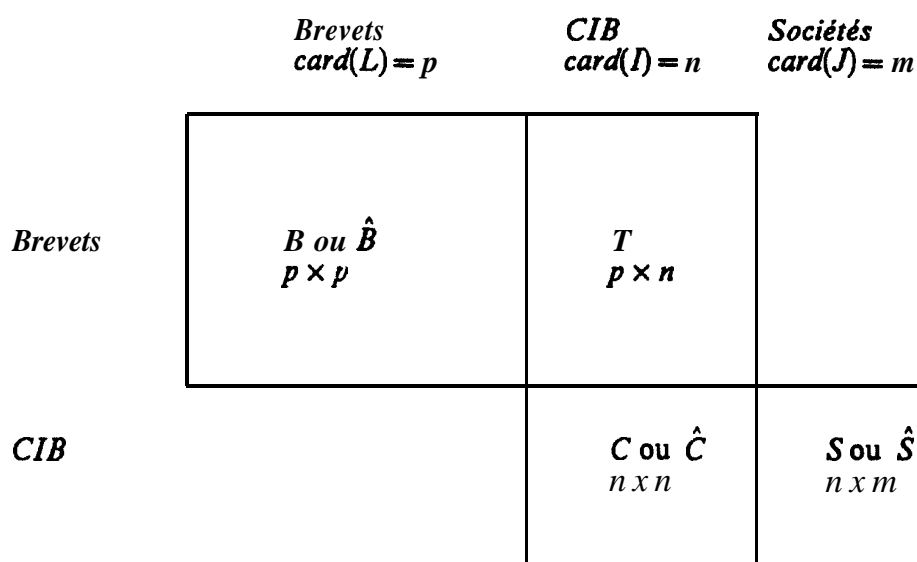
Le choix du critère de sériation peut ensuite s'effectuer comme dans la partie précédente, en fonction du problème que l'on veut cerner. Mais quelle que soit la préoccupation, on sait que les résultats feront émerger une correspondance entre domaines d'activités et sociétés, non plus en termes de couverture dominante, mais en termes de simple présence dans un domaine.

Le choix de cette similarité,  $\hat{S}$ , peut-être moins intuitif que celui de  $S$ , permet de comprendre un autre aspect de la répartition des activités des sociétés. Certes avec  $S$  on répond à des questions plus immédiates de positionnement stratégique, mais les informations apportées par la pondération permettent de mieux cerner les phénomènes de couverture des technologies et donc de variété des activités.

### 3.3.3 Récapitulatif

Le schéma suivant présente les différents croisements que l'on peut effectuer avec les trois ensembles, et les structures des matrices relationnelles directement extraites des données de base ou bien dérivées de celles-ci.

- Optique Classification directe :
  - 1) Classification des brevets ( sur  $B$  ou  $\hat{B}$  )
  - 2) Classification des codes ( sur  $C$  ou  $\hat{C}$  )
- Optique Classification croisée :
  - 3) Sériation Brevets x CIB ( sur  $T$  )
  - 4) Sériation CIB x Noms des sociétés ( sur  $S$  ou  $\hat{S}$  )



A chacun de ces problèmes, posé en termes mathématiques, correspond une préoccupation d'analyse de l'information extraite du corpus :

- 1) Positionnement relatif des brevets vis à vis de la stratégie d'interrogation (connaissance des liens entre **différents** brevets, regroupement de brevets en classes indépendantes,...).
- 2) Connaissance des relations entre les divers domaines de recherche ou d'applications.
- 3) Connaissance des liaisons spécifiques entre les brevets et les domaines d'applications.
- 4) Bien cerner la stratégie de recherche commune ou spécifique des entreprises.

### 3.4 Ouvertures

Nous nous sommes limités, dans le présent article, à l'étude des relations entre les champs **nom des sociétés, numéro des brevets** et codes **CIB décrivant ces brevets**. **Mais il** est tout à fait envisageable d'appliquer les mêmes méthodes sur le croisement d'autres champs descriptifs. Les résultats obtenus donneront des informations stratégiques d'une autre nature. Enfin tout type d'information restructurée sous une forme conventionnelle peut

également faire l'objet de tels traitements. Nous pensons par exemple aux informations dites **informelles** pour effectuer des recoupements et augmenter ainsi leur niveau de **sûreté**. Parmi les questions prioritaires en matière de veille technologique, nous pensons à deux types de **problèmes** qu'il serait intéressant d'aborder; d'une part le travail sur les pays d'extensions et d'autre part l'analyse comparative entre **codification** interne et CIB.

Rappelons que l'une des règles d'or dans le domaine de la propriété industrielle est la suivante :

**«Le brevet délivré dans un Etat (ou groupe d'Etats) ne protège l'invention que sur son territoire.»**

Ainsi lorsqu'un industriel français dépose un brevet en France il peut, pendant une période de douze mois, faire valoir ce droit de **dépôt** prioritaire pour **étendre** son brevet dans d'autres pays membres de **l'Union Internationale pour la Protection de la Propriété Industrielle (Union de Paris)**. A la suite de cette procédure, l'invention est protégée dans ces pays où la **délivrance** a été **effectuée**.

Cette information, contenue dans la référence du brevet, est **très** intéressante puisqu'elle donne une bonne idée de la stratégie de **dépôt** de l'entreprise sur le plan international.

A partir de ces faits, nous pouvons concevoir des traitements analogues à ceux que nous avons présentés, en croisant cette fois les noms des sociétés avec les pays d'extensions de leurs brevets." Nous pourrions ainsi définir une classification des sociétés en fonction de leurs axes de stratégie en **matière** de **dépôts**.

Il arrive souvent que des entreprises ou des organismes d'état, travaillant dans des domaines **très** pointus, ne soient pas satisfaits de la CIB et adoptent, en interne, de nouvelles classifications. Nous pensons qu'il est possible d'établir une correspondance entre ces classifications internes et la CIB.

L'ensemble des codes représente la totalité de la description de l'activité technique et technologique puisqu'ils servent à juger de l'activité innovante d'un brevet. L'avantage que nous voyons à définir cette correspondance se situe au niveau de la formulation des stratégies d'interrogations des serveurs de bases de données brevets.

---

<sup>22</sup> voir le chapitre **sur "la protection internationale des inventions"** dans le livre de B. **Phelip, Brevets d'invention, J.Delmas** et Ci, 3<sup>ème</sup> Edition, (1989).

<sup>23</sup> L'idée originale de cette matrice est due à W. **Nivol** ancien étudiant de la première promotion (1989-1990) du **DEA** d'Information **Stratégique**, Veille Scientifique et Technologique, de l'**Université** d'Aix Marseille III et qui **effectue actuellement** une **thèse** sur l'analyse statistique de base de **données** industrielles.

Il apparaît clairement qu'une multitude de traitements sont envisageables et notre objectif, ici, n'était pas de tous les énumérer mais plutôt de proposer des ouvertures par l'utilisation d'une méthodologie d'analyse des données encore peu connue dans le domaine de la veille technologique.



# *Troisième Partie*

Application d'une nouvelle  
méthode de  
classification automatique en veille  
technologique  
**l'Analyse** Factorielle-Relationnelle

*Charles Huot*

Revue Française de Bibliométrie, n°8, pp  
78-105, décembre 1990



# Application d'une nouvelle méthode de classification automatique en veille technologique: l'analyse factorielle relationnelle

Charles HUOT  
Centre Européen de Mathématiques Appliqués **IBM**<sup>®</sup>

## 4.1 Introduction

Aujourd'hui tous les services de veille technologique des grandes entreprises utilisent les serveurs de banques de données pour avoir **accès à** l'information scientifique et technique . **Qu'il** s'agisse d'information concernant des brevets ou des articles scientifiques, les  **systèmes** d'interrogation en ligne permettent de bien cerner une question, et de fournir **à** l'utilisateur les **références** concernant le sujet qui l'intéresse. Cependant il arrive souvent que cet ensemble de **références** soit assez important. Le professeur H. Dou et son équipe du Centre de Recherche Rétrospective de **Marseille** (CRRM) se sont donc attachés, depuis 10 ans déjà, à traiter en mode local sur micro-ordinateur ces **références** téléchargées [**Dou89, Quon90a**] Ces traitements permettent notamment de reformater les références des documents pour les présenter sous une forme matricielle, directement traitable par des méthodes d'analyse des données. La thèse de L. Quoniam [**Quon88**] montre combien il est justifié et souhaitable d'utiliser ces méthodes pour obtenir une information plus pertinente. De nombreux auteurs ont travaillé dans ce domaine parmi lesquels nous citerons C. Paoli [**Paol87**] C. Dutheil [**Duth90**], W.A. Turner [**Turn89**], B. Dousset [**Dous87**]

Les méthodes le plus couramment utilisées sont l'analyse factorielle des correspondances multiples et les méthodes de classification hiérarchique. L'un des problèmes de ces méthodes est lié d'une part à la faible densité de remplissage des matrices traitées et d'autre part à la **difficulté** d'interprétation des résultats obtenus. La méthode que nous utilisons dans cet article est issue de la théorie de l'**Analyse** Relationnelle. Cette théorie fut développée par F. Marcotorchino et P. **Michaud** dans les années 70 [**Marc78, Marc81**]. Elle a donné lieu **à** un grand nombre de travaux de recherche, **thèses** et articles. Une partie de cette théorie a trait **à** la classification automatique. En effet l'analyse relationnelle propose un certain nombre d'outils de classification, comme la sériation [**Marc87**], la quadri-décomposition [**Bede89b, Bede89a**] ou l'analyse factorielle-

---

% **CEMAP** IBM, 3-5 Place Vendôme, 75001 Paris

relationnelle [Marc89, Marc91a] qui classifient des ensembles sans **fixation** a priori du nombre de classe. L'une des plus remarquables applications de l'analyse relationnelle a été faite dans le domaine de la lexicographie ou 1. Warnesson a utilisé ces techniques pour restructurer des dictionnaires, en particulier les dictionnaires de synonymes [Warn90].

Le but de cet article est double. Il s'agit tout d'abord de montrer les possibilités de ces méthodes de classification automatique dans le domaine de la veille technologique et de la documentation automatique, de mettre en évidence leurs points forts: la non **fixation** du nombre de classe, la facilité d'interprétation des résultats, la possibilité de prise en compte d'une grande variété de données.

Le second aspect que nous avons voulu présenter ici est directement lié à la position de ce type de méthode dans le processus qui transforme l'information brute en information pour décideur. L'étude que nous avons effectuée nous permet de mettre en évidence des corrélations entre des domaines divers n'ayant a priori pas de rapport entre eux. Cela montre également la position du brevet d'une société vis à vis de la concurrence. Des rapports de ce type, pour qu'ils soient exploitables par un décideur, doivent obligatoirement transiter par l'intermédiaire d'un expert <sup>25</sup> du domaine qui portera son appréciation sur l'étude. Mais il est clair qu'une analyse bien conçue et réalisée **efficacement** apportera un plus considérable au travail de l'expert.

#### 4.2 Stratégie d'interrogation

Le serveur qui a été utilisé est le serveur américain **SDC-ORBIT**. Ce serveur est le principal concurrent du serveur Lockheed (USA). Le SDC Search Service de System Development Corporation propose un service de recherche automatique dans un certain type de bases sélectionnées.

Les principales bases de données utilisées systématiquement lorsqu'on parle de brevets sont celles de DERWENT. Constituées à partir du fonds documentaire élaboré par DERWENT Publications Ltd (Londres) depuis 1963 pour la pharmacie, 1965 pour l'agriculture, 1966 pour les **polymères**, 1970 pour la chimie et 1974 pour le reste des activités industrielles, elles sont très largement utilisées dans l'ensemble des pays industrialisés: Europe, USA, Japon. WPI (WORLD PATENTS INDEX) est relatif aux années antérieures à 1981, WPIL (WORLD PATENTS INDEX **LATEST**) couvre les périodes

---

<sup>25</sup> Nous nous plaçons ici dans un **système** de veille technologique fonctionnant sur le **modèle défini** par F.JAKOBIAK dans son ouvrage sur l'exploitation systématique des informations industrielles [Jako90].

postérieurs à 1981. L'ensemble comprend plus de 3 millions de documents et s'enrichit de 300 000 **références** par an, par mises à jour hebdomadaires.

Les multiples possibilités d'interrogation en ligne (vocabulaire **contrôlé**, codes divers, classification internationale des Brevets...) sont complétées par une impressionnante quantité de supports imprimés mis à la disposition des souscripteurs payant une redevance annuelle assez élevée. La série des Basic Abstracts Journals comporte douze éditions, une par grand domaine, et contient les titres et résumés Derwent de bonne **qualité**. A ces informations s'ajoute le World Patent Index Gazette qui alerte, un mois avant les Basic Abstracts Journals, les veilleurs Technologiques. Si elle ne comporte pas de résumé, cette publication comprend des titres très informatifs, réécrits par DERWENT et les références pour les demandes publiées (**unexamined**) et les brevets délivrés (**granted**). Le choix de cette base de données est donc lié à son sérieux et au nombre de références qu'elle contient.

Nous avons également utilisé la base de données USPA (**USPAtent**), pour la recherche des citations de brevets.

Le cas que nous allons traiter est issu d'une recherche documentaire effectuée au **C.R.R.M** de Marseille par Luc Quoniam. Il s'agissait de réaliser une étude sur les brevets déposés dans le domaine du traitement de surface des lentilles de contact. La substance utilisée pour ces traitements est la **Papaïne**. Elle permet d'obtenir la dissolution des protéines se déposant sur les lentilles les rendant opaques et irritantes.

L'extraction des **références** s'est effectuée en deux phases.

1. Interrogation sur le concept "Lentilles de contact et Papaïne". (WPI)

«CONTACT LENS **AND PAPAINE** ?»

En réponse à cette question, la base de données nous a renvoyé 4 numéros de brevets.

- US4757014
- US474951 **1**
- US4614549
- us4609493

Nous les avons appelés les **brevets de départ** et nous leur avons affecté le code d'ensemble **SS1** (Search Strategy 1).

2. Nous avons ensuite voulu connaître les brevets qui citent ces brevets de départ. Nous avons donc interrogé sur les champs citations des brevets de la base USPA.

**«(US4757014/CT OR US4749511/CT OR US4614549/CT OR US4609493/CT)»**

Cette fois-ci nous avons obtenu 9 réponses.

- US4872965
- us4710313
- US4767559
- US4784790
- US4808239
- US4829001
- US4832754
- US4839082
- US4855234

Nous avons appelé ces 9 nouveaux brevets, les **brevets citants** et nous leur avons associé le nom d'ensemble SS2 (Search Strategy 2).

L'ensemble des références issues de cette recherche est présenté en annexe.

C'est à partir de ces données brutes que nous avons constitué la matrice qui a été soumise à la classification.

### **4.3 Analyse par les codes DERWENT**

#### **4.3.1 La Matrice Initiale**

Partant de notre ensemble de **références**, nous avons pris les deux ensembles SS1 + SS2, des brevets de base et des brevets citants.

A l'aide d'outils informatiques d'extraction et de retraitement des champs contenus dans les **références** des brevets (programmes du **C.R.R.M**), nous avons extrait les codes

DERWENT (DC-) contenus dans les 13 **références** (SS1 + **SS2**). Les **références** dont il est question (“SS1 dans **WPI**” et “SS2 dans WPI”) figurent en annexe.

Les codes DERWENT utilisés dans WPI se trouvent dans un champ intitulé “DC”. Ces codes sont mis par les personnes qui indexent les documents brevets. Ils permettent de savoir dans quel domaine se situe le brevet. Bien **sûr**, il peut se trouver, et c’est généralement le cas, qu’un même brevet comprenne plusieurs codes. Dans cette codification aucun élément n’a plus de poids qu’un autre. **Il** existe des codes pour la description de l’ensemble de l’activité scientifique et technique.

Le traitement des 13 **références** fait **apparaître** au total 15 codes DERWENT **différents**. C’est à partir de ce moment que l’on construit une matrice. Nous mettons en ligne les **numéros** de nos 13 brevets et en colonne les 15 codes DERWENT.

Nous noterons  $c_{ij}$  la case du tableau située à l’intersection de la ligne  $i$  et de la colonne  $j$ .

L’information est restituée sous la forme:

$c_{ij} = 1$  si le brevet  $i$  possède le code DERWENT  $j$

$c_{ij} = 0$  sinon.

Nous obtenons une matrice relativement creuse de type «**présence-absence**». Les sommes en ligne ne sont pas constantes, (elles varient entre 3 et 6); chacune représente le nombre de codes qu’il faut **à** un indexeur pour caractériser 1 brevet. Cela donne une idée de la richesse et de la **spécificité** des brevets.

### Matrice initiale

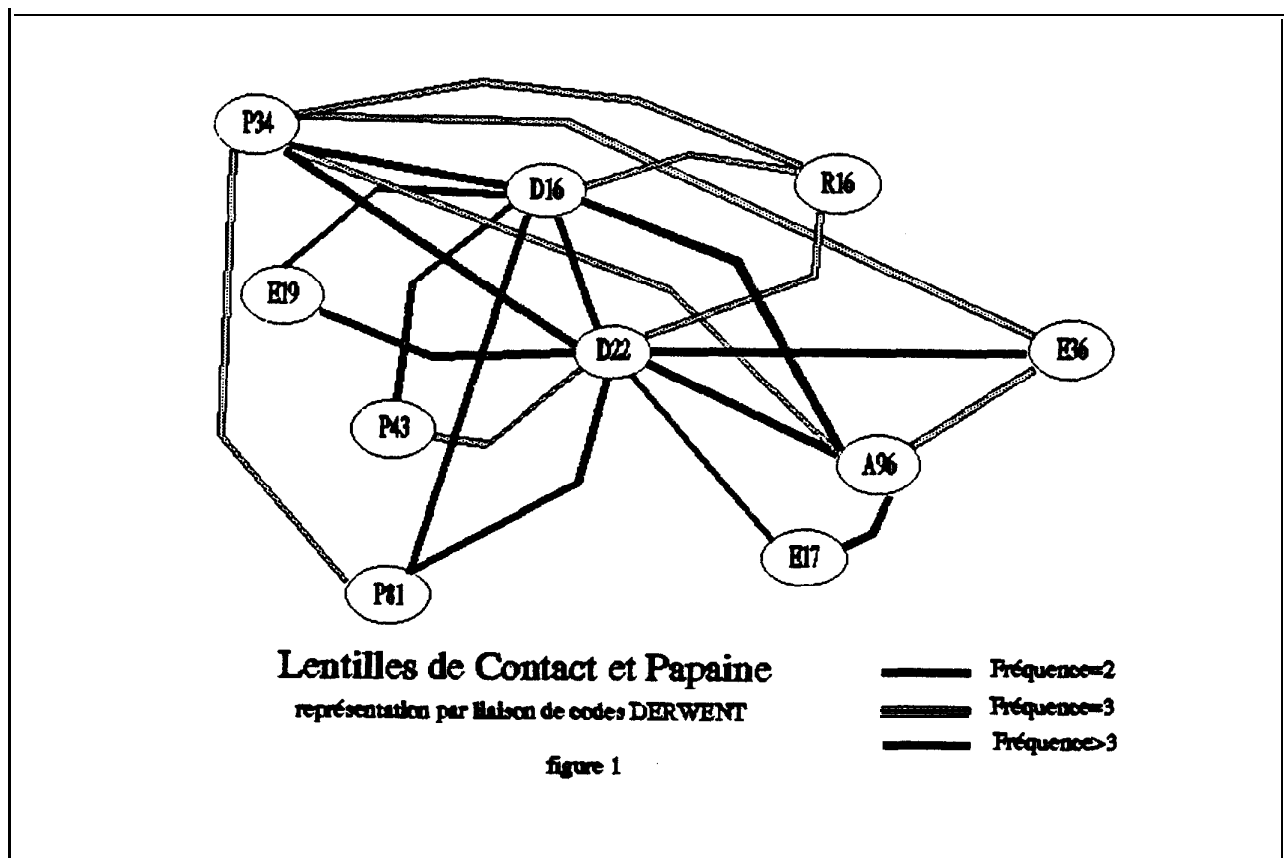
	D	P	D	P	A	E	R	E	D	P	S	A	E	P	E	T
	2	3	1	8	9	3	1	1	2	4	0	9	1	3	3	0
	2	4	6	1	6	6	6	7	5	3	5	7	9	1	7	L
<b>US474951</b>	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	4
<b>us7757014</b>	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	6
<b>US4609493</b>	1	0	1	1	1	0	0	1	1	0	0	0	0	0	0	6
US4614549	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	5
US4872965	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
US4839882	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
US4832754	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	4
US4808239	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	4
US4767559	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	3
US4784798	1	1	1	0	0	0	0	0	0	0	0	1	1	1	0	6
US4829801	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	6
US4855234	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	6
us4710313	1	0	1	1	0	0	0	0	0	0	0	0	1	0	1	5
<b>Total</b>	12	7	9	5	6	4	3	2	1	4	1	1	2	1	1	59

### Signification des codes DERWENT de notre analyse

- D22 **Desinf, Deter**, Dental, Sterilizing, bandages, sutures, plaster **casts**, prostheses. (les lentilles de contacts sont assimilables à des prothèses oculaires)
- P34 Health, amusement. Sterilizing, Syringes, Electrotherapy
- D16 Food, Fermentation industry, Brewing, **Yeast**, Pharmaceuticals **Alcohol**
- P81 Optics, Photography, General. Optics
- A96 Veterinary, **medical**, dental.
- E36 General inorganic, **None-metallic** elements
- R16 Measuring, testing, Investigating **chem./phys.props.**
- E17 General organic, Other aliphatics
- D25 Desinf.deter. soap.including** metal salt and fatty **acids** used in soaps
- P43 Separating, mixing. Sorting, cleaning
- SO5 Electromedical
- A97 Miscellaneous goods
- E19 General organic, other organic compounds general
- P31 Health , amusement. Diagnosis, Surgery

E37 General inorganic, mixtures of many components

Nous avons représenté en figure 1 la fréquence des liaisons (> 1) entre les différents codes DERWENT des 13 brevets.



#### 4.3.2 L'Analyse Factorielle Relationnelle

##### 4.3.2.1 Partition des Brevets (figure 2)

Nous avons utilisé ici le critère de Condorcet pondéré pour effectuer une classification des brevets en fonction des codes qu'ils possèdent. Par maximisation du critère de Condorcet pondéré nous obtenons une classification de notre ensemble de brevets en 9 classes constituées de la façon suivante:

- Classe 1  
US47495 11 (brevet de SS 1)

US4614549 (**brevet de SS1**)

**U S4767559** (brevet SS2)

- Classe 2

US7757014 (**brevet de SS1**)

US4829001 (brevet SS2)

US4855234 (brevet SS2)

- Classe 3

**US4609493** (**brevet de SS1**)

- Classe 4

**US4832754** (brevet SS2)

- Classe 5

**US4872965** (brevet SS2)

- Classe 6

US4839082 (brevet SS2)

- Classe 7

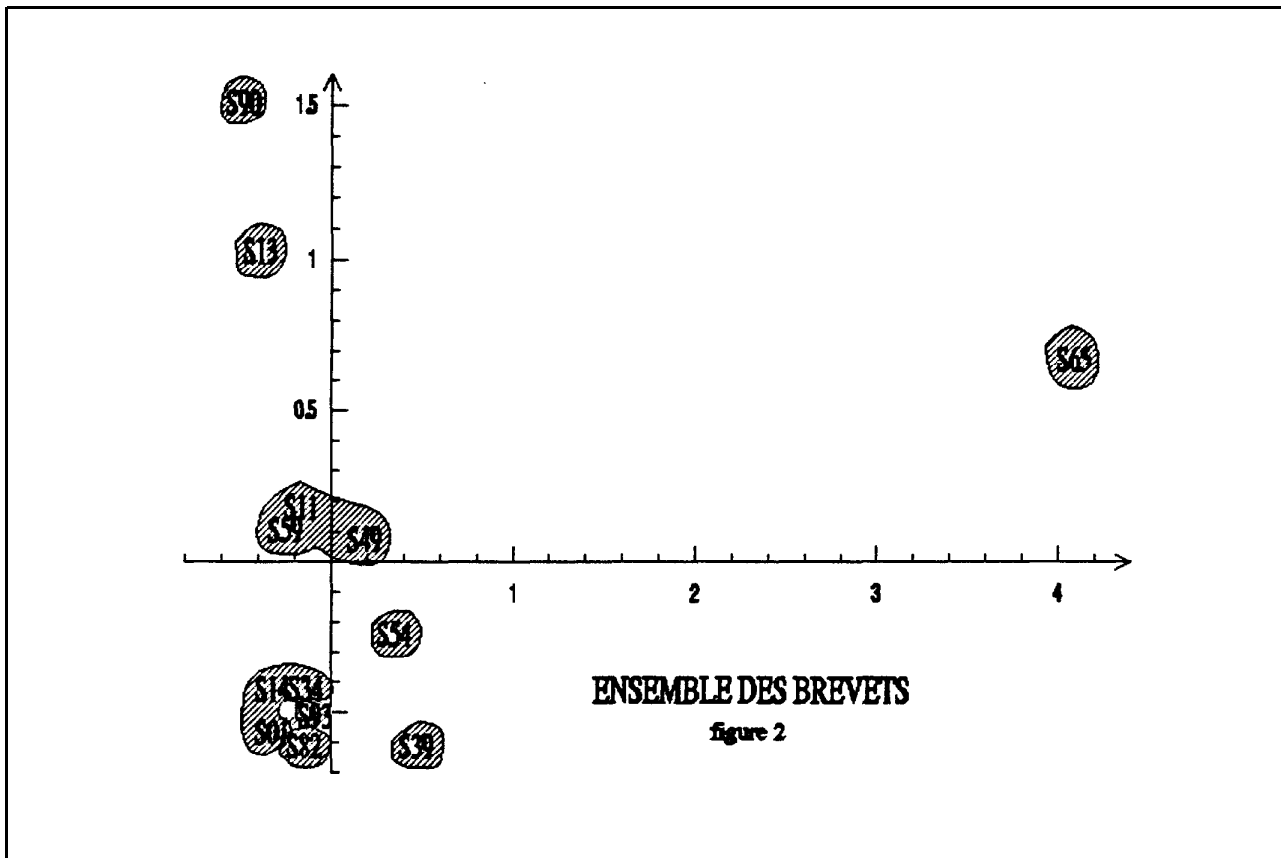
US4808239 (brevet SS2)

- Classe 8

**US4784790** (brevet SS2)

- Classe 9

US4710313 (brevet SS2)



L'analyse nous donne une répartition des brevets en 9 classes dont 7 ne possèdent qu'un seul élément. Ce résultat s'explique par le fait que notre matrice comporte peu de brevets, que certains sont seuls à posséder un code et donc cela va fortement jouer pour isoler ces individus. Ce type de phénomène serait moins marquant si la matrice était de plus grande taille. Néanmoins nous avons deux classes très intéressantes. 'L' une contient notamment deux brevets de départ.

Etudions de plus près ces classes

- Classe I

US47495 **11 (brevet SS 1)**

US4614549 (**brevet SS1**)

US4767559 (brevet SS2)

Ces brevets traitent de la stérilisation en général, des éléments qui touchent l'homme de près (bandages, sutures, lentilles,...) (**D22**), mais également de la stérilisation **d'instru-**

ments (seringues, électrothérapie) (P34). Ils traitent également de sujets rattachés aux problèmes de fermentation, de stérilisation alimentaire et alcools pharmaceutiques (D16) (à l'exception du brevet **US4767559**), ainsi que des domaines liés à l'optique générale ou la photographie (P81).

Le brevet US4614549 porte également sur les problèmes liés à la séparation de mélange (P43).

Cette classe contient l'ensemble des domaines que l'on pouvait s'attendre à trouver dans notre examen. En effet, les **thèmes** de la stérilisation et de l'optique sont très bien couverts par nos brevets. Nous pouvons prendre cette classe comme **référence** de l'état global de la technologie à posséder impérativement.

- Classe 2

US7757014 (brevet **SS1**)

US4829001 (brevet SS2)

US4855234 (brevet SS2)

Cette classe est un peu particulière dans la mesure où les trois brevets possèdent exactement les mêmes codes DERWENT. C'est bien sûr cette identité qui a motivé leur regroupement.

Ces brevets traitent, comme les précédents, de domaines liés à la stérilisation et aux lentilles (**D22**), (**P34**), à la fermentation (D16). Mais ils se distinguent par le fait qu'ils touchent également au domaine dentaire (vétérinaire, et médical) (**A96**), au domaine de la chimie inorganique générale des éléments non-métalliques (E36) et enfin à la recherche et à la mesure de propriétés physiques et chimiques de composés (R16). Ce sont d'ailleurs les seuls brevets à couvrir ce dernier domaine.

- Classe 3

US4609493 (brevet **SS1**)

C'est une classe singleton comme toutes celles qui suivent. Ce brevet recouvre les domaines de stérilisation de prothèses, bandages, lentilles, etc (**D22**), de la fermentation (**D16**), ainsi que celui de l'optique générale et de la photographie (P81). D'ailleurs sur les 5 brevets qui couvrent le domaine de l'optique, 3 sont des brevets de départ (SSI).

Il a trait ensuite au domaine dentaire (vétérinaire, **médical**)(**A96**).

Enfin cette classe est caractérisée par les deux domaines suivants: chimie organique générale autre que les aliphatiques (**E17**) et désinfectants, détergents et savons, incluant également les sels métalliques et les acides gras utilisés pour les savons (D25).

- Classe 4

US4832754 (brevet SS2)

On retrouve un brevet qui traite des lentilles (**D22**), et fermentation (**D16**), mais aussi de chimie inorganique générale des éléments non-métalliques (E36) et enfin le domaine de séparation de phases liquides (P43).

- Classe 5

US4872965 (brevet SS2)

C'est une classe curieuse, car ce brevet est le seul qui ne traite pas du domaine de la stérilisation. Il ne couvre que deux domaines qui sont l'électromédical (**S05**) et la séparation de phases Liquides (P43).

- Classe 6

US4839082 (brevet **SS2**)

Cette classe comporte un brevet qui couvre les lentilles (D22) et le domaine dentaire (vétérinaire et **médical**)(**A96**).

- Classe 7

US4808239 (brevet SS2)

Ce brevet couvre les domaines de la stérilisation des lentilles (**D22**), du dentaire plus général (vétérinaire et **médical**)(**A96**), de la chimie organique générale (sauf le domaine des aliphatiques) (**E17**) et du domaine de la séparation de phases liquides (P43).

- Classe 8

US4784790 (brevet SS2)

Cette classe couvre les domaines classiques de stérilisation de lentilles (**D22**), (P34) et de fermentation (**D16**), mais également d'autres domaines qu'il est pratiquement le seul à posséder dans notre ensemble. Ainsi il possède les codes relatifs à la santé et à la chirurgie (**P31**), aux biens divers (A97) et à la chimie organique (autres que les composés généraux) (E 19).

- Classe 9

US4710313 (brevet SS2)

La classe couvre le domaine de stérilisation générale (**D22**), de fermentation (D16) mais également d'optique générale et photographie (P81). De plus elle traite de la chimie organique (moins les composés organiques généraux) (**E19**) et de la chimie inorganique (mélange d'un grand nombre de **composés**)(**E37**). Notons que c'est le seul brevet qui couvre cette **dernière** catégorie.

Les classes 1 et 2 sont celles qui sont les plus représentatives du domaine d'application choisi. La taille de ces classes mesure l'activité industrielle ainsi que l'effort de recherche

relativement à ce domaine. Ces classes importantes sont créées car les brevets qu'elles contiennent possèdent des profils proches en terme de codes DERWENT. C'est cette ressemblance qui détermine le partitionnement. Mais à la différence de la quasi-sérialisation, l'analyse factorielle-relationnelle accorde une grande importance au fait que 2 brevets sont les seuls à posséder le même code. C'est la notion de "**présence-rareté**" définie par F.Marcotorchino [Marc89] qui joue dans ce cas.

Le fait que nous retrouvons dans les classes principales 3 brevets de base sur 4 est tout à fait caractéristique de ce phénomène. Nous voyons ainsi tout l'avantage de la technique que nous utilisons et qui consiste à aller chercher l'information plus difficile à obtenir. **En Veille Technologique, il ne faut jamais se contenter de l'information facile à obtenir.**

Cette carte de décision où nos classes sont représentées sur l'ensemble projeté  $n^\circ$  de brevet / codes DERWENT permet de bien représenter la situation industrielle dans le domaine interrogé.

Ainsi une  **firme**  peut facilement positionner son brevet par rapport à la concurrence et détecter des brevets qui se rapprochent de "trop près" des siens. Ce mode de surveillance sectorielle doit être systématisé si l'on veut avoir une bonne connaissance de l'état de la technique. Il s'agit d'un formidable outil d'analyse et de décision qui facilitera très fortement le travail du réseau des experts, tout en diminuant les risques d'erreurs et les oublis éventuels liés à un manque de finesse de la méthode d'analyse.

Les classes de très petite taille ( ici 1 élément ) contiennent les brevets qui traitent du domaine mais avec une  **différence**  sensible. Cela correspond à de nouveaux axes de recherche ou bien à des axes de recherche qui sont abandonnés, ce qui est vérifiable par la date. Donc c'est ici que nous allons trouver les brevets qui sont des charnières entre des technologies ou des domaines d'applications.

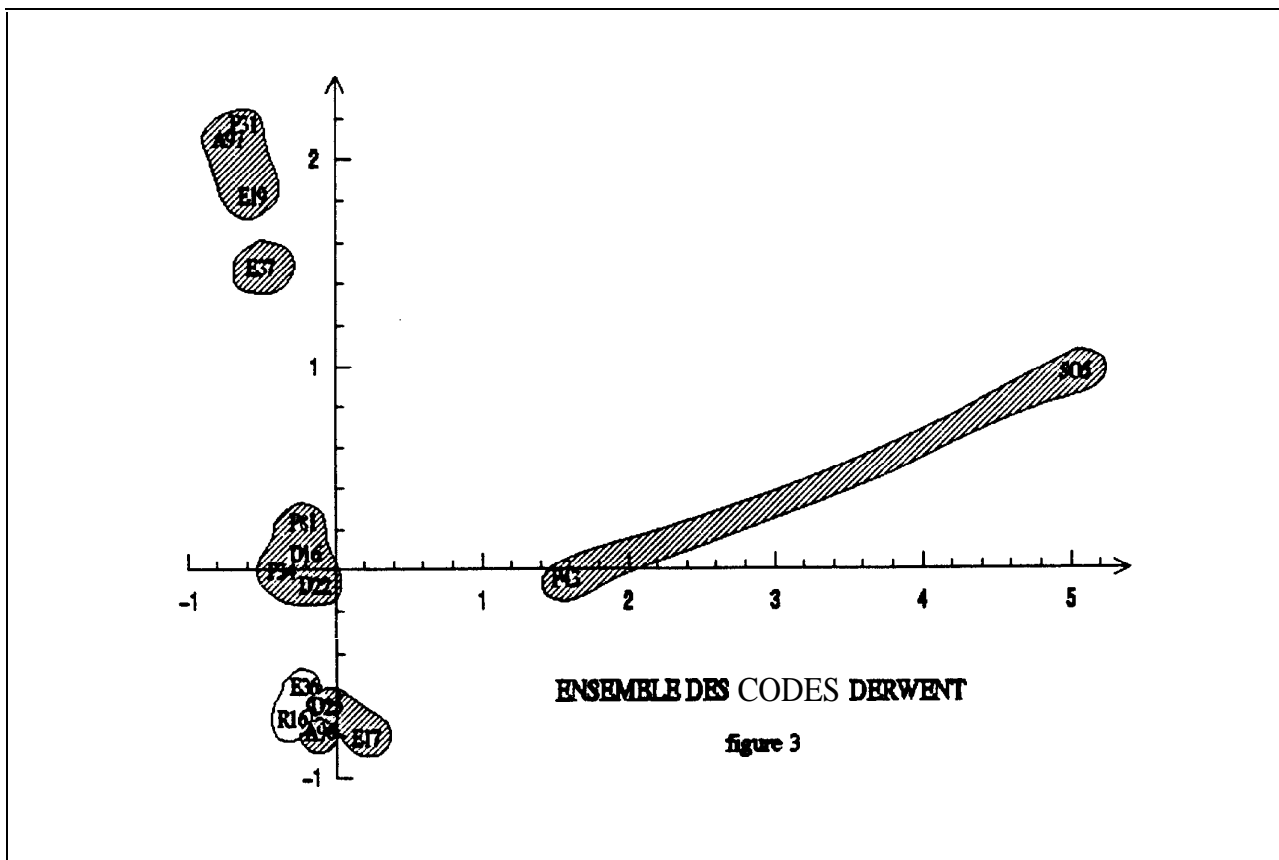
#### 4.3.2.2 Partition des codes DERWENT (figure 3)

**Nous** avons utilisé ici le critère de Burt pondéré pour effectuer une  **classification**  des codes DERWENT en fonction des brevets qui les possèdent. Rappelons que le nombre de classes obtenu, ici 7, n'est pas fixé a priori.

- Classe 1
  - D22  **Desinf, lens, ...**
  - P34 Sterilising, ...
  - D16 Pharmaceuticals,  **Alcohol, ...**
  - P81**  Optics, ...
- Classe 2

A96 **Medical**,...

- Classe 3
  - E36 General inorganic,...
  - R16 Investigating chem./phys.props.**,...
- Classe 4
  - E17 General Organic (non aliphatics)
  - D25 Desinf. deter.** soap including metal salt and fatty **acids** used in saops
- Classe 5
  - P43 Cleaning,...
  - SO5 Electromedical
- Classe 6
  - A97 Miscellaneous Goods
  - E 19 General organic,.. .
  - P31 Health**,...
- Classe 7
  - E37 General inorganic, mixtures of **many** components



L'étude classe par classe permet de comprendre le domaine que nous étudions.

• Classe 1

- D22 Desinf,lens,...
- P34 Sterilising,...
- D 16 Pharmaceuticals Alcohol,...
- P8 1 Optics,...

Cette première classe comporte 2 codes traitant de la stérilisation dans les domaines qui touchent à la santé de l'homme (D22), (P34), 1 autre touchant plus aux problèmes de fermentation industrielle, levure ou alcools touchant à la nourriture (D16) et un dernier relevant de l'optique générale (P81). Ces codes sont regroupés par le fait qu'ils apparaissent souvent ensemble. Ils représentent pour nous les codes «triviaux». En effet il n'est pas concevable qu'en interrogeant le domaine des lentilles de contact et de la papaine on ne retrouve pas ce qui touche à la stérilisation, la santé et l'optique. Ainsi la majorité des brevets possèdent naturellement ces codes.

- Classe 2

A96 Dental,...

Il s'agit, dans cette classe, du code relatif au problème vétérinaire et médical du domaine dentaire. Son isolement dans une classe est dû à sa liaison avec des codes ayant une faible fréquence d'apparition.

- Classe 3

E36 General inorganic,...

R16 Investigating chem./phys.props.,...

Cette classe composée de la chimie inorganique des éléments non métalliques (E36) et des méthodes de mesure et de recherche de propriétés physique et chimique d'éléments (R16) est essentiellement due au fait que ces codes sont possédés par une seule classe de brevet (classe 2).

- Classe 4

E17 non-Aliphatics,...

**D25 Desinf. deter.** soap including metal salt and fatty **acids** used in soaps

Une classe qui relie le groupe des détergeants (D25) et celui de la chimie organique (sans les **aliphatiques**)(E 17).

- Classe 5

P43 mixing. Sorting,...

SO5 Electromedical

Cette classe montre le rapprochement de l'électromédical (SO5) et des méthodes de séparations de phases liquides (P43).

- Classe 6

A97 Miscellaneous goods

**E19** General organic,...

P3 1 Health,...

Cette classe se justifie par le brevet US4784790 qui traite à la fois des biens divers (**A97**), de la chirurgie et de la santé (P3 1) ainsi que de la de chimie organique (sans les composés organiques **généraux**)(P3 1).

- Classe 7

E37 General inorganic, mixtures of **many** components

Ce code n'apparaît que dans un seul brevet ce qui lui vaut l'exclusion des autres classes. Il est relatif aux mélanges de divers composés en chimie inorganique.

La partition sur les codes DERWENT nous apporte 2 types informations.

1. Tout d'abord nous obtenons une topographie de la situation actuelle, cela par les grosses classes qui sont caractéristiques d'une grande activité dans l'ensemble et le mélange de ces domaines. Ainsi si nous retrouvons dans une même classe les codes D22, P34, **D16**, P8 1, c'est parce qu'ils sont les éléments de base de cette technologie (ce fait est confirmé par l'analyse plus **fine** des références où ils apparaissent dans la majorité des brevets). Cette méthode semble donc donner de bonnes indications quant à la situation présente. Il faut rappeler que notre matrice initiale n'est pas très importante et que de nouvelles analyses sur de plus grosses matrices sont en préparation pour confirmer nos hypothèses. De même **il** ne s'agit ici que d'une analyse des données, or dans le processus de la veille technologique tel que nous l'avons rapporté en introduction, ce type de résultat doit être transmis à un expert du domaine qui confirmera l'intérêt du résultat obtenu.
2. Mais un autre aspect à plus long terme, tout aussi avantageux et stratégique pour l'entreprise, réside dans la recherche de l'innovation. Nous l'avons dit plusieurs fois, l'innovation est une nécessité absolue pour la firme si elle veut subsister. Partant du principe que l'innovation se mesure aux faibles fréquences, les toutes- petites classes contiennent les éléments qui sont faiblement liés avec les autres codes, les parties d'innovations se trouvent dans ces codes marginaux. On retrouve le résultat obtenu pour l'analyse des brevets à savoir qu'un élément à faible fréquence va donner de la valeur au brevet qui le détient et va tendre à l'isoler dans une classe. Mais dans cette partie comme dans la précédente il faudrait un expert pour confirmer le résultat.

## 4.4 Annexes

---

### 4.4.1.1 SS1 dans base WPI

-1-

AN - 88-058018/09

XRAM- C88-025833

XRPX- N88-044088

TI - Cleaning contact lenses - by treatment with protease and endo-proteinase lys-C

DC - D22 D16 P34 P81

PA - (GENE-) GENENCOR INC

IN - LAD PJ, WOODHOUSE LR

NP - 4

PN - EP-257821-A 88.03.02 (8809)

AU87762880A 88.02.04 (8813)

-----> US4749511-A 88.06.07 (8825)

563158528-A 88.07.01 (8832) éJPè

LA - E

DS - CH DE FR GB IT LI NL SE

CT - (E)GB2083477

PR - 86.07.31 86US-892528

AP - 87. 87. 29 87EP-306721 86.07.31 86US-892528 87. 07. 29

87JP-189985

IC - A61L-002/18 C11D-003/38 G02C-013/00 C11D-007/42 D06M-016/00

G02C-007/04

MC - D09-C01A

AB - (EP-257821)

Contact lenses are cleaned by treatment with a protease (I) and endoproteinase lys-C (II). Specifically, (I) is an oxidn.-resistant form of subtilisin. (II) is derived from *Lysobacter enzymogenes*. The lenses may be treated simultaneously with (I) and (II), or pretreated with (II) and then treated with (I). The concn. of (I) and (II) is 0.1-20 mcg/ml. The lenses may also be treated with a disulphide-cleaving reagent, e.g. 2-mercaptoethanol.

ADVANTAGE - (II) renders lysozyme (the main protein component of tear fluid) more susceptible to attack by (I), thus inhibiting the opacifying effect of lysozyme on soft contact lenses. (9pp Dwg.No.0/0)

-2-

AN - 87-144798/21

XRAM- C87-060339

XRPX- N87-108624

TI - Hydrogen peroxide disinfecting of medical devices - using immobilised protein, esp. enzyme, to decompose excess peroxide

DC - A96 D22 E36 D16 P34 R16

PA - (MINN ) MINNESOTA MINING MFG CO  
IN - HENDRICKSO CE,MENCKE AJ  
NP - 7  
PN - EP-223479-A 87.65.27 (8721)  
362114557-A 87.05.26 (8726) éJPè  
AU8664243-A 87.05.14 (8726)

-----> US4757014-A 88.67.12 (8830)

US4829001-A 89.05.09 (8922)  
US4855234-A 89.08.08 (8939)  
CA1261254-A 89.89.26 (8945)

LA - E

DS - DE FR GB IT SE

CT - (E)No-SR. Pub A3...8827 W08607264 EP-209071 US4025667  
FR2883867 FR2573772 EP--3923 EP--46613 US4210722 1.Jnl.Ref

PR - 85.11.08 85US-796274 85.11.08 85US-796272 88.03.17  
88US-169832

AP - 86.11.03 86EP-308559 86.11.07 86JP-265412 85.11.88  
85US-796274 85.11.08 85US-796272

IC - A61L-002/18 C07K-017/08 C12N-011/08 G01N-033/54

MC - A11-C A12-V02 A12-V03 D09-A01A E31-E

AB - (EP-223479)

Method for disinfecting a medical device comprises: (a) immersing the device in H2O2 soln. for sufficient time to disinfect; and (b) decomposing any residual H2O2 using a catalytically effective amt. of a protein (I) capable of decomposing H2O2, (I) being immobilised on a composite article (II) comprising, in sequence: (i) a support, opt. surface modified to provide binding sites for a protein immobiliser cpd. (II); (ii) a layer of (III); and (iii) a biologically active protein (I). Composite (II) is claimed where the support is fibrous and surface modified. Also claimed is a kit for use as above and comprising H2O3 soln. and a composite (II) where (I) is catalase or peroxidase, in a single or in separate packages.

USE/ADVANTAGE - Useful for disinfecting medical and dental instruments, surgical staples, implants and esp. contact lenses. Excess H2O2 is decomposed, making the device safe for use in or on the body, and the immobilised enzyme may be present in the same package, simplifying use, by suitable choice of enzyme amts. or by forming into a slow release prepn.

The kit may be re-usable or disposable. (18pp Dwg.No.0/0)

-3-

AN - 86-196878/30

XRAM- C86-084924

XRPX- N86-147098

TI - Aq. contact lens cleaning soins - comprises anionic or nonionic surfactant, and proteolytic enzyme

DC - A96 D22 E17 D16 D25 P81

PA - (ALCO-) ALCON LABS INC

IN - SCHAFFER R

NP - 5

PN - W08604083-A 86.07.17 (8630)

----> US4609493-A 86.09.02 (8638)

AU8653082-A 86.87.29 (8641)

EP-207144-A 87.01.07 (8701)

562501651-W 87.07.02 (8732) éJPè

LA - E

DS - \*AU \*JP AT BE CH DE FR GB IT LU NL SE AT BE CH DE FR GB IT LI  
LU NL SE

CT - (E)GB2088581 DE2854278 US4448662 US4285738 US3918296 358864383  
357848712 553125412 DE3328348 US3882836 (E)GB2088581 DE2854278  
US4448662 US4285738 US3918296 J50064303 557848712 553125412  
DE3328348 US3882836

PR - 84.12.28 84US-687275

AP - 85.12.24 85W0-U04083 84.12.28 84US-687275 85.12.24

85EP-900549 86.00.00 86JP-500298

IC - C11D-001/06 C11D-003/33 G02C-007/04 G02C-013/00 C11D-010/02

MC - A05-H03 A05-H04 A10-E08A A12-V03C1 D11-A01A D11-A03A1 D11-D01C  
D11-De7 E10-C04D E10-E04H

AB - (W08604083)

Aq. contact lens cleaning compsn. comprises as surfactant a nonionic surfactant  $\text{HO}(\text{CH}_2\text{CH}_2)_x - (\text{CHMeCH}_2)_y - (\text{CH}_2\text{CH}_2)_x\text{H}$  (I) or an anionic surfactant  $\text{RO} - (\text{CH}_2\text{CH}_2)_z\text{CH}_2\text{CO}_2\text{H}$  (II), together with a proteolytic enzyme (III). In the formulae,  $y = 10-50$ ;  $x = 5-20$ ;  $z = 1-25$ ;  $R = 8-18\text{C}$  hydrocarbon.

Prof. compsn. contains (w/v) 0.01-5 (0.05-1)% (III), 8.82-1 (0.2-0.6)% (I) and (II), and pref. also 0.005-0.5 (0.05-0.2)% of a Ca chelating agent (pref. a polycarboxylic acid, esp. citric acid or EDTA), and 8.82-1 (0.2-0.6)% urea. Prof. compsns. are aq. solns. contg. 0.03-7.5 (0.25-2.4)% of the solid mixt.

ADVANTAGE - The soln. is economical, convenient and efficient. The soln. is capable of removing both surface and sub-surface deposits of proteins, mucins, lipids, and Ca from soft contact lenses. (20pp Dwg.No.0/0)

-4-

AN - 85-117757/20

XRAM- C85-050898

**XRPX- N85-088596**

**TI - Cleaning-disinfecting contact lenses - by heating in aq. proteolytic enzyme soln.**

**DC - D22 D16 P34 P81 P43**

**PA - (BAUL ) BAUSCH & LDMB INC**

**IN - OGUNBIYI L, RIEDHAMMER TM, SMITH FX**

**NP - 11**

**PN - EP- 141607-A 85. 85. 15 (8520)**

**AU8434541-A 85. 95. 82 (8525)**

**N08404223-A 85. 05. 20 (8527)**

**360121416-A 85. 06. 28 (8532) éJPè**

**DK8405077-A 85. 04. 25 (8533)**

**ES8601493-A 86. 02. 16 (8618)**

**-----> US4614549-A 86. 09. 30 (8642)**

**CA1231070-A 88. 01. 05 (8885)**

**EP- 141607-B 88. 12. 21 (8851)**

**DE34757330G 89. 01. 26 (8985)**

**J89001773-B 89. 01. 12 (8906) éJPè**

**LA - E**

**DS - AT BE CH DE FR GB IT LI LU NL SE AT BE CH DE FR GB IT LI LU NL SE**

**CT - (E)No-SR-Pub A3. . 8526 EP- - - 5131 US3598121 (E)EP---5131 us3598121**

**PR - 83. 10. 24 83US-545314 85. 01. 09 85US-690364**

**AP - 84. 10. 22 84EP-307264 84. 10. 23 84JP-221373 84. 10. 23**

**84ES-537002 85. 01. 09 85US-690364 84. 10. 22 84EP-307264**

**84. 10. 23 84JP-221373**

**IC - A61L-002/18 C11D-007/42 G02C-013/00 C11D-000/00 G02C-007/04**

**C11D-000/00 G02C-000/00 B08B-003/10**

**MC - D05-A02 D09-A01 D09-C01**

**AB - (EP-141687)**

The known cleaning of contact lenses by sequential soaking in aq. proteolytic enzyme (1) then disinfecting is improved by heating the lenses in the aq. (1) soln. to achieve cleaning and disinfection in one step.

Uses as (1), papain, pancreatin, or a protease (amylase etc.) which may be derived from Bacillus, Streptomyces, or Aspergillus spp. (esp. B. subtilis). Heating is to 60-100 esp. 65-85 deg. C for up to 60 min. Prepfd. solns. also contain tonicity agents (esp. made isotonic), 0.00001-0.5 % esp. 0.0001-0.1% by wt. preservatives (thimerosal, sorbic acid, etc.), 0.1-2% by wt. chelating agents (esp. EDTA or di-Na salt), 0.5-2.5% esp. 0.1-1.5% by wt. buffers, and up to 15% by wt. surfactants (polycarbonate 20, polyoxyethylene (4) stearate etc.).

**USE/ADVANTAGE - The process allows the single-step**

cleaning and disinfecting or hard, soft, gas-permeable, and textured-wear lenses. The solns. used are effective, safe and non-toxic. (15pp Dwg.N .0/0)

#### 4.4.1.2 SS2 dans base WPI

-1-

AN - 89-370170/50

XRPX- N89-281798

TI - Soft contact lens cleaning appts. - has device for providing DC charge to substance that when immersed in fluid will transmit electric charge

DC - se5 P43

PA - (PANK/) PANKOW M L

IN - PANKOW M L

NP - 1

PN - US4872965-A 89.10.10 (8950)

LA - E

PR - 88.02.10 88US-154790

AP - 88.02.10 88US-154790

IC - B08B-003/12 B08B-006/00 B08B-007/00

MC - S05-G

AB - (US4872965)

The soft contact lenses cleaning apparatus comprises a base which is made of non-conducting material for holding a lens, a fluid capable of transmitting the electrical charge, a device for containing the electrical charge transmitting fluid and means for providing a DC charge and creating separate electrical poles. Another device is composed of a substance which when immersed with the electrical charge transmitting fluid is capable of transmitting the electrical charge to one lens for the purpose of cleaning the lens.

When in operation, deposits or contaminants will migrate from the surface of the contact lens (24) or from within the body of the contact lens and onto the surface of the contact lens (24) or onto the transmission device (28). The deposits or contaminants may be removed from the surface of the contact lens such as by washing the surface of the contact lens. Where deposits or contaminants migrate on to the surface of the transmission device they may be washed.

ADVANTAGE - Prevents damage to lens. (7pp Dwg.No.3/3)

-2-

AN - 89-220116/30

XRAM- C89-097821

TI - Cleaning of contact lenses - using aq. compsn. contg. carboxy vinyl polymer, and using pptd. material as abrasive

DC - A96 D22

PA - (ALCO-) ALCON LABS 1 NC  
IN. - BHATIA RP  
NP - 1  
PN - US4839882-A 89.06.13 (8938)  
LA - E  
PR - 86.12.24 86US-946343 88.08.04 88US-229766  
AP - 88.08.04 88US-229766  
IC - C11D-003/14  
MC - A04-A03 A04-E04 A11-C04 A12-V02A A12-W12B D11-D01C A04-F04  
AO - (US4839882)

In the cleaning process (a) a small amt. of an aq. cleaning soln. comprising H<sub>2</sub>O and a carboxy vinyl polymer (1) (mol. wt. (1-6) x 18 power 6 is applied to the lens; (b) the compsn. is rubbed over the lens surface to form an abrasive ppte. of (1); and (c) the lens is rinsed to resolubilise the abrasive (1) ppte. and remove the remaining compsn. and debris from the surface of the lens.

ADVANTAGE - The method completely and rapidly removes deposits of proteins, lipids and other materials from contact lenses. (3pp Dwg.No.0/0)

-3-

AN - 89-177796/24  
XRAM- C89-078521  
XRPX- N89-135814  
TI - Peroxide and heat stable protease soft contact lens cleaning soln. - for elevated temp. use giving simultaneous cleaning and disinfection of lens surfaces  
DC - D22 E36 D16 P43  
PA - (DART-) DARTMOUTH COLLEGE  
IN - TRUMPOWER BL  
NP - 1  
PN - US4832754-A 89.05.23 (8924)  
LA - E  
PR - 87.01.38 87US-008744 84.05.04 84US-606942  
AP - 87.01.30 87US-008744  
IC - B08B-003/10  
MC - D09-A02 E31-E  
AB - (US4832754)

Soft contact lenses are cleaned and disinfected by (i) immersing in an H<sub>2</sub>O<sub>2</sub> soln. contg. a thermophilic protease which is stable and active at over 70 deg. C; (ii) heating to above 78 deg. C and (iii) maintaining this temp. until the protease has hydrolysed and removed protein and simultaneously disinfected the lenses.

ADVANTAGE - Prior art proteases would denature and deposit on lens surfaces at 78-80 deg. C. Chemical disinfection leaves irritants in the lenses which are not completely removed

by rinsing in saline. Use of heat stable protease avoids both problems. (8pp Dwg.No. 0/5)

-4-

AN - 89-084902/11

XR - 86-196542

XRAM- C89-037670

XRPX- N89-064826

TI - Cleaning contact lens - using compns. contg. polyether carboxylic acid surfactant

DC - A96 D22 E17 P43

PA - (ALCO-) ALCON LABS INC

IN - SCHAFFER D, SCHAFFER R

NP - 1

PN - US4808239-A 89.02.28 (8911)

LA - E

PR - 87.85.22 87US-053982 84.12.28 84US-687274 86.07.07  
86US-882671

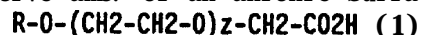
AP - 87.85.22 87US-053982

IC - B08B-007/00 C11D-001/08

MC - A10-E08A A12-V A12-V02A A12-W12 D11-A01A1D11-D01CE10-C04D4

AB - (US4808239)

A method of cleaning a contact lens comprises applying to the contact lens an aq. contact lens cleaning compsn. comprising an effective amt. of an anionic surfactant of formula (I).



where R = 8-18C hydrocarbon z = 1 to 25; pref. R = 12C hydrocarbon z = 10.

USE/ADVANTAGE - The compsn. removes protein, lipid and calcium deposits from contact lenses (esp. soft contact lenses) surfaces and subsurfaces. The compsns. are also adapted for cleaning contact lenses directly in the eye and the compsn. can be formulated as isotonic or hypotonic solns. (4pp Dwg.No.0/0)

-5-

AN - 88-213534/31

XRAM- C88-095255

XRPX- N88-162830

TI - Tablet for cleaning, disinfecting contact lens - with slowly dissolving core of neutralising agent and rapidly dissolving casing of active material

DC - D22 P81 P34

PA - (HENK ) HENKEL KGAA

IN - KRUSE H, JACOBS J, WISOTZKI KD, THUL J

NP - 4

PN - DE3701129-A 88.07.28 (8831)

EP-278224-A 88.08.17 (8833)

-----> US4767559-A 88.08.30 (8837)

563199799-A 88.08.18 (8839) éJPè

LA - G; E

DS - AT BE CH DE ES FR GB GR IT LI LU NL SE

CT - (G)DE3329922 EP--82798 GB2117534 US3884826

PR - 87.01.16 87DE-701129

AP - 87.01.16 87DE-701129 88.01.09 88EP-100215 87.04.08  
87US-035848 88.01.14 88JP-007384

IC - C11D-007/38 C11D-017/00 G02C-013/00 A61L-002/00

MC - D05-A02C D09-A01 D09-C01A

AB - (DE3701129)

A tablet for diinfecting and cleaning contact lenses is made by (a) forming a stable core mixt., which can be tabletted, forming the neutralising agent, (b) forming a core tablet from this, (c) pref. coating this with a lacquer layer, (d) forming a casing mixt., comprising the actual cleaning agent, and (e) pressing the core tablet into the casing mixt.

In the core mixt., contg. catalase, NaHCO<sub>3</sub> and opt. other substances, the NaHCO<sub>3</sub> and opt. other additives are coated with a dense, acid or neutral, lacquer layer, sol. under the conditions of use.

ADVANTAGE - In use, the outer layer dissolves in water and acts on the contact lens in the soln. The core dissolves more slowly, and neutralises excess cleaning agent. The active materials dissolve in the required order, there is no loss of activity due to premature reaction, and the cleaned lens does not irritate the eye, indicating complete neutralisation of the disinfectant agent. (5pp Dwg. No 0/0)

-6-

AN - 88-141719/21

XRAM- C88-063083

XRPX- N88-108229

TI - Cleaning and disinfecting endoscope - in soln. contg. disinfectant soln. comprising aldehyde and complex former, and drying in hot sterile air

DC - A97 D22 E19 D16 P34 P31

AW - POLYETHYLENE GLYCOL

PA - (HENK ) HENKEL KGAA

IN - DISCH K, HACHMANN K, BANSEMIR K

NP - 11

PN - EP-268227-A 88.85.25 (8821)

DE36393220A 88.85.26 (8822)

AU87812350A 88.05.19 (8828)

563135123-A 88.06.07 (8828) éJPè

N08704768-A 88.06.13 (8829)

BR8706157-A 88.86.21 (8838)

DK8705940-A 88.05.18 (8832)  
ZA8708576-A 88.05.17 (8835)  
FI8705049-A 88.05.18 (8835)

----> US4784790-A 88.11.15 (8848)

DE3639322-C 89.08.17 (8933)

LA - G; E

DS - AT BE CH DE ES FR GB GR IT LI LU NL SE

CT - (G)No-SR.Pub A3...8951 DE3327466 EP-118933 EP-117183  
US3697222 EP--28865 EP-158869 EP-224971

PR - 86.11.17 86DE-639322

AP - 87.11.13 87EP-116781 86.11.17 86DE-639322 87.11.17  
87JP-291641 87.11.16 87ZA-008576 87.11.17 87US-121492  
86.11.17 86DE-639322

IC - A61L-002/18 C11D-003/38 A61B-001/00 C11D-001/72 A61L-000/00  
A61B-000/00 A61L-000/00 C11D-000/00

MC - A12-V03C1 D09-A01 D11-A01E D11-A03A1 D11-B02 D11-D01 E10-A07  
E10-A09B4 E10-C02B E10-D01D E10-E04J E10-G02H E10-H01D

AD - (~~-268227)

Endoscopes are cleaned and disinfected by: (a) contacting the surfaces for 1-15 mins. with a soln. at 55-65 deg.C, contg. a low-foam nonionic surfactant, a proteolytic enzyme, a complex-former, and opt. the usual cleaning components, and with pH 6-8, and sepg. the soln., (b) contacting the surfaces for 1-15 mins. with a disinfecting soln. at 55-65 deg.C, contg. HCHO or a 2-8C aliphatic dialdehyde, and a coaplex-former, and with pH 6-8, (c) rinsing at least twice with water with pH 6-8, using water at 55-65 deg.C in at least the last rinse, and (d) drying with sterilised air at 55-65 deg.C. The hardness of the water in (a)-(c) is 3-8 deg. d.

(a) The cleaning soln. contains 8.1-g/l of low-foam surfactant, esp. an alkoxylate or a polyethylene glycol ether, 0.03-0.3 Anson units/l of proteolytic enzyme, and 0.03-0.3 g/l of a complex-former, esp. Na gluconate. (b) The disinfecting soln. contains 0.5-5 g/l of an aldehyde, pref. glutaraldehyde, and 0.02-0.25 g/l of complex-former, esp. a Na salt of phosphonobutane tricarboxylic acid. Opt., before the cleaning soln. in (a) is sepd., it is treated with a disinfecting soln. contg. HCHO or a 2-8C aldehyde and a complex-former, esp. to give a total of 8.25-2.5 g/l of aldehyde and a complex-former, esp. to give a total of 0.25-2.5 g/l of aldehyde and 0.01-0.13 g/l of complex-former. The endoscope may be subjected to ultrasonic vibration in stages (a) and/or (b), and may be rinsed with water with pH 6-8 between these stages. (d) The air is sterilised with a aicrofilter.

ADVANTAGE - Cleaning and disinfection is quick, and does not harm the instrument. The process may be automatic. Used

solns. can be released into the waste water. (9pp Dwg.No.0/0)

-7-

AN - 87-144798/21

XRAM- C87-060339

XRPX- N87-108624

TI - Hydrogen peroxide disinfecting of medical devices - using immobilised protein, esp. enzyme, to decompose excess peroxide

DC - A96 D22 E36 D16 P34 R16

PA - (MINN ) MINNESOTA MINING MFG CO

IN - HENDRICKSO CE,MENCKE AJ

NP - 7

PN - EP-223479-A 87.05.27 (8721)

562114557-A 87.05.26 (8726) eJPè

AU8664243-A 87.05.14 (8726)

US4757014-A 88.07.12 (8830)

====> US4829001-A 89.05.09 (8922)

====> US4855234-A 89.08.08 (8939)

CA1261254-A 89.09.26 (8945)

LA - E

DS - DE FR GB IT SE

CT - (E)No-SR. Pub A3...8827 W08607264 EP-289071 US4025667

FR2083867 FR2573772 EP--3923 EP--46613 US4218722 1. Jnl. Ref

PR - 85.11.08 85US-796274 85.11.68 85US-796272 88.03.17

88US-169832

AP - 86.11.03 86EP-308559 86.11.07 86JP-265412 85.11.08

85US-796274 85.11.08 85US-796272

IC - A61L-002/18 C07K-017/08 C12N-011/08 G01N-033/54

MC - A11-C A12-V02 A12-V03 D09-A01A E31-E

AB - (EP-223479)

Method for disinfecting a medical device comprises: (a) immersing the device in H<sub>2</sub>O<sub>2</sub> soln. for sufficient time to disinfect; and (b) decomposing any residual H<sub>2</sub>O<sub>2</sub> using a catalytically effective amt. of a protein (1) capable of decomposing H<sub>2</sub>O<sub>2</sub>, (1) being immobilised on a composite article (II) comprising, in sequence: (i) a support, opt. surface modified to provide binding sites for a protein immobiliser cpd. (II); (ii) a layer of (III); and (iii) a biologically active protein (1). Composite (II) is claimed where the support is fibrous and surface modified. Also claimed is a kit for use as above and comprising H<sub>2</sub>O<sub>2</sub> soln. and a composite (II) where (1) is catalase or peroxidase, in a single or in separate packages.

USE/ADVANTAGE - Useful for disinfecting medical and dental instruments, surgical staples, implants and esp. contact lenses. Excess H<sub>2</sub>O<sub>2</sub> is decomposed, making the device safe for

use in or on the body, and the immobilised enzyme may be present in the same package, simplifying use, by suitable choice of enzyme amts. or by forming into a slow release prepn.

The kit may be re-usable or disposable. (18pp Dwg.No.0/0)

-8-

AN - 87-041049/06

XRAM C87-017535

XRPX- N87-031143

TI - Cleaning agent for contact lenses - contg. at least one glycosidase e.g. amylase and at least one activating agent e.g. urea

DC - DZ2 E19 D16 E37 P81

PA - (T0C0-) TOYO CONTACT-LENSE; (LLOY) LION CORP

NP - 2

PN - 562080913-A 87.01.06 (8786) éJPê

====> US4710313-A 87.12.91 (8758)

LA - E

PR - 85.06.26 85JP-139799

AP - 85.06.26 85JP-139799 86.86.12 86US-873351

IC - C11D-007/60 G02C-007/04 G02C-013/00

MC - D05-A02C D09-A01A D09-C01A E10-A13A E10-A13B E10-A17 E10-B02D5  
E10-B02D6 E31-Q07 E33-C

AB - (562880913)

Cleaning agent for washing contact lenses contains at least one species of glycosidase (A) selected from amylase, cellulase, pectinase, hemicellulase, alginase, heparinase, and dextrinase, and at least one activating agent (B) selected from urea, thiourea, guanidinate, reducing agents and amino acids.

Content of glycosidase (A) is 0.005-10 wt.%, pref. 0.05-5 wt.%, and that of activating agent (B) is 0.01-20 wt.%, pref. 0.1-10 wt.%. In case of reducing agent, sulphite salts, hydrogen borate salts etc. can be used.

Amino acids e.g. DL-aspartic acid, L-glutamic acid, glycine, DL-alanine etc. can be used. Cleaning agent may further contain buffer component e.g. sodium citrate, boric acid or mixed phosphate buffer etc., and stabiliser e.g. sodium bicarbonate, content of these together is in range between about 0.001-2.5 wt.%.

USE - Stain of contact lenses comprising proteins, lipids and mucopolysaccharides can be readily removed in short period of time. (4pp Dwg.No.0/0)



# *Quatrième Partie*

## **L'exploitation systématique des bases de données : des analyses stratégiques pour l'entreprise (1)**

*Chantal Bédécarrax, Charles Huot, Luc  
Quoniam, Hewé Rostaing et William Nivol*

Journées d'étude **ADEST**, Paris, juin 1992



# L'exploitation systématique des bases de données: des analyses stratégiques pour l'entreprise

Chantal **BEDECARRAX**, Charles HUOT  
Centre Européen de Mathématiques Appliqués **IBM**<sup>26</sup>

Hervé ROSTAING, Luc QUONIAM, William NIVOL  
Centre de Recherche Rétrospective de Marseille<sup>27</sup>

## 5.1 Introduction

Aujourd'hui tout industriel se pose des questions sur l'environnement scientifique, technologique et concurrentiel de son entreprise. Pour répondre à ces questions il met en place une structure dite de veille technologique ou de veille stratégique. Outre une vigilance de tous les instants cette structure doit répondre à des questions assez générales sur l'état de la concurrence dans un domaine, établir des cartographies d'une technologie ou encore analyser des évolutions à travers le temps. Ces travaux synthétisent le maximum de connaissance que l'on peut avoir sur un thème à un moment précis. L'une des matières premières permettant d'effectuer de telles analyses est l'information contenue dans les bases de données scientifiques et techniques.

L'objectif du travail que nous présentons ici est de répondre, par l'utilisation de la méthode de classification relationnelle, aux questions que se posent les industriels sur la prise en compte de leur environnement dans sa globalité.

Nous décrirons le minerai, ainsi que les divers traitements que nous lui ferons subir pour le transformer en un produit **raffiné** à haute valeur ajoutée. Ce dernier sera nécessaire au bon fonctionnement de l'entreprise en lui apportant un plus inestimable dans la compréhension de l'environnement présent et lui permettra de préparer l'avenir en toute connaissance de cause.

Pour illustrer cette méthodologie nous avons choisi d'analyser le domaine des «Systèmes Thérapeutiques Transdermiques à base de **Patches**» (STTP).

---

<sup>26</sup> CEMAP IBM, 68-76 quai de la **Rapée**, 75592 Paris CEDEX 12, Tel: **40.01.57.11/40.01.53.37**, Fax: 49.28.08.60

<sup>27</sup> CRRM, Université Marseille St **Jérôme**, 13397 Marseille CEDEX 13 Tel: 91.28.86.77, Fax: 91.28.30.80

Les bases de données accessibles en ligne représentent la première et la plus importante source d'information scientifique et technique analysable par des systèmes automatiques. Une référence signalétique dans une base de données contient diverses informations réparties en plusieurs rubriques. Ces rubriques, nommées champs, ont des portées significatives différentes. Il est rare que la richesse de cette diversité d'informations soit pleinement exploitée dans les études bibliométriques de corpus de références. La raison de cette omission est due à la complexité des relations engendrées par toutes les combinaisons de ces informations. L'étude exposée dans cet article traite de la complémentarité d'informations apportées par la prise en compte simultanée de trois champs de la base brevets de Derwent. Ces trois champs correspondent à trois codifications documentaires indexées parallèlement par les producteurs Derwent. Tout brevet **référéncé** dans la base voit son contenu, au sens informationnel, qualifié selon les trois types de codes. Une **codification** documentaire découpe les domaines scientifiques en sections, sous-sections, classes, sous-classes, groupes, sous-groupes... A une **référence** de brevet dans la base Derwent est donc affectée, dans plusieurs champs, **différents** codes qui décrivent les thèmes abordés par l'invention.

Les trois champs "codes" que l'étude prend en considération sont les champs:

1. DC (Dexwent Codes) : Codification établie par **Derwent**
2. MC (Manuel Codes) : **Codification** établie par Dexwent
3. IC (International Patent Classification) : Codification établie par l'Office Européen des Brevets.

Pour estimer l'apport de chacune de ces codifications, nous allons les confronter autour d'un thème bien ciblé. Le choix (arbitraire) de ce thème s'est porté sur une technologie charnière entre la médecine et la pharmacie: les systèmes thérapeutiques transdermiques sous forme de **patch**. Le STTP est un système adhésif qui assure une absorption contrôlée d'un principe actif par voie transdermique. Les méthodes **d'Analyse** de Données Relationnelles vont nous permettre d'évaluer les complémentarités de ces trois modes de codification pour l'ensemble des références concernant **les** STTP. Ces méthodes font partie du **panel** des analyses statistiques qui ont été élaborées pour mieux cerner les phénomènes complexes.

## 5.2 Constitution du corpus de références

Cette étape, la première dans toute analyse bibliométrique, est certainement celle qui influence le plus la validité des résultats. Il est donc indispensable de constituer un ensemble homogène de références tout en assurant une parfaite couverture du sujet. Le corpus extrait doit, autant que possible, être **suffisamment** large pour couvrir le thème étudié et suffisamment restreint pour présenter un “bruit” aussi faible que possible. La stratégie d’interrogation est affinée selon une méthode itérative de type “coup de sonde”. Chaque itération permet, après lecture des échantillons de références, de dégager de nouvelles pistes pour enrichir la stratégie. Les itérations sont répétées tant que les échantillons, obtenus par les croisements des nouvelles pistes, laissent apparaître des références non pertinentes. Les échantillons ont été estimés pertinents pour la stratégie d’interrogation suivante:

QUESTION 1: PATCH ou PATCHES <b>ou</b> PATCHS	<b>(2106)</b>
QUESTION 2 : 1 et TRANSDERM:	(103)
QUESTION 3 : <b>1</b> et THERAPEUTIC:	<b>(36)</b>
QUESTION 4 : 1 et <b>PERCUTANE:</b>	<b>(15)</b>
QUESTION 5 : 1 et (DRUG ou DRUGS)	(102)
QUESTION 6 : 1 et <b>MEDICIN:</b>	(14)
QUESTION 7 : <b>2 ou 3 ou 4 ou 5 ou 6</b>	<b>(160)</b>
QUESTION 8 : 7 <b>sans (CARDIAC</b> à coté de PATCHff)	(159)
QUESTION 9 : 8 <b>sans (VASCULAR</b> à coté de GRAFTf)	<b>(158)</b>
QUESTION 10: 9 <b>sans (PATCHf à coté de GRAFT</b> )	<b>(156)</b>
QUESTION 11: 10 <b>sans (FASTENING</b> à coté de PATCHff)	<b>(155)</b>
QUESTION 12: 11 <b>sans (CARRY à coté de PATCH</b> )	<b>(155)</b>
QUESTION 13: 12 <b>sans (CARRIES à coté de PATCHff)</b>	<b>(154)</b>
QUESTION 14: 13 <b>sans PROTHES:</b>	<b>(148)</b>
QUESTION 15: 14 <b>sans CAMERA</b>	(1471)
QUESTION 16: 15 <b>sans (TEST</b> à coté de PATCHff)	<b>(146)</b>

Remarque: les questions sont posées sur l’**Index** de Base de la base Detwent, c’est-à-dire sur les champs: Résumé, Résumé équivalent, Titre et Titre normalisé Denvent.

L’examen de cette stratégie finale montre que la simple utilisation des **termes** “PATCH” et “THERAPEUTIC” ne nous permettait pas de couvrir la totalité des brevets du domaine. Par contre l’emploi d’autres termes, pour élargir la recherche, laissait apparaître des références hors-sujet car le terme PATCH recouvre des significations multiples et variées (rustine, greffon, pièce de prothèse, chute de **film**, test d’allergie...). Les brevets qui font référence à ce terme dans un sens autre que celui recherché ont donc été “désélectionnés”. Cette stratégie d’interrogation a permis de dégager un corpus de 146 brevets sur le thème choisi. Le téléchargement de ces **références** a été réalisé non seulement pour les trois champs des codifications mais aussi pour tous les champs qui **four-**

nissent des renseignements pouvant aider à la compréhension des résultats des analyses automatiques.

### 5.3 Les premiers résultats

Avant de s'engager plus en avant, il est bon de montrer quelles informations nous pouvons extraire des données dont nous disposons déjà. En effet outre les champs présentés au paragraphe précédent (champs de codes), il existe dans chaque référence de brevet des champs contenant des informations sur l'année de dépôt du brevet, le pays et la société déposante. En se situant sur un plan global, il est possible d'étudier l'évolution des dépôts de brevets en fonction du temps, ce qui permet de situer le niveau de maturité du sujet sur lequel on travaille. De même nous pouvons établir la répartition du nombre de dépôts prioritaires dans le monde ce qui permet de dégager les pays leaders dans le domaine étudié. **Enfin** il est intéressant de voir, au niveau des sociétés, quelles sont celles qui déposent le plus de brevets autour du thème de l'étude. L'obtention de ces informations n'est pas difficile une fois que l'on est en possession du corpus de base. Il ne s'agit en effet dans cette phase que d'analyser les distributions de certains champs sélectionnés.

### 5.4 Construction des tableaux à analyser

Pour confronter les différentes informations apportées par chacune des **codifications**, nous allons reproduire leurs interactions par la construction de tableaux. De tels tableaux sont nommés "matrices". Selon les croisements des éléments dans ces matrices et selon les méthodes d'analyse employées nous essayerons de dégager pour ces trois codifications:

- les recouvrements de sens,
- les complémentarités de sens,
- les qualités discriminantes (**taille** et homogénéité des groupes obtenus par l'analyse).

#### 5.4.1 Choix des niveaux hiérarchiques

Les trois codifications brevets possèdent leur propre hiérarchie de codes. Un code est assimilable à un **chemin** pris parmi les branches de la hiérarchie. Dans cette hiérarchie, les branches sont réparties d'un niveau de signification très large à des niveaux de signification de plus en plus fins. Plus on descend dans les branches de la hiérarchie, plus le code a une représentation complexe et plus son sens est pointu.

La codification Derwent (Champ DC): deux niveaux de hiérarchie (3 digits = un caractère alphabétique + 2 caractères numériques)

Les manuels codes Derwent (Champ MC): 6 niveaux (section, sous-section, groupe, sous-groupe, division, sous-division)

La codification CIB (Champ IC): 7 niveaux (section, sous-section, classe, sous-classe, groupe, sous-groupe, subdivision).

Il est important de mentionner que tous les codes rencontrés dans les références ne sont pas systématiquement renseignés jusqu'au dernier niveau de la hiérarchie. Sur notre corpus si 66 % des codes CIB sont renseignés au dernier niveau, seulement 12 % des Manuels Codes le sont. Mais parallèlement, plus le niveau hiérarchique considéré est fin, plus la diversité des codes augmente. On comprend donc l'importance du choix des niveaux des codes pour satisfaire le compromis entre la perte d'information et le gain en signification.

#### **5.4.2 Construction des matrices de présence/absence**

Le point d'entrée de toute méthode d'analyse des données est un tableau croisant un ensemble *d'individus*, noté *I*, et un ensemble de *variables* descriptives, noté *3*. Les individus sont naturellement ici les 146 références de brevets, mais en ce qui concerne les variables descriptives, un choix préalable s'impose.

Rappelons en effet que nous avons à notre disposition trois types de codifications décrivant le contenu des brevets du corpus: les Derwent Codes, les Manuels Codes et les codes de la Classification Internationale, et que par ailleurs les MC et les IC se hiérarchisent à leur tour en sous rubriques comme il l'a été expliqué plus haut.

La première étape des traitements a donc consisté en la confrontation des ces différents ensembles de descripteurs. Ce travail préliminaire, basé sur des calculs statistiques tout à fait classiques (distributions, moyennes, indices d'associations, tests d'indépendance) nous a conduit à la conclusion suivante :

aucun des 3 modes de codification ne semble prévaloir en termes qualitatifs ou en termes quantitatifs, chacun apporte sa part d'informations sur le contenu des **références** et rien ne nous autorise, a priori, à favoriser l'un ou l'autre des trois types.

Nous avons donc opté pour un ensemble de descripteurs combinant les sources d'informations. Pour les DC la codification est unique, en revanche, en ce qui concerne les IC et les MC nous avons dû sélectionner un niveau de troncature des codes. Notre choix s'est porté sur les IC à 7 digits et sur les MC à 3 digits, tant pour des considérations de cohérence que pour garantir un compromis entre la précision des codes et la quantité des informations disponibles.

Finalement l'ensemble  $J$  que nous avons retenu pour décrire les informations contenues dans notre corpus est :  $J = DC \cup MC3 \cup IC7$ , qui comporte au total 197, répartis en 52 DC, 51 MC3 et 94 IC7.

Nous pouvons maintenant construire le tableau des données de départ  $I \times J$  en faisant figurer la valeur 1 à l'intersection de la ligne  $i$  et de la colonne  $j$  si le code  $j$  figure dans la **référence** du brevet  $i$  et la valeur 0 dans le cas contraire. Ce tableau est la simple restitution des données élémentaires contenues dans les **références**, rien n'est omis, ni ajouté, ni transformé.

Il est à la base de tout traitement mais ses lignes (références de brevets) d'une part et ses colonnes (codes descripteurs) d'autre part peuvent constituer les points d'entrées de deux types d'analyses.

## 5.5 La classification des brevets

---

On va chercher ici à dégager des classes de brevets qui s'apparentent par les codes descripteurs qu'ils ont en commun, autrement dit des familles de brevets caractérisées par leurs similitudes en termes de technologies partagées.

C'est sur ces familles de brevets que l'on peut dans un deuxième temps greffer des analyses répondant à des préoccupations générales d'évolution chronologique, de positionnement des sociétés les unes par rapport aux autres, de stratégies d'extensions internationales ou encore de réseau d'inventeurs.

La technique que nous allons mettre en oeuvre pour construire la partition des brevets en classes disjointes s'inscrit dans le cadre méthodologique général de l'**Analyse Relationnelle**. Nous avons pris l'option, dans ce résumé, de bannir toute formule mathématique et de réduire au minimum les explications et les justifications méthodologiques. Tous les détails figureront dans l'article qui sera publié à l'issue des journées d'étude **ADEST**.

Nous nous contenterons de parler, ici, du point fondamental qui préside au déroulement du traitement, à savoir : le critère de classification qui mesure les similarité entre les objets que l'on compare. C'est incontestablement la question qu'il faut à tout prix se poser pour appliquer la procédure de classification dans un cadre clairement défini. Faute de quoi, on ne sait pas, a posteriori, expliquer la structure que l'on a mis en évidence.

Les spécialistes de l'infométrie, de la scientométrie, de la documentation ou encore de la lexicographie mathématique, ont depuis fort longtemps mis en évidence et étudié les caractéristiques des distributions que l'on rencontre dans ces domaines (lois de Zipf, Bradford ou Lotka).

Les données extraites de télédownload telles que celles dont nous disposons sont de cette nature. Ainsi, un certain nombre de codes se retrouvent dans la grande majorité des références (ce sont les thèmes qui ont présidé à la construction du corpus), d'autres apparaissent de façon moins systématique et d'autres enfin ne figurent que dans un faible, voire très faible, nombre de références. Il importe que le critère de classification qui va permettre de mesurer les ressemblances entre brevets tienne compte de ces phénomènes. Notre choix s'est porté sur le critère de Burt pondéré, qui outre ses bonnes propriétés axiomatiques, se base sur un indice dit de **présence-rareté** répondant parfaitement à notre problème. En effet, ce critère a pour effet d'attribuer aux codes un poids inversement proportionnel à leur présence dans le corpus. Ainsi, deux brevets qui partagent un code rare seront "plus similaires" que deux brevets qui auraient en commun un code présent dans de nombreuses références.

Les classes de brevets issues de la classification pilotée par le critère de Burt pondéré sont donc formées de brevets qui se ressemblent non seulement parce qu'ils partagent les mêmes codes mais encore parce que ces codes les caractérisent tout particulièrement. On garantit le découpage du corpus en familles de brevets homogènes et discriminantes. L'analyse du résultat permet en outre d'expliquer chacune de ces classes en fonction des codes ou groupes de codes qui ont présidé à leur création, autrement dit d'attacher à chacune d'elle une étiquette synthétique résumant les thèmes technologiques qu'elle recouvre.

Cette classification, déjà en **soi** intéressante, gagne encore à être combinée à d'autres données pour dégager de l'information stratégique. Ainsi, la partition des brevets peut être confrontée à celle qui est fournie par les années de dépôt des brevets. On peut alors faire émerger des phénomènes d'évolution dans le temps sur les familles de brevets et voir se déplacer les centres d'intérêt au **fil** du temps.

On a également la possibilité de mettre en regard de chacun des brevets sa (ou ses) société(s) **déposante(s)** afin d'étudier la répartition de ces sociétés dans les familles de brevets. On peut ainsi mesurer la variété ou au contraire la concentration du portefeuille de brevet par société, dégager des apparentements, a priori non évidents, entre sociétés et naturellement surveiller le positionnement relatif de sa société vis à vis de concurrents.

Ces différents traitements, qui relèvent de techniques tout à fait classiques de mesures d'associations entre partitions, ne sont envisageables que parce qu'on a réalisé en amont la classification des brevets.

## 5.6 La classification des codes

---

L'objectif de ce traitement est de découper l'ensemble de codes descripteurs en classes homogènes sur la base de leurs cooccurrences dans les **références**. Autrement dit, on cherche ici à dégager les pôles technologiques autour desquels s'articule le corpus.

Cette fois ce sont les colonnes du tableau de départ, c'est-à-dire l'ensemble des codes descripteurs, qui sont soumis au processus de classification.

Les considérations qui avaient induit le choix du critère de classification sur les brevets sont encore valides dans ce nouveau contexte. Il n'est toutefois plus possible d'utiliser le critère de Burt pondéré. Celui-ci aurait en effet pour conséquence de faire jouer un rôle à la richesse de description des brevets. Or le fait qu'un brevet possède un plus ou moins grand nombre de codes ne doit pas être pénalisant. En revanche, il convient toujours de prendre en compte la fréquence d'apparition des codes qui, elle, reste tout à fait pertinente.

Notre choix s'est finalement porté sur le critère de Burt, très couramment utilisé en Analyse Relationnelle pour ses bonnes propriétés de règle d'agrégation.

A l'issue de ce traitement, nous obtenons une partition des codes descripteurs. Chacune des classes de cette partition regroupe un certain nombre de codes qui ont pour caractéristique d'apparaître conjointement dans les **références** de brevets.

On a donc mis en évidence des combinaisons de technologies qui présentent un fort taux de corrélation à l'échelle du corpus étudié. Ces pôles, disjoints les uns des autres par construction, peuvent être reliés en réseau à l'aide de la mesure des liaisons inter classes. On peut ainsi dessiner une cartographie des thèmes technologiques intervenant dans la mise au point des STTP.

Un autre regard sur l'analyse détaillée du résultat montre que l'on a pu établir une forme de correspondance ou de synonymie entre les codes des trois origines (DC, MC3 et IC7). Certaines classes sont formées de codes des trois types, d'autres en revanche n'en regroupent que deux. Outre le fait que l'on justifie, a posteriori, le choix de travailler sur les trois sources d'information, on comprend comment ce type de résultats peut être utilisé à des  **fins**  documentaires pour générer des "noyaux" de codes qui donneront lieu à de nouvelles stratégies d'interrogation.

## **5.7 Conclusion**

---

Nous avons présenté succinctement la chaîne des traitements nécessaires pour analyser en profondeur une information issue des bases de données. Ces traitements nous fournissent, suivant l'orientation choisie, une réponse à l'un des problèmes que l'on se pose. Ainsi la classification des brevets nous apporte des réponses sur la répartition des sociétés **concurrentes** en termes de positionnement dans des niches technologiques spécifiques. S'appuyant sur la même classification, nous **affinons** l'analyse de l'évolution d'une technologie au cours du temps.

La classification des codes, quant à elle, nous permet de déterminer des **thèmes** technologiques majeurs ou **innovants** et de les relier entre eux **afin** de réaliser une carte synthétique des technologies mises en oeuvre. Ces analyses qui combinent des codes de **différentes** natures, permettent de dégager des thèmes qui ne seraient pas apparus si l'on s'était contenté d'un seul type de code descripteur.



# *Cinquième Partie*

**A new method for analysing  
downloaded data for  
strategic decision**

*Charles Huot, Luc Quoniam et Henri Dou*

Scientometrics, Vol 22, n° 2, pp 279-294



# A new method for analysing downloaded data for strategic decision

Ch. HUOT \*,  
L. QUONIAM \*\*, H. DOU \*\*

★ - IBM CEMAP (Centre **Europeen** de **Mathematiques** Appliquees), 68-76 quai de la rapee, 75592 Paris CEDEX 12 (France)

★★ - CRRM **Faculte** de St **Jerome**, 13397 Marseille CEDEX 13 (France)

(Received July 16, 1991)

Technology assessment survey is nowadays a **specific** and **scientific** subject that **any** manufacture needs for increasing **productivity**. This **function** was initially reserved to experts of the studied **field**. But the increase of information volume has called for a change. Now, we need specialists of technology assessment survey which know **about** sophisticated methods to **extract** strategic information from downloaded data. We **will** explain how to **build** strategic information. We present here a new and original method of data analysis. This **Factorial** Relational Analysis is bom **after** 15 years of IBM France mathematics research **center** works on qualitative data analysis. The method is based on Relational Analysis. The **particularity** of this method is to work with sparse matrices and to obtain the best classification without **any** ‘a priori **fixation** of **number** of classes’. Relational Analysis is used in other **sectors** than the analysis of matrices issued from downloaded data. For example it is also used in computational **lexicography** or in **credit** scoring or in **any domain** where classification is concemed. Here we **choose** to present an example of an application in patent analysis.

## 6.1 Introduction

To face the «**technological war**» that has begun all **around** the world, without **any** exempted country, it is crucial for the **chief executives** to be always **very** well **informed** on the few subjects that **could** have great **consequences about** their **decisions** to be always more **competitive**. These **few** subjects have been called by **Rockart [Rock79]** the “Critical Success Factors” (CSF). As soon as defmed, these “CSF must be overlooked by specialists of information retrieval. To do this they need acess to international **specific**

databases. **As** soon as they have **found** the database, they **can** query and obtain a lot of **bibliographic** references that would **provide** the requested information. The problem nowadays is that the amount of available information increases [Dou83] in **such** a rate that a lot of problems occur **during** information retrieval.

-**First** there are problems concerning the construction and use of databases.

-Second, **after** getting information, it is more and more **difficult** to analyze the information with accuracy (the brain ability is rather constant for a continual increase of the amount of information).

In this paper the second point will be developed and an application to a case study **concerning** patents analysis will be treated. We **choose** patents information because of its major importance in manufacture strategy. The **database** used was the WPIL (World Patents Index **Latest**) which **contains** patents **issued after** 1981. The request was **about** the problem of cleaning the contact lens by a chemical process (more specially with **papain** which is a proteolytic enzyme). The exact query formulation was:

«CONTACT LENS AND **PAPAIN** ?»

The **answer** gave 4 patents and we added the 9 patents that **contained one** of the 4 **formers** in their citation **field** (all the patents **provide** information upon the technology and the technologies which use **it**). **The** analysis treats those 13 patents. We chose an application with few patents because of **space** problem but the method we use **could** be applied to a large number of patents (> 1000).

## **6.2 Basic Information**

Once the query is formulated on the host, it is possible to download the data: it **means** that **one** gets the data in your laboratory, on a microcomputer. **Afterwards**, in the laboratory **you may** analyze the **bibliographic** data the way **you** want (Post Processing of **Online** search: PPOS Concept [Dou90b]). The **bibliographic** data **you** get are divided into fields, **each** of them having a **specific mean** as shown on table 1.

There are two ways to look at the references [Dou90c]:

-**To** consider them as **bibliographic** tools and then whatever combination of the **fields you** make, the information will always be a **bibliographic reference**.

-**To** consider them as a sum of **specific** strategic information and in this case combination **and** analysis of constitutive **fields provide** strategical information (like the frequency of

patents owners). Strategic information is a part of information for IHDS [Tela87] (Interactive Helps for **Decision** Systems, **SIAD** in **French**).

**Many** authors working on bibliometrics (reviewed in [Whit89]) and scientometrics **all around** the world already use this concept with **different** methods to analyze separated fields.

In our case study, we developed the Derwent codes field analysis. We chose this field for several reasons:

-Codes are interesting to study because they **provide** short, concise information.

-Figures processing is casier and faster than **word** processing.

-More important, codes are independent from linguistic vogues, almost not related to from period considerations (changes in code signification are uncommon) and independent from **space** considerations (the codes affectation is the **same** for US patents and USSR patents).

-Codes are **quite** “objective” when **you** consider that the **person** who **abstracts** a large number of patents has surely the best overview on a subject.

So they are **very** powerful for quick analysis,

The Derwent code classification is made of 330 codes which **can** be divided into 8 non equal parts (cf table 2).

The detailed exhaustive list of the codes for this analysis is given in the table 4.

### 6.3 Different ways to analyze a specific field

To analyze a **specific** field there are **many** methods that we are going to overview in an increasing **complexity** order.

#### 6.3.1 Frequency analysis

First of **all** it is possible to **count** the frequency of **each** constitutive code. The result concerning our data is given in table 3.

This result **provides** a **scan over** three diierent zones:

1. **Evident information:** **this** is information which is present in almost all **references** (high frequency codes). It just gives information **about** the subject we work on.
2. **Potentially innovative information:** medium frequency codes which **provide** information only present in some **references** and that **could** be specific to some **new** technical **consideration**.

3. Noisy **information**: so low frequency codes that it is impossible to **say** if they are accidental data or it is real innovation. **Usually** this part is the most important but not **in our** sample because of the few patents we used.

In our case we could divide in codes D22 to **P8** 1; E36 to **E19**; E37 to P3 1. It is possible to determine graphically the frontiers between zones. **Many** authors worked on these aspects of information retrieval that could be **named** “Zipf or Bradford or Lotka or Informetrics” distributions [**Eggh89, Lafo90, Broo84**].

### 6.8.2 Pairing techniques

A second aspect of the exploration of a **specific field** is to analyze the relations between the **different constituents**. More than knowing which codes are present, it is important to see the connections between the **different** techniques. It introduces a multidimensional view of the problem. Links **may** be studied in **different ways**. The simplest **one consists** of counting the links and presenting them on a graph as shown in Fig. 1.

The permuted pairs are equivalent. This graph is easy to **construct** but has two major drawbacks: '

- It is not easy to represent the **differences** in pairs frequencies.
- The** pairs frequencies are not relevant to the importance of the pairs.

Let us examine the example in table 5.

In the example the (BC) pair seems to be more important (higher frequency) but in **fact** the (AD) pair is **much** more important (quasi 100 % of the constitutive codes are engaged in the pair formation).

Another inconvenience is the **difficulty** to draw the graph for a **very** large amount of downloaded **references** that leads us to **cut** and draw only high frequency pairs, but we will explain **later** that the low frequency pairs are not inevitably interesting. These inconveniences are largely **balanced** by the easiness and rapidity of the method (few minutes for automatic treatment of 1000 **references** on a micro computer). The results concerning our case study are shown in Fig 1. On the graph, we see the **different** links between codes. We **can** divide these links in several parts:

- Links** between the main frequency codes (that we **defined** as evident information) which are not necessarily interesting (seem to be evident links). They explain the subject we work on.
- Links** between high frequency and medium frequency codes (that we define as innovative information) or between the medium frequency codes and themselves which are relevant to interesting links (strong or not) because of their potential innovative aspects.

-**Links** between **any** frequency codes and low frequency codes which are generally not drawn due to the **fact** that they complicate the graph with potentially noisy information (constitutive codes **may** be noisy).

### 6.3.3 Matrix building

To overcome the drawbacks of pairing analysis, **many** authors build up matrices and **analyze** the result with classical data analysis methods, which **provide** metric solutions.

The results dependant, in part, on the matrix analyzed:

-**«Co-frequencies» matrices**: are matrices that **contain** in **each** row column intersection the of co-frequency of two concepts, **one** is a row header, the other a column header. This kind of matrix is symmetric if row headers are equal to column headers.

-**«Frequencies» matrices**: *are* built with the concepts of column headers and the reference numbers as row headers. If just considering the presence of a concept, the matrix will only **contain 1** for presence and 0 for absence (in our case study, we built up **such** a matrix which is given in the **annex**). This former matrix form is called a "**presence/absence** matrix" and is always used for code analysis (**each** code is present only once by reference).

It is important to notice that a symmetric co-frequencies matrix is **computable** from a frequencies matrix (it is simply a Burt matrix [**Benz80**]). The **bibliographic references** matrices are **very «sparsed»**, i.e. they **contain** few numbers (due to information properties developed when we described codes frequencies). This **consideration** is **very** important to understand the problems in using classical data analysis [**Quon90b**]. Building a matrix must always be an automatic process regarding the amount of data and possible **human** errors.

## 6.4 Classical methods of data analysis

There are two families of data analysis which are complementary and **often** consecutively used.

**1) Inertia analysis**: *they* represent the whole matrix in a reduced **space** obtained by computation of eigenvalues and eigenvectors of a calculated distance matrix. It introduces a simplification of the information but an increase of **significance**. The problem is that the reduced **space** must be small. The sparser the matrix, the more the reduced **space** is similar to the original **one**; so graphs are numerous and analysis **difficult** to understand [**Esco88**] (dispersion concept). In spite of this major restriction several authors use this type of analysis to **provide** strategic information from downloaded data [**Duth90, Doré86**]. Some others overcome this drawbacks by grouping variables [**Paol87**]

but we think it **induces** too **much** important loss of information to be an **efficient** method. **Factorial** analysis is an example of inertia analysis. An other example of **recent** inertia analysis is the quasi-correspondence analysis (QCA) used in [**Leyd87**] described in [**Van 89, Tijs88**]

2) **Classification analysis:** are techniques that classify either rows or columns of a matrix using an aggregation criterion **over** a **computed** distance. They are used separately [**Tijs88**] or as a complementary tool for the interpretation of inertia analysis [**Todo90**] (multidimensional scaling and cluster analysis), [**Duth87**]. In this former case the classification is calculated with the position in the reduced space. At the end, classes repartition is available with a major restriction: **you** must **specify** the amount of classes before standing the process. Ascendant Hierarchical Classification (AHC) is an example.

All the restrictions mentioned, induced the **co-word** approach [**Rip88, Cour90**].

3) **Co-word analysis method:** The authors calculate an AHC but instead of working in a matrix space they use a chained-word notion. They solve the problem of **class** number determination with a **special** representation of the results. According to us the problem seems to be **about** the classification criterion **choice** which is not the most impressive [**Rous90, Cele89**].

## 6.5 Relational Analysis

The mathematic department of the IBM Scientific Center of Paris has been working on a new method of qualitative data classification for 15 years. This method **named** Relational **Analysis** was developed by Marcotorchino and **Michaud** [**Marc79, Marc87, Bede89a**]. Relational Analysis groups together a pull of techniques to modelize and solve problems defined like:

«**Find** a **structured** relation **Y** which is the **closest** to a set of **any** relation **R**»

The method keeps data under a relational form and modelize the problem by using linear programming.

### **6.5.1 Data representation, basic tables**

To build the **different** matrices used by the methodology we define

1.  $N$  = number of **objects**
2.  $A_4$  = number of variables
3.  $P$  = total number of modalities

With these elements we define 3 tables.

#### **6.5.1.1 The complete disjunctive table K**

The table **K** (dimension =  $N \times P$ ) has for general term  $k_{ij}$  with:

$k_{ij} = 1$  if  $i$  «is in relation with»  $j$   
 $k_{ij} = 0$  otherwise

$$\sum_{j=1}^P k_{ij} = M$$

$$\sum_{i=1}^N \sum_{j=1}^P k_{ij} = M \times N$$

Factorial Relational Analysis is a part of the Relational Analysis developed by Marcotorchino **since** 1989 [Marc89]. This method **does** not use the original «*Condorcet Criterion*» but weighted criteria. These criteria were introduced for **different** notions of classification, and to **create** a bridge, mathematically validated, between Multiple Correspondence Factorial Analysis and Relational Analysis.

#### **6.5.1.2 The weighted Condorcet table**

The weighed criteria allow for the introduction of weights in the **object comparisons**; they are based on the traditional «*presence-rareness index*».

##### **The presence-rareness index:**

*It is because of the data nature we need that we used weighted criteria. The whole data are very poor in information (there are many 0 in the matrix), so we need a measure which reflects this phenomenon of «presence-rareness». By «presence-rareness» index we mean a similarity index  $S(x,y)$ , between two objects  $x$  and  $y$ :*

$$S(x,y) = \frac{N(x,y)}{D(x,y)} \text{ varying from 0 to 1.}$$

*Tk numerator equals 0 or 1 depending if  $x$  is similar or not with  $y$ , the denominator  $D(x,y)$  equals number of objects  $y$  with:*

$$N(x,y) = 1, \text{ so } D(x,y) = |\{y \mid N(x,y) = 1\}| \text{ (cardinal set)}$$

So  $s(x,y) = 1$ , if  $x$  is alone and  $s(x,y) \leq 1$  if  $N(x,y) = 1$  et  $D(x,y) > 0$ .

If  $N(x,y) = 0$  :  $\forall$  the denominator value,  $S(x,y) = 0$  That is the origin of its name «presence-rareness» index, because it measures the presence of a similarity according to its rarity. The principle of a such measure is to consider two objects all the more similar than they are rare in the studied population.

We consider that two **objects** are **very similar** when they share a characteristic which is rare in the population to **classify**.

The matrix  $\hat{C}$  of pairwise **comparisons** between **objects** (using «presence-rareness index») has as general term:

$$\hat{c}_{ir} = \sum_j \frac{k_{ij} k_{rj}}{k_{\bullet j}}$$

where  $k_{\bullet j} = \sum_i k_{ij}$  = number of **objects** which possess the **form j**

The **similarity**  $\hat{c}_{ir}$  between two **objects**  $i$  and  $i'$  will be as greatest as they share  $(k_{ij}, k_{rj})$  rare forms (divided by  $k_{\bullet j}$ ).

The weighted Condorcet criterion, based on this similarity, takes the following **form**:

$$\hat{C}(X) = \sum_{i \in I} \sum_{r \in I} (\hat{c}_{ir} - \bar{\hat{c}}_{ir}) x_{ir} + \sum_{i \in I} \sum_{r \in I} \bar{\hat{c}}_{ir}$$

$$\text{with } \hat{c}_{ir} = \frac{\hat{c}_{ii} + \hat{c}_{rr}}{2} - \hat{c}_{ir}.$$

### 6.5.1.3 The weighted Burt table

When **analyzing** the modalities of **classification** using the **presence** rareness index the data **matrix** we process is the  $\hat{B}$  matrix, defined as follows:

$$\hat{b}_{jr} = \sum_i \frac{k_{ij} k_{ir}}{k_{i\bullet}}$$

where  $k_{i\bullet} = \sum_j k_{ij}$  = number of modalities of the **object**  $i = M$  (number of variables).

So the weighted **Burt** criterion takes the **form**:

$$\hat{B}(Y) = \sum_{j \in J} \sum_{r \in J} (\hat{b}_{jr} - \bar{\hat{b}}_{jr}) y_{jr} + \sum_{j \in J} \sum_{r \in J} \bar{\hat{b}}_{jr} \text{ with } \hat{b}_{jr} = \frac{\hat{b}_{jj} + \hat{b}_{rr}}{2} - \hat{b}_{jr}.$$

It is precisely on the use of both weighted Burt criterion and weighted Condorcet criterion Relational Factorial Analysis is based on.

## 6.6 Relational Factorial Analysis

Unlike to other non hierarchical clustering method based upon inertial criterion, the relational analysis methodology **does** not oblige to **fix a priori** the number of classes of the solution.

According to the Huyghens **principle**, we know that the total inertia of a partition  $P$  noted  $I_T$ , is the sum of its within cluster inertia  $I_w(P)$  and its between clusters inertia  $I_B(P)$ :  
 $I_T = I_B(P) + I_w(P) \forall P$

This is a well known result to **maximize**  $I_B(P)$  or to **minimize**  $I_w(P)$ . Also a trivial solution exists if we do not place constraints upon the number of clusters, the solution **consists** of a partition in  $N$  clusters, where all the **objects** are isolated. In the case of qualitative data and for a «trivial» partition  $P_{iso}$ , solution of the inertia maximization problem, without constraints upon the number of clusters, the between cluster inertia keeps the value  $P/M - 1$ , with:

$$I_T = I_B(P_{iso}) = \frac{P}{M} - 1 \text{ and } I_w(P_{iso}) = 0$$

This result, well-known the **AFCM's** specialists (Analyse Factorielle des Correspondances Multiples), explains why it is necessary to fix the number of clusters when using an inertial index in non hierarchical cluster analysis. That is the reason why Marcotorchino [Marc91a, **Ayou90**] proposed to keep the inertial properties to define a criterion with «**natural intuitive properties**» but which **does** not require the fixation **a priori** the number of clusters.

The within and between cluster inertia **can** simply be written with relational notations, as follows.

$$I_B(P) = \sum_i \sum_r \frac{\hat{c}_{ir}}{M} \frac{x_{ir}}{x_i} - 1 \text{ and } I_w(P) = \frac{P}{M} - 1 - I_B(P)$$

where  $X$  is the binary relational matrix of the **equivalence** relation searched (unknown partition) and  $\hat{C}$  the weighted Condorcet matrix.

In multiple correspondence factorial analysis the processed matrix, (AF), has for general term:

$$AF_{ir} = \sum_i k_{ij} \frac{k_{rj}}{\sqrt{k_{i\bullet}} \sqrt{k_{r\bullet} k_{\bullet j}}} = \frac{1}{M} \sum_j \frac{k_{ij} k_{rj}}{k_{\bullet j}}$$

$$\text{So } AF_{ir} = \frac{\hat{c}_{ir}}{M}$$

The weighted Condorcet matrix is the basic matrix of the multiple correspondence factorial analysis. In practice a multiple correspondence factorial analysis **consists** in projecting points on the **first** factorial axis ( $r = 1, 2, 3, \dots$ ) and generally making the interpretation for small values of  $r$ . **One** qualitative indicator of an analysis is the percentage of inertia explained by the primary axis:

$$\frac{\lambda_1 + \lambda_2 + \lambda_3 + \dots + \lambda_r}{\sum \lambda_i}$$

In AFCM,  $I_r = \sum_{i=1}^q \lambda_i$  (where  $q$  is the number of eigen-values non equal to 0) and if we note  $I_F(r)$  the **inertia** explained by the  $r$  **first** axis we would like to compare  $I_F(r)$  and  $I_B(P)$ . If  $I_B(P)$  is greater than  $I_F(r)$  we are in a **configuration** where the partition  $P$  **provides** «more information» on the data structure than the AFCM **does**.

Maximizing  $I_B(P)$  is equivalent to **maximize**  $\sum \sum \frac{\hat{c}_{ir}}{M} \frac{x_{ir}}{x_{i\bullet}} - 1$ , and this inertia (if we do not add **constraints**) is maximal and equal to  $I_r$  **when all the** points are separated.

To avoid the simple response presented previously, it is necessary to find a partition without **fixing** the number of clusters, and which is the nearest from the inertia research. This explains the use of the weighted Condorcet criterion, because it generally gives an inertia greater than  $I_F(r)$  when  $r$  is **small**. **So** we obtain a clustering without **fixing** hypothesis «a priori», and we **can surround**, on the factorial graph, realistic clusters given by the classification, and compatible with the factorial data analysis, because the **cost**

$(\frac{\hat{c}_{ir}}{M} - \frac{\hat{c}_{ir}}{M}) x_{ir}$  of the weighted Condorcet Criteria works on the **same** data as the AFCM (table 6).

With **references** we **can** build a matrix crossing the patent numbers **fields** and the Derwent codes **fields**. This matrix is the basic table called  $K$  in our relational notations.

This table is used to **create** the  $\hat{B}$  and  $\hat{C}$  table. Our **purpose** is not to analyze the result of the **analysis**, because it is the job of experts in contact lens and chemical or enzymatical treatment. But we **can say**, when we look at the **graphics** that the partition we have on our projection, **does** have a logical explication when analyzing primary data. For example, codes P43 and **S05** are in the **same** cluster because only **one** patent has in its description the couple (P43, S05). And if we just analyze the AFC projection in the **first** two axes it is impossible to **reach** this conclusion.

## 6.7 Conclusions

---

This case study was developed to show how **difficult** the analysis of downloaded information **can** be. This is due to information properties and to the **final** user of the information which is **usually** an expert of the analyzed subject or a chief executive but not an expert in information science. This point **confirms** that technology assessment is a real interface science which needs experts in information science whose job is to look for sophisticated methods, test and use them. The results must then be presented to the chief executive in a literal form. The other important point is to show that nowadays new computerized tools exist for solving what was considered a problem not so long **ago**. We do hope to go on with **such** operations.



The authors would like to thank the referees for useful suggestions and help concerning **references**, and they thank **Orbit** Information Technologies for providing access to various **database** files.

Table 1: Constitutive fields of a **bibliographic reference** and their meanings

-1-	
AN -89-354640/48	Accession Number in <b>the database</b>
TI -Wall pane1 - has slits joining inter-pane <b>space</b> to interior, made in top of window <b>frame</b> , and slit joining it to ventilation <b>cavity</b>	<b>Title of the patent</b>
DC -Q44 Q48	Derwent Codes
PA -(EVEN/) EVENTOV V S	<b>Patent Owner</b>
IN -EVENTOV VS	<b>Inventor</b>
PN -SU1479589-A 89.05.15 (8948)	Patent Number
PR -86.0508 86SU-075778	<b>Priority Number</b>

Table 2: Number of codes by section

Section	Number of codes
PLASDOC	41
FOODDET	13
<b>ELECTRIC</b>	44
FARMAG	11
CHEMDOC	56
GENERAL	46
MECHANIC	<b>70</b>
<b>SX-ELECT</b>	49

Table 3: Frequency of **each** Derwent codes in **our** downloaded data

Derwent code	Frequency
D22	12
D16	9
P34	7
A96	6
<b>P81</b>	5
E36	4
P43	4
R16	3
E17	2
E19	2
E37	1
D25	1
<b>S05</b>	1
A97	1
P31	1

Table 5

Code	Frequency	Pairs	Frequency
A	20	AD	20
B	100	BC	30
C	200		
D	25		

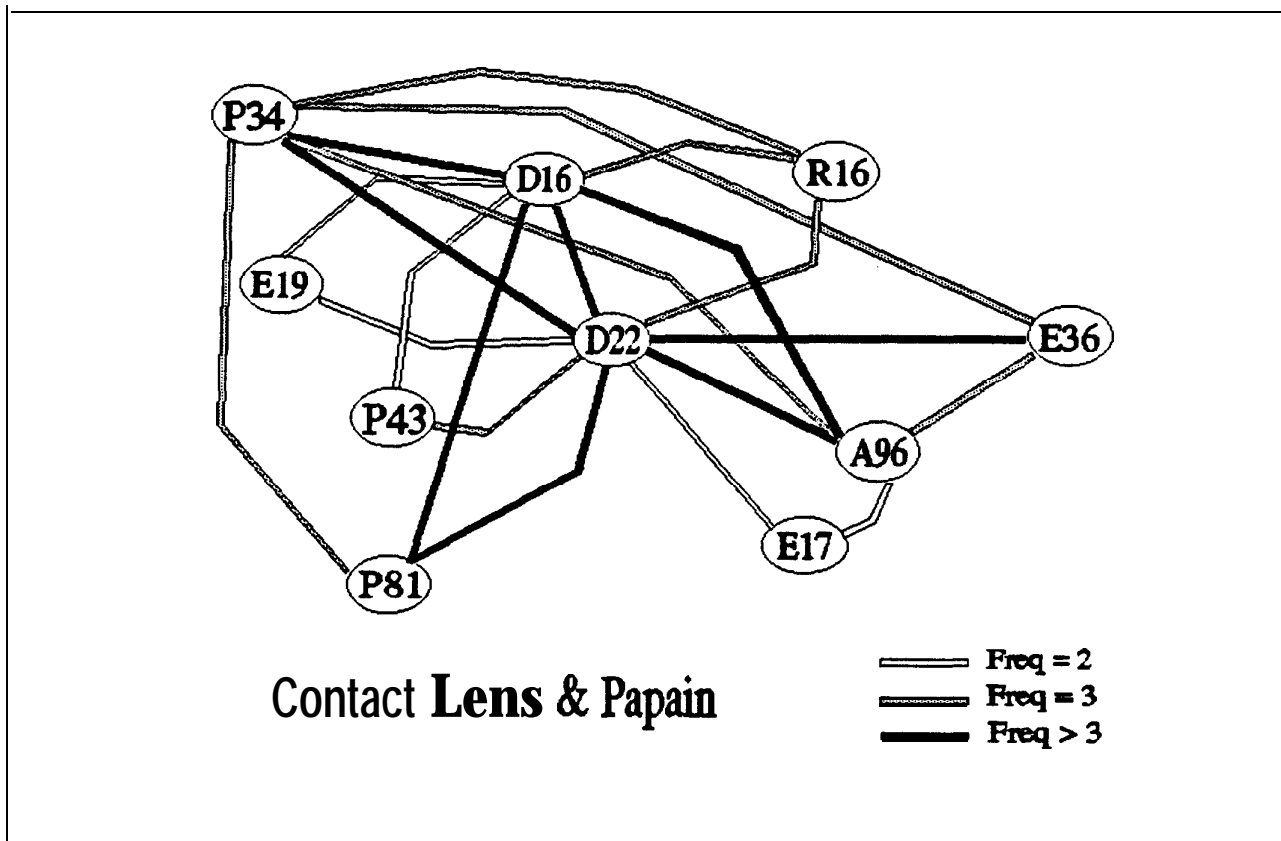


Fig. 1: Links between Derwent codes in our downloaded database

## Presentation of downloaded data

Table 4: Derwent codes

Letter	Signification
A	Codes PLASDOC
D	Codes FOODDET
E	Codes CHEMDOC
P	Codes GENERAL
R	Codes ELECTRIC
S	Codes SX-ELECT
DZ2	<b>Desinf, deter.</b> , dental, sterilizing, bandages, sutures, <b>plaster casts</b> , prostheses (lens).
<b>P34</b>	Health. amusement. <b>sterilizing</b> , syringes, electrotherapy
D16	<b>Food</b> , fermentation industry, brewing, <b>yeast</b> , pharmaceuticals <b>alcohol</b>
<b>P81</b>	Optics, <b>photography</b> , general. optics
A %	Veterinary, <b>medical</b> , dental.
E36	General inorganic. <b>none-metallic</b> elements
R16	<b>Measuring</b> , testing, <b>investigating chem./phys.props.</b>
E17	General organic, <b>other aliphatics</b>
<b>D25</b>	<b>Desinf.deter.</b> soap.including <b>metal salt and fatty acids</b> used in soaps
P43	Sepuating. mixing. <b>sorting</b> , cleaning
<b>S05</b>	<b>Electromedical</b>
<b>A97</b>	<b>Miscellaneous</b> goods
E19	General organic, other organic compounds <b>general</b>
P31	Health, amusement <b>diagnosis</b> , surgery
E37	General inorganic, mixtures of <b>many</b> components

Table 6: Initial **matrix**

	D	P	D	P	A	E	R	E	D	P	S	A	E	P	E	L
	<b>2</b>	<b>3</b>	<b>1</b>	<b>8</b>	<b>9</b>	<b>3</b>	<b>1</b>	<b>1</b>	<b>2</b>	<b>4</b>	<b>0</b>	<b>9</b>	<b>1</b>	<b>3</b>	<b>3</b>	<b>A</b>
	<b>2</b>	<b>4</b>	<b>6</b>	<b>1</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>7</b>	<b>5</b>	<b>3</b>	<b>5</b>	<b>7</b>	<b>9</b>	<b>1</b>	<b>7</b>	<b>L</b>
US474951 1	<b>1</b>	<b>1</b>	1	1	0	0	0	0	0	0	0	0	0	0	0	4
us7757014	<b>1</b>	<b>1</b>	1	0	1	1	1	0	0	0	0	0	0	0	0	6
US4609493	1	0	1	1	<b>1</b>	0	0	1	1	0	0	0	0	0	0	6
US46 14549	1	1	1	1	0	0	0	0	0	1	0	0	0	0	0	5
US4872965	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	2
US4839082	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
US4832754	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0	4
US4808239	1	0	0	0	1	0	0	1	0	1	0	0	0	0	0	4
US4767559	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	3
US4784790	1	1	1	0	0	0	0	0	0	0	0	1	1	1	0	6
US482900 1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	6
US4855234	1	<b>1</b>	1	0	1	1	1	0	0	0	0	0	0	0	0	6
US4710313	1	0	1	1	0	0	0	0	0	0	0	0	1	0	1	5
Total	<b>1</b>	7	9	5	6	4	3	2	1	4	1	1	2	1	<b>1</b>	59
	<b>2</b>															

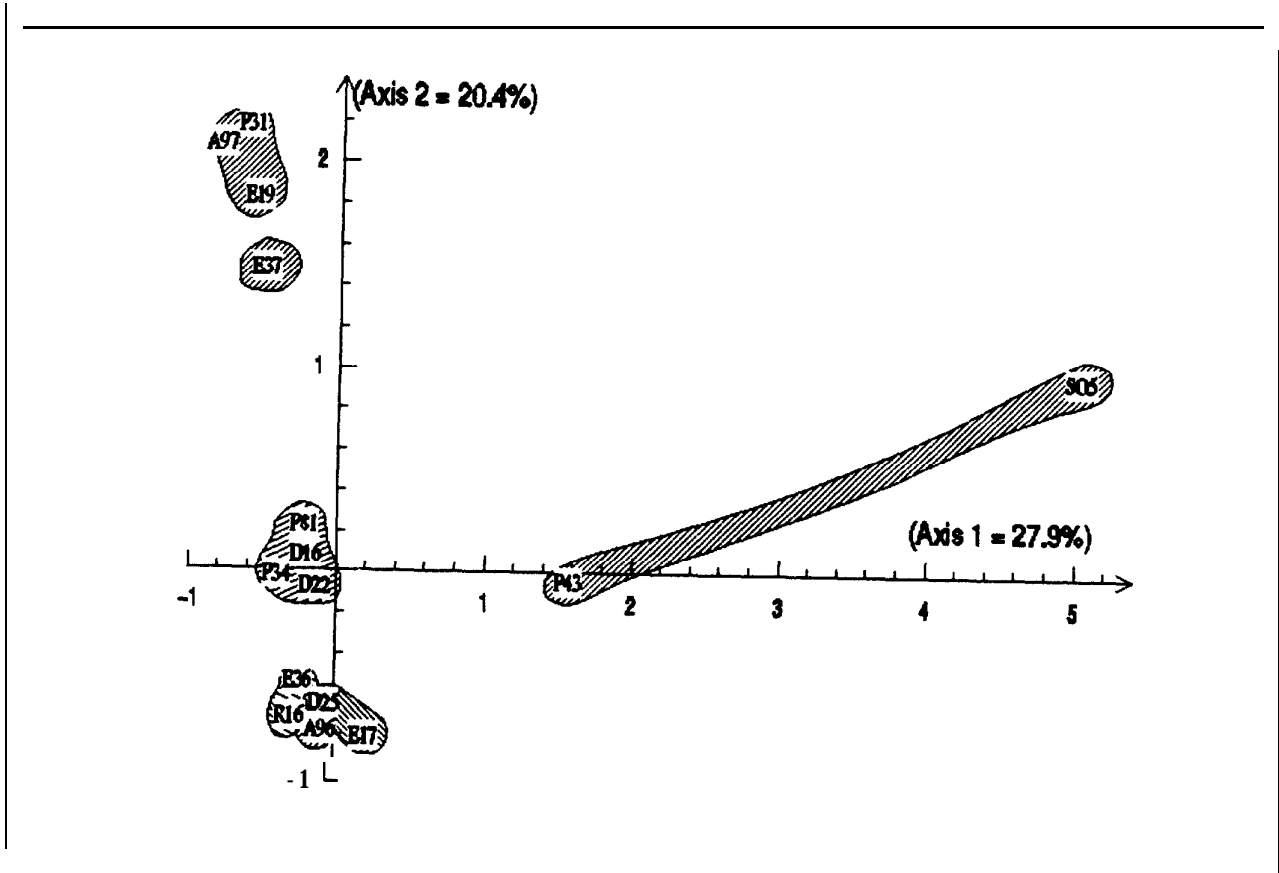


Fig. 2: Graph of the factorial relational analysis applied to our downloaded data

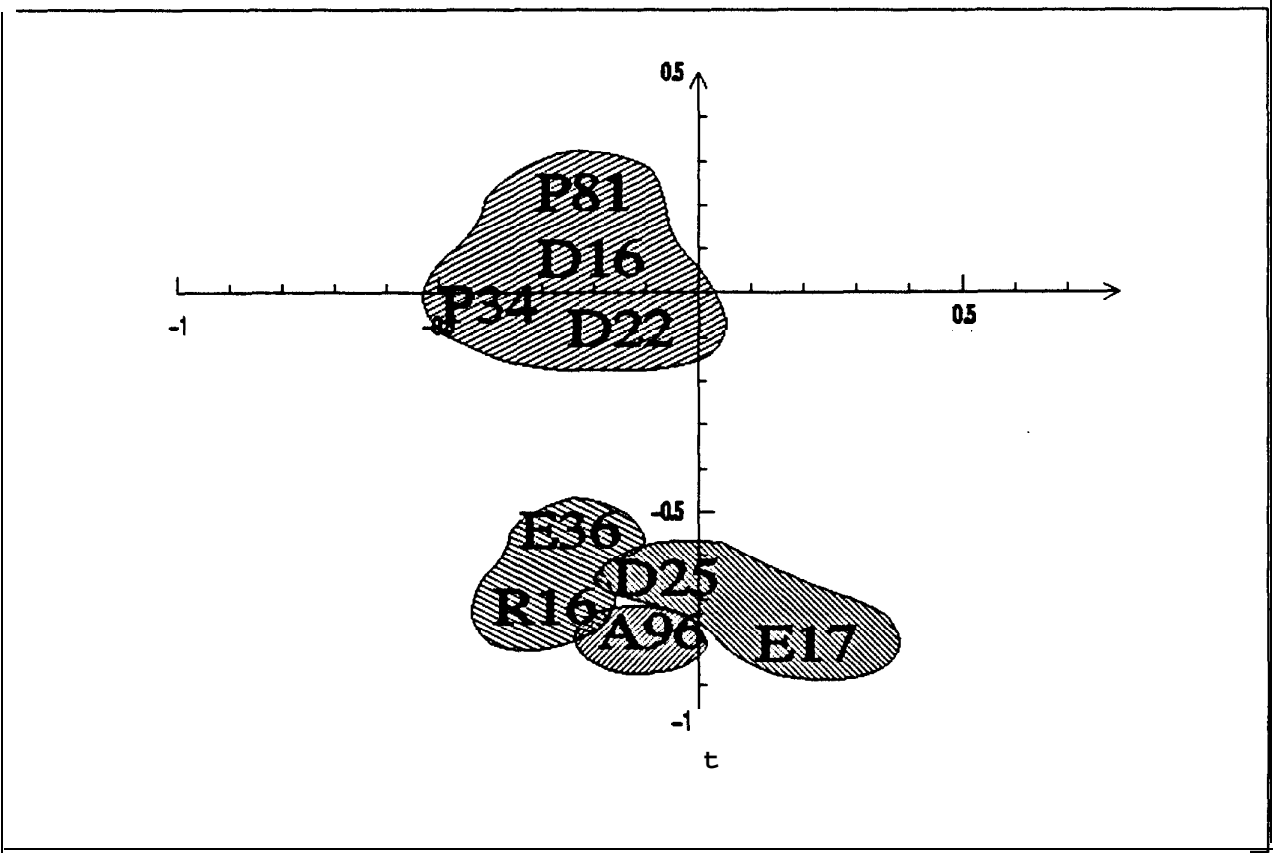


Fig. 3: Zoom of the graph central part



# *Première annexe technique*

**Analyse Relationnelle : des outils  
pour la documentation automatique**

*Chantai Bédécarrax et Charles Huot*

dans Hélène Desvals et H.Dou: "La veille  
Technologique", Dunod, pp 348- 368,  
janvier 1992.



# Analyse Relationnelle: des outils pour la Documentation Automatique

Chantal BEDECARW , Charles HUOT  
Centre Européen de Mathématiques Appliqués **IBM**<sup>28</sup>

## 7.1 Introduction

L'analyse relationnelle regroupe un ensemble de techniques d'analyse des données qui permettent de modéliser et de résoudre des problèmes relevant de la formulation générale suivante :

**Recherche d'une relation structurée  $S$  qui s'ajuste "au mieux" à une relation (ou un ensemble de relations) quelconque donnée  $R$ .**

Cette formulation couvre une vaste gamme de **problèmes**, posés pour certains depuis fort **longtemps**<sup>29</sup>, et touchant à des domaines très variés.

C'est vers la fin des années 70 que F. Marcotorchino et P. **Michaud** établissent une jonction entre ces **différents** sujets de recherche en proposant une méthodologie générale de modélisation et de résolution : **l'analyse relationnelle**. Elle unifie tous ces problèmes en les situant sur une base commune qui combine une approche relationnelle pour la prise en compte des données et une modélisation sous forme de programme **linéaire** pour la résolution.

Les domaines d'application de l'analyse relationnelle sont trop nombreux pour que nous les présentions, ici, de façon exhaustive. En outre, les **problèmes** les plus courants dans le domaine de la **Veille Technologique** relèvent de la classification, c'est donc sur cette problématique que nous mettrons **l'accent**<sup>30</sup>.

**La prise en compte des données est basée sur des principes logiques de comparaisons par paires** : on s'intéresse aux relations qu'entretiennent, deux à deux, les objets **à classer**.

---

<sup>28</sup> CEMAP IBM, 68-76 quai de la Rapée, 75592 Park CEDEX 12, Tel: 40.01.57.11/ 40.01.53.37, Fax: 49.28.08.60

<sup>29</sup> Le premier apparaît dans les travaux du marquis de Condorcet sur la théorie des votes, en 1785. Il s'agit de trouver le meilleur classement collectif des candidats à une élection, à partir des classements individuels fournis par le corps électoral (ensemble des votants). Les travaux de Condorcet ont été à l'origine du développement de la méthodologie relationnelle, comme le rappelle P. Michaud dans son "Hommage à Condorcet" [Mich82].

<sup>30</sup> On trouvera des développements très complets sur le traitement des données ordinales dans les travaux de F. Marcotorchino et P. Michaud [Marc78, Marc81] ou dans la thèse de S. Ghoshghaie [Ghas90].

Cette présentation offre une grande souplesse. D'une **manière** générale, les données sont traitées sous leur **forme brute**, sans recours à des techniques de codage pour les adapter au modèle, ce qui évite toute perte d'information.

Tous les **problèmes** d'analyse relationnelle ont donc, pour point de départ, une **matrice de base** qui contient l'**information** sur les relations à structurer. Il s'agit **ensuite d'écrire** le **modèle** adapté au **problème** que l'on se pose. La modélisation, comme nous l'avons signalé, s'effectue à l'aide d'un programme linéaire. Celui-ci se compose de deux parties :

- **une fonction économique linéaire** à maximiser; son **rôle** est tenu par le critère d'ajustement de la solution aux **données**<sup>31</sup>,
- **un système de contraintes linéaires** qui permettent de caractériser la solution en fonction du problème **posé**<sup>32</sup>.

La variété des **critères** disponibles, combinée à l'éventail des relations binaires **définies** linéairement, permettent à l'analyse relationnelle de couvrir une vaste gamme de problèmes relevant de la formulation énoncée plus haut.

Le but de l'analyse relationnelle, comme celui de la statistique en général, est de fournir une solution qui **résume** un **ensemble de données**, trop vaste pour être appréhendé dans sa globalité. Il est donc fondamental d'être à même de **vérifier**, *a posteriori*, la validité de cette solution relativement au problème posé et aux données en jeu. La méthodologie relationnelle offre des outils qui permettent une analyse détaillée de la solution, aidant aussi bien à sa validation qu'à son interprétation [Mich85].

## 7.2 Méthodologie

Les données sont prises en compte dans une matrice de comparaisons par paires *R*, contenant l'information sur les relations que l'on cherche à modéliser. Cette matrice peut se présenter sous **différentes** formes. Selon les problèmes, on aura affaire à :

- **une matrice carrée**, si l'on s'intéresse à un ensemble unique et aux relations qu'entretiennent les objets à l'**intérieur** de cet ensemble;
- une matrice rectangulaire, si l'on travaille sur deux ensembles distincts et que les relations que l'on cherche à analyser portent sur le croisement des objets issus des deux ensembles.

---

<sup>31</sup> La **linéarisation des critères** d'ajustement de la **solution aux données** constitue un atout majeur de l'analyse relationnelle. En réunissant les travaux de F. Marcotorchino [Marc84, Marc85], S. Chah [Chah86] et H. Messatfa [Mess90] sur le sujet, on se trouve en possession d'un grand nombre de **critères**, pour la plupart fort **classiques**, écrits sous forme **linéaire**.

<sup>32</sup> La formulation **linéaire** des **propriétés** de relations **binaires** a permis d'établir des expressions **caractérisant**, sous forme d'**égalités** ou d'**inégalités**, des propriétés **essentielles** telles que la **transitivité**, la **triade impossible** (ou **transitivité généralisée**), la **symétrie**, l'**antisymétrie**, la **totalité**, l'**intermédiarité**, la **cohérence**, l'**affectation**, etc. Par la **combinaison** de **ces expressions** de base on **définit** des relations binaires **classiques**.

Dans chacun des deux cas,  $R$  peut renfermer des données de type :

- binaires, si l'on ne traite qu'une seule relation;
- non binaires, si l'on prend en compte plusieurs relations.

Quelle que soit donc la forme de  $R$ , pourvu qu'elle renferme des informations de nature relationnelle que l'on veut structurer, elle peut être soumise à la méthodologie de l'analyse relationnelle.

En fonction du problème à résoudre, on doit choisir le type de relation  $S$  qui va être ajustée aux données. D'une manière générale, cette solution  $S$  est une relation binaire qui vérifie un certain nombre de propriétés caractérisées par un système de contraintes linéaires (égalités ou inégalités).

A tout problème relevant de l'analyse relationnelle des données est associé, comme nous l'avons dit, un programme linéaire. Si l'on note formellement  $F(R,S)$  le critère, linéaire en  $S$ , de mesure d'adéquation entre les données  $R$  et la solution  $S$ , ce programme prend la forme générale :

**Max  $F(R,S)$  sous les contraintes (linéaires) générées par les propriétés de  $S$**

Le premier critère utilisé pour la résolution des problèmes d'analyse relationnelle a été le critère de Condorcet ou **critère de la majorité**. Basé sur la règle de la majorité, il s'avère parfaitement adapté à tout problème d'agrégation, c'est-à-dire aux cas où l'on a affaire à une matrice de base qui se présente comme somme d'un certain nombre de matrices binaires. C'est aussi un excellent critère d'ajustement pour le traitement de données binaires, en tant que cas particulier des problèmes précédents.

Si le critère de Condorcet est historiquement le premier exploité en analyse relationnelle, il est important de souligner qu'il en existe beaucoup d'autres. Nous avons déjà signalé que de nombreux critères pouvaient s'exprimer à l'aide des notations relationnelles, dont un grand nombre sous forme linéaire. Il en est ainsi, par exemple, pour la plupart des critères usuels de la statistique des contingences. Ces différents critères offrent des alternatives à la règle de la majorité, mal adaptée dans certains contextes, voire quelquefois, inapplicable ou dénuée de sens.

Nous présenterons en particulier les critères de type **burtien** et les **critères pondérés** qui s'adaptent particulièrement bien aux données que l'on rencontre dans le domaine de la Veille Technologique.

Dans la problématique générale de la classification, nous distinguerons deux types de problèmes :

- ceux qui relèvent de la **classification simple**,
- ceux qui relèvent de la classification croisée ou **sériation**.

Dans la **première** catégorie, on traite des **matrices carrées** prenant en compte les relations à l'intérieur d'un ensemble unique; c'est le cas par exemple en **agrégation des similarités** ou en ajustement de relations.

Dans la seconde entrent naturellement les **tableaux rectangulaires**<sup>33</sup> portant sur le croisement de deux ensembles, nous sommes alors dans le domaine de la **sériation par blocs**.

Dans les deux cas, on peut résumer l'essence même des problèmes par la fameuse maxime: "**Qui se ressemble s'assemble**".

Lorsque l'on cherche à effectuer une classification, préciser clairement ce que l'on entend par se **ressembler** constitue une étape fondamentale; on doit déterminer un critère qui permettra de mesurer la ressemblance entre les objets. La notion **d'assembler** doit également être **définie** car elle peut **faire** l'objet de **différentes** stratégies; il faut opter pour un mode de réunion des objets en fonction du problème auquel on se trouve confronté. C'est l'un des atouts de l'analyse relationnelle que d'offrir un choix dans les réponses à apporter à ces deux types de questions.

Dans la **dernière** partie de ce chapitre, nous présenterons la méthode de **quadri-décomposition** qui est une méthodologie dérivée de l'analyse relationnelle. Elle englobe les deux types de classification (simple et croisée) dans une approche méthodologique **unifiée et permet** en outre de traiter des données plus complexes, en particulier le cumul de relations sur les croisements de plusieurs ensembles **différents** et leur traitement simultané. De fait, elle offre une ouverture de l'analyse relationnelle à des **problèmes** de classification plus généraux.

### 7.21 Classification simple

Le problème porte sur la modélisation des relations de similarités qu'entretiennent  $n$  objets à l'intérieur d'un ensemble  $I$ . Les informations sur cette (ces) relation(s) sont donc prises en compte dans la matrice carrée de comparaisons par paires  $R$ , croisant l'ensemble  $I$  avec lui-même. Son terme général  $r_{ii'}$  peut se présenter, selon la nature du problème traité, sous différentes formes. Parmi les plus fréquentes, citons :

- **présence/absence** de relation entre les objets  $i$  et  $i' \rightarrow r_{ii'} \in \{0, 1\}$ ,
- intensité de la relation entre les objets  $i$  et  $i' \rightarrow r_{ii'} \in \mathbb{R}^+$ ,
- nombre de fois où les deux objets  $i$  et  $i'$  sont en relation  $\rightarrow r_{ii'} \in \mathbb{N}^+$ .

*On* cherche une partition de  $I$ , en classes d'équivalence, donc une relation, que nous noterons  $X$ , dotée d'une certaine structure, qui s'ajuste au mieux aux informations contenues dans  $R$ .

---

<sup>33</sup> Ils peuvent bien sûr avoir un nombre identique de lignes et de colonnes, mais être structurellement rectangulaires pu k frit qu'ils croisent deux ensembles distincts.

Une partition  $X$  d'un ensemble  $I$  correspond, en termes relationnels, à une relation d'équivalence sur  $I \times I$ , qui se présente sous la forme d'une matrice carrée binaire. Les caractéristiques d'une telle relation sont :

- **la binarité** :  $x_{ii'} \in \{0, 1\}$
- **la réflexivité** : tout objet  $i$  est en relation avec lui-même
- **la symétrie** : si  $i$  est en relation avec  $i'$  alors  $i'$  est en relation avec  $i$ , autrement dit  $x_{ii'} = 1 \Leftrightarrow x_{i'i} = 1 \forall i, i' \in I$ .
- **la transitivité** : si  $i$  est en relation avec  $i'$  et  $i'$  est en relation avec  $i''$  alors  $i$  est en relation avec  $i''$ , autrement dit  $x_{ii'} = 1$  et  $x_{i'i''} = 1 \Rightarrow x_{ii''} = 1 \forall i, i', i'' \in I$ .

Ces propriétés s'expriment **linéairement** de la façon suivante :

- **réflexivité** :  $x_{ii} = 1 \forall i \in I$
- **symétrie** :  $x_{ii'} - x_{i'i} = 0 \forall i, i' \in I$
- **transitivité** :  $x_{ii'} + x_{i'i''} - x_{ii''} \leq 1 \forall i, i', i'' \in I$

#### 7.2.1.1 Critères de type Condorcéen

Ils **opèrent** sur des données résultant de la somme d'un certain nombre de relations binaire: de départ. Ainsi, le terme général de la matrice de données, notée ici  $C$ , s'écrit :

$$c_{ii'} = \sum c'_{ii'} \text{ avec la caractéristique essentielle que } c_{ii} = m = \text{constante } \forall i \in I.$$

Cette 'configuration caractérise les matrices de comparaisons par paires dérivées de la réunion d'opinions individuelles émises par un nombre **fixe** de "juges" ou variables. Elle apparaît par exemple dans le traitement des éléments lignes (ou individus) des tableaux **disjonctifs**, très courants en Analyse des Données. Ces tableaux binaires sont issus de la mise sous forme disjonctive complète de  $m$  variables à  $p$  modalités décrivant une population  $I$ .

Si l'on note  $K$  un tel tableau, on aura en l'occurrence :

$$c_{ii'} = \sum_{j=1}^m c'_{ii'} = \sum_{j=1}^p k_{ij} k_{i'j}, \text{ avec } k_{ij} = \begin{cases} 1 & \text{si l'individu } i \text{ possède la modalité } j \\ 0 & \text{sinon} \end{cases}$$

et  $c'_{ii'} = \begin{cases} 1 & \text{si les individus } i \text{ et } i' \text{ ont la même modalité de la variable } l \\ 0 & \text{sinon} \end{cases}$

On a ici :  $c_{ii} = \sum_{j=1}^p k_{ij} = m$ , car chaque individu ne possède qu'une modalité de chaque variable.

Pour exploiter ce type de données dans une optique classificatoire, les critères condorcéens sont de la forme :  $C(X) = \sum_{i \in I} \sum_{i' \in I} [c_{ii'} x_{ii'} + \bar{c}_{ii'} (1 - x_{ii'})]$ , ou la matrice  $\bar{C}$

représente le complémentaire de C au maximum de liens possibles entre deux individus  $i$  et  $i'$ , en l'occurrence  $\bar{c}_{ii'} = m - c_{ii'}$ <sup>34</sup>.

$C(X)$  n'est autre que le critère de Condorcet appliqué à la matrice de comparaisons par paires  $C$ . Quant à  $X$ , elle représente la relation d'équivalence sur  $I \times I$  ou partition des  $n$  éléments de  $I$ .

Examinons de plus près la signification du critère  $C(x)$ .

Maximiser  $C(X)$  revient à maximiser simultanément, pour une paire donnée  $(i, i') \in I$ , le nombre d'accords internes (représenté par  $c_{ii'} x_{ii'}$ ) et le nombre d'accords externes (représenté par  $\bar{c}_{ii'} (1 - x_{ii'})$ ); ce qui conduit, au niveau de la matrice toute entière, à réunir les individus qui se ressemblent (maximisation des liens intra classe) tout en s'opposant aux autres (maximisation des non liens inter classes).

Le critère  $C(X)$  se développe en :  $C(X) = \sum_{i \in I} \sum_{i' \in I} (2 c_{ii'} - m) x_{ii'} + \sum_{i \in I} \sum_{i' \in I} (1 - c_{ii'})$ .

Il apparaît donc que :  $\max_x C(X) \Leftrightarrow \max_{i, i'} \sum_{i \in I} \sum_{i' \in I} (2 c_{ii'} - m) x_{ii'}$ .

On retrouve bien la règle de la majorité qui consiste à regrouper deux individus  $i$  et  $i'$  (soit

$x_{ii'} = 1$ ) si  $c_{ii'} \geq \frac{m}{2}$ , c'est-à-dire dès qu'ils ont été réunis par au moins la moitié des variables.

Remarquons que ce critère s'interprète également en termes d'indice de similarité entre profils des individus dans le tableau disjonctif  $K$ , dont la matrice de données  $C$  serait issue.

En effet, si l'on note  $S_D(i, i')$  l'indice de Dice entre les deux individus  $i$  et  $i'$ , par définition :

$$S_D(i, i') = \frac{2 (\text{nombre de 1 communs aux lignes de } i \text{ et de } i')}{\text{nombre de 1 de la ligne de } i + \text{nombre de 1 de la ligne de } i'} = \frac{c_{ii'}}{m}$$

Selon l'approche indicielle de Solomon-Fortier, on considère que deux individus  $i$  et  $i'$  se ressemblent si  $S_D(i, i') \geq 1/2$ ; ce qui nous ramène naturellement à la règle de la majorité de Condorcet :  $c_{ii'} \geq \frac{m}{2}$ .

Le critère  $C(X)$  admet également d'autres interprétations. Ainsi, il est équivalent au critère de la différence symétrique qui sert, dans une approche métrique, à mesurer la distance entre deux partitions. C'est encore ce même critère que l'on retrouve derrière l'indice de Rand, utilisé en statistique des contingences pour comparer des structures de partitions.

<sup>34</sup> Nous considérerons, ici, que  $\bar{c}_{ii'} = m - c_{ii'} \forall i, i' \in I$ , c'est-à-dire que l'on dispose de toutes les données. C'est dans un souci de simplification que nous avons choisi cette option, mais la méthodologie permet de traiter des problèmes plus généraux avec données manquantes.

### 7.2.1.2 Critères de type Burtien

Formellement, la structure des critères burtiens est la suivante :

$$B(X) = \sum_{j \in J} \sum_{j' \in J} [ b_{jj'} x_{jj'} + \bar{b}_{jj'} (1 - x_{jj'}) ] .$$

C'est dans la manière de calculer les  $\bar{b}_{jj'}$ , qu'ils se distinguent des critères de type condorcéen. Le mode de construction de ces quantités est une nouvelle fois dicté par la nature des données. Ici, la matrice de comparaison par paire  $B$  ne dégage pas de notion de majorité constante. Ainsi, si l'on reprend l'exemple des tableaux **disjonctifs**, la matrice de Burt  $B$  correspondante est issue de la comparaison des colonnes, c'est-à-dire des modalités. Autrement dit :

$$b_{jj'} = \sum_{i=1}^n k_{ij} k_{ij'} , \text{ avec cette fois : } b_{jj} = \sum_{i=1}^n k_{ij} = \text{nombre d'individus possédant la modalité } j .$$

La quantité  $b_{jj}$  variant pour chaque  $j$ , il faut adopter une nouvelle stratégie pour construire les

$\bar{b}_{jj'}$ : c'est la stratégie burtienne. Elle consiste à poser :  $\bar{b}_{jj'} = \frac{b_{jj} + b_{j'j'}}{2} - b_{jj'}$ . L'approche indicielle de **Solomon-Fortier confirme** ce choix.

En effet, l'indice de **Dice** entre les **profils** des modalités  $j$  et  $j'$  dans le tableau disjonctif est **défini** par :

$$S_D(j, j') = \frac{2 (\text{nombre de 1 communs aux colonnes de } j \text{ et de } j')}{\text{nombre de 1 de la colonne de } j + \text{nombre de 1 de la colonne de } j'} = \frac{2 b_{jj'}}{b_{jj} + b_{j'j'}}$$

On considère que les deux modalités  $j$  et  $j'$  se ressemblent dès que  $S_D(j, j') \geq \frac{1}{2}$ .

A cette borne indicielle correspond la borne relationnelle  $b_{jj'} \geq \frac{b_{jj} + b_{j'j'}}{4}$ , qui amène naturellement à la définition de  $\bar{b}_{jj'}$ , présentée plus haut.

Ainsi, les critères de type burtien s'écrivent :

$$B(X) = \sum_{j \in J} \sum_{j' \in J} (2 b_{jj'} - \frac{b_{jj} + b_{j'j'}}{2}) x_{jj'} + \sum_{j \in J} \sum_{j' \in J} \bar{b}_{jj'}$$

Malgré la similitude de leurs expressions<sup>35</sup>, les critères de Burt et de Condorcet, ne remplissent pas exactement la même fonction. Le critère de Condorcet s'adapte parfaitement aux cas où l'on dispose d'une majorité constante sur l'ensemble des paires, alors que celui de Burt, plus général, s'intègre aux **problèmes** où l'on a **affaire** à une "majorité variable" (ou un équilibre) d'une comparaison à l'autre.

<sup>35</sup> On pourra remarquer que les **critères condorcéens** sont des cas particuliers des **critères burtiens** en ce sens que si l'on pose  $\bar{c}_{ii'} = \frac{c_{ii} + c_{i'i'}}{2} - c_{ii'}$ , du fait que  $c_{ii} = c_{i'i'} = m$ , on retrouve bien  $\bar{c}_{ii'} = m - c_{ii'}$ .

### 7.2.1.3 Les critères pondérés et l'Analyse Factorielle-Relationnelle

Les critères pondérés ont été introduits d'une part pour répondre à des stratégies de classification sensiblement **différentes** de celles dictées par les deux approches précédentes et d'autre part pour montrer qu'il existe un pont, mathématiquement validé, entre les approches relationnelle et factorielle [Marc89, Marc91a, Ayou90].

#### 7.2.1.2.1 Présentation de ces critères

Dans les deux cas précédents, pour que deux objets soient considérés comme ressemblants il suffisait qu'ils possèdent une majorité de descripteurs communs, que cette majorité soit constante (Condorcet) ou variable par paire (Burt).

Les critères pondérés ont pour objectif **d'affecter des poids aux comparaisons entre objets; ils** sont basés sur ce que l'on appelle traditionnellement des **indices de présence-rareté**. Ainsi, on considère que deux objets sont d'autant plus semblables qu'ils partagent des caractéristiques rares dans la population à classer.

Pour reprendre l'exemple général des tableaux disjonctifs, la matrice  $\hat{C}$  de comparaison par paires entre les individus, en terme d'indice de **présence-rareté**, a pour terme général :

$$\hat{c}_{ii'} = \sum \frac{k_{ij} k_{i'j}}{k_{.j}} \text{ où } k_{.j} = \sum k_{ij} = \text{nombre d'individus possédant la modalité } j.$$

Ainsi, la **similarité**  $\hat{c}_{ii'}$  entre 'deux individus  $i$  et  $i'$  sera d'autant plus élevée qu'ils partageront

$(k_{ij} k_{i'j} = 1)$  des modalités rares (division par  $k_{.j}$ ).

Le **critère** de Condorcet pondéré, basé sur cette similarité, prend la forme suivante:

$$C(X) = \sum_{i \in I} \sum_{i' \in I} (\hat{c}_{ii'} - \bar{\hat{c}}_{ii'}) x_{ii'} + \sum_{i \in I} \sum_{i' \in I} \bar{\hat{c}}_{ii'} \text{ avec } \bar{\hat{c}}_{ii'} = \frac{\hat{c}_{ii} + \hat{c}_{i'i'}}{2} - \hat{c}_{ii'}$$

De façon analogue, si l'on s'intéresse à la classification des modalités en **termes** d'indice de présence-rareté, on aura pour matrice de données la matrice  $\hat{B}$  définie par

$$\hat{b}_{jj'} = \sum \frac{k_{ij} k_{i'j'}}{k_{i.}} \text{ où } k_{i.} = \sum_j k_{ij} = \text{nombre de modalités de l'individu } i = m \text{ (nombre de variables).}$$

Le critère de Burt pondéré s'écrit alors :

$$\hat{B}(Y) = \sum_{j \in J} \sum_{j' \in J} (\hat{b}_{jj'} - \hat{b}_{jj}) y_{jj'} + \sum_{j \in J} \sum_{j' \in J} \hat{b}_{jj'} \text{ avec } \hat{b}_{jj'} = \frac{\hat{b}_{jj} + \hat{b}_{j'j'}}{2} - \hat{b}_{jj'}$$

C'est en fait la simultanéité d'utilisation des **critères** de Condorcet pondéré et de Burt pondéré qui nous permet d'introduire la notion d'analyse factorielle-relationnelle.

#### 7.2.1.3.2 L'Analyse Factorielle-Relationnelle

Les méthodes relevant de l'**Analyse** Relationnelle en Classification Automatique, ont permis depuis quelques années de pallier l'obligatoire **fixation** du nombre de classes de la

partition souhaitée de la population analysée, communément pratiquée dans bon nombre de méthodes usuelles de classification non hiérarchique.

Cette fixation a *priori* du nombre de classes étant due, dans les méthodes non hiérarchiques, à l'usage systématique de critères de classification s'appuyant sur l'**approche inertielle** du partitionnement. En effet, fort du principe de Huyghens, nous savons que l'inertie totale d'une partition P, notée  $I_T$  est la somme de son inertie intra-classes  $I_w(P)$  et de son inertie inter-classes  $I_b(P)$ , soit :  $I_T = I_b(P) + I_w(P) \quad \forall P$ .

Il est dès lors connu que, d'une part, il est équivalent de maximiser  $I_b(P)$  ou de minimiser  $I_w(P)$  et que, d'autre part, la solution triviale obtenue, si l'on ne met pas de contrainte sur le nombre de classes, consiste à "isoler" tous les objets de la population en une partition dite "triviale" à  $n$  classes (ou  $n$  est le nombre total d'individus de la population). Dans le cas de variables qualitatives, qui nous intéresse, et pour la partition "triviale"  $P_{iso}$ , solution du problème de maximisation de l'inertie sans contrainte de nombre de classes, l'inertie inter-classes prend la valeur  $p/m - 1$ , avec :  $I_T = I_b(P_{iso}) = \frac{p}{m} - 1$  et  $I_w(P_{iso}) = 0$

Ce résultat, fort connu, en particulier par les spécialistes de l'**Analyse Factorielle des Correspondances Multiples**, explique pourquoi il est indispensable de se **fixer** a priori un nombre de classes quand on utilise un **critère** de type **inertiel** en classification automatique non hiérarchique. C'est à cause de cette constatation restrictive que nous proposons dans cette partie de conserver l'approche "inertielle" (forte des considérations statistiques, mécaniques, mathématiques qui lui sont associées) pour définir un critère qui s'appuie sur l'inertie et qui possède des propriétés "intuitives" naturelles tout en évitant le passage obligé par la fixation du nombre de classes.

Les inerties inter et intra classes associées à une partition s'écrivent simplement en notations relationnelles sous la forme suivante:

$$I_b(P) = \sum_i \sum_{i'} \frac{\hat{c}_{i i'}}{m} \frac{x_{i i'}}{x_{i \cdot}} - 1 \quad \text{et} \quad I_w(P) = \frac{p}{m} - 1 - I_b(P)$$

où X représente la matrice relationnelle binaire de la relation d'équivalence cherchée (partition inconnue) et C la matrice de Condorcet pondéré présentée plus haut.

En Analyse Factorielle des Correspondances Multiples la matrice de l'analyse, notée AF, a pour terme général :

$$AF_{ii'} = \sum_j \frac{k_{ij} k_{i'j}}{\sqrt{k_{i \cdot}} \sqrt{k_{i' \cdot}} k_{\cdot j}} = \frac{1}{m} \sum_j \frac{k_{ij} k_{i'j}}{k_{\cdot j}}, \quad \text{soit} \quad AF_{ii'} = \frac{\hat{c}_{ii'}}{m}$$

donc la matrice de Condorcet Pondéré n'est autre que la matrice de base de l'**Analyse factorielle des correspondances multiples**. En pratique une analyse factorielle des correspondances multiples revient à projeter des points sur les  $r$  premiers axes factoriels ( $r = 1, 2, 3, \dots$ ) et en général on se contente d'une interprétation sur  $r$  faible. L'un des indicateurs de la qualité d'une analyse factorielle est le pourcentage d'inertie apporté par les premiers axes, soit le rapport :

$$\frac{\lambda_1 + \lambda_2 + \dots + \lambda_r}{\sum \lambda_i}$$

On sait en AFCM que  $I_T = \sum 1$ , (où  $q$  est le nombre de valeurs propres non nulles) et si l'on note  $I_R(r)$  l'inertie **expliquée** par les  $r$  premiers axes, on aimerait pouvoir comparer  $I_R(r)$  à  $I_B(P)$ . Si  $I_B(P)$  est supérieure à  $I_R(r)$  nous sommes dans une configuration où la partition  $P$  apporte "plus d'information" sur la structuration de la population que l'AFCM.

Nous avons vu que maximiser  $I_B(P)$  revient à maximiser  $\sum_i \sum_{i'} \frac{\hat{c}_{ii'}}{m} \frac{x_{ii'}}{x_{i' \cdot}} - 1$  or cette inertie (si

l'on n'ajoute pas de contrainte) est maximale et égale à  $I_T$  dès que l'on "isole" tous les points.

Pour ne pas tomber dans le processus trivial évoqué plus haut, il convient de trouver une partition sans avoir à fixer le nombre de classes et qui soit voisine de la structure de l'inertie cherchée. C'est ce que propose le critère de Condorcet Pondéré car il fournit une partition dont l'inertie associée est en général supérieure à  $I_R(r)$  pour  $r$  faible. Ceci permet donc d'obtenir une partition de la population sans hypothèses **fixées** a priori permettant **d'entourer sur le diagramme factoriel les classes réalistes** issues de la classification, tout en étant compatibles avec les données traitées par l'AFCM puisque le coût

$(\frac{\hat{c}_{ii'}}{m} - \frac{\hat{c}_{ii'}}{m}) x_{ii'}$ , du critère de Condorcet **Pondéré**<sup>36</sup> manipule les **mêmes** données que l'AFCM.

### 7.2.2 Sériation par blocs

On travaille sur deux ensembles simultanément et non plus sur un seul, comme c'était le cas pour les **problèmes** de classification développés dans les parties précédentes. De fait, la relation cherchée n'a plus les propriétés d'une relation d'équivalence. Elle se caractérise par une structure en blocs diagonaux disjoints qui mettent en correspondance les classes des partitions des deux ensembles. C'est cette description qui a inspiré son nom : **correspondance par blocs**.

**Comme** toutes les relations solutions intégrables à la méthodologie de l'analyse relationnelle, elle répond à une définition en termes relationnels. Nous allons la présenter maintenant.

Soit donc  $Z$  une relation de correspondance par blocs sur le croisement de deux ensembles  $Z$  et  $J$ . Elle est **définie** par trois types de contraintes :

- **la binarité** ( $z_{ij} \in (0, 1)$ ),
- **les contraintes d'affectation**,

---

<sup>36</sup> La division par la constante  $m$  nous **place** directement dans le contexte factoriel, sans **affecter** l'optimisation du **critère**.

• les contraintes de la triade impossible.

Les **contraintes d'affectation** garantissent une correspondance bijective entre les classes des deux partitions, c'est-à-dire qu'à toute classe de la partition sur  $I$  correspond une et une seule classe de la partition sur  $J$ , et inversement. Ces contraintes s'expriment linéairement comme suit :

$$\left\{ \begin{array}{l} \sum_{j \in J} z_{ij} \geq 1, \forall i \in I \\ \sum_{i \in I} z_{ij} \geq 1, \forall j \in J \end{array} \right.$$

Remarquons que les contraintes d'affectation ont été introduites, à l'origine, pour répondre exactement au problème de sériation posé dans le domaine du *manufacturing*. Mais on peut tout à fait les exclure du modèle, auquel cas on autorise la création de classes sans correspondance tant du côté de  $I$  que de celui de  $J$ ; on parlera alors de **quasi-sériation**<sup>37</sup>.

Les **contraintes de la triade impossible** ont pour rôle d'assurer la structure en blocs rectangulaires disjoints. Elles tirent leur nom de la propriété caractéristique de ce type de relation, à savoir : pour toute configuration de quatre points  $(i, i') \in I$  et  $(j, j') \in J$ , il ne peut pas y avoir simultanément trois paires en correspondance sur  $I \times J$ .

L'expression linéaire de ces contraintes est donnée par le système d'inégalités :

$$\begin{array}{l} (I) \quad z_{ij} + z_{i'j'} + z_{i'j} - z_{ij'} - 1 \leq 1 \quad \forall (i, i') \in I, \forall (j, j') \in J \\ (II) \quad z_{i'j'} + z_{i'j} + z_{ij} - z_{ij'} - 1 \leq 1 \quad \forall (i, i') \in I, \forall (j, j') \in J \\ (III) \quad z_{i'j} + z_{ij} + z_{i'j'} - z_{ij'} - 1 \leq 1 \quad \forall (i, i') \in I, \forall (j, j') \in J \\ (IV) \quad z_{ij'} + z_{i'j'} + z_{i'j} - z_{ij} - 1 \leq 1 \quad \forall (i, i') \in I, \forall (j, j') \in J \end{array}$$

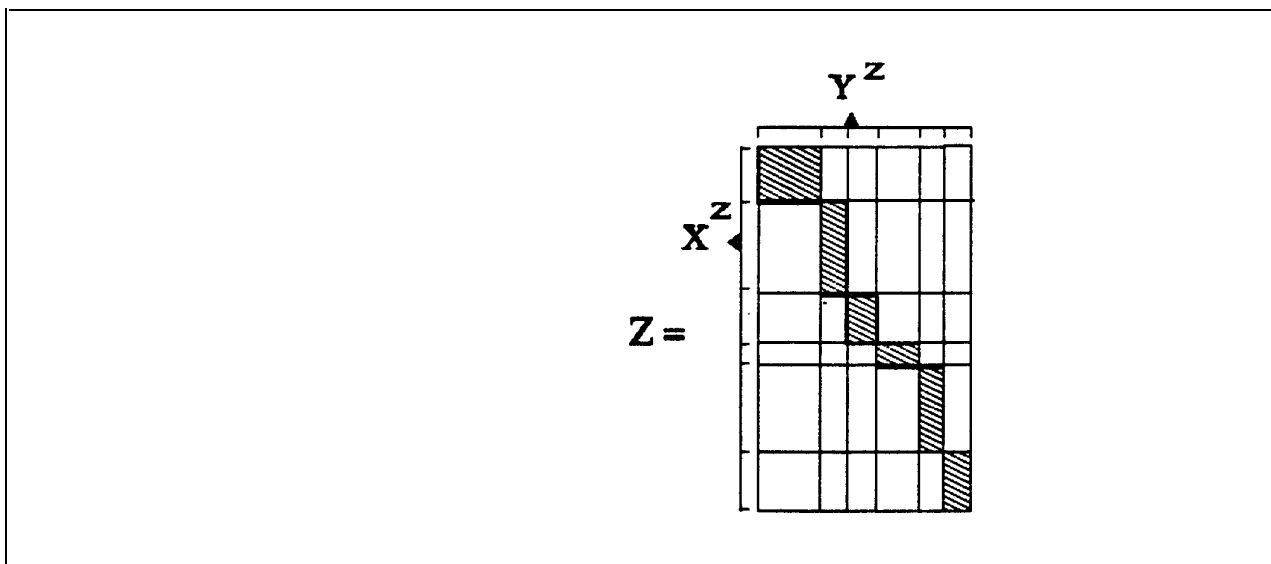
Ces contraintes généralisent la propriété de transitivité à des données rectangulaires. Du fait de cette analogie avec les problèmes de classification, nous parlerons souvent de **classification croisée** pour désigner le processus de sériation par blocs.

Intéressons-nous de plus près à la structure d'une relation de correspondance par blocs  $Z$  sur  $I \times J$ . Elle se décompose en trois parties :

- une partition de l'ensemble  $I$ , notée  $X^Z$ , "projection" de  $Z$  sur  $I$ ,
- une partition de l'ensemble  $J$ , notée  $Y^Z$ , "projection" de  $Z$  sur  $J$ ,
- une correspondance entre les classes de ces deux partitions (bijective si l'on obtient une sériation, non bijective si l'on a affaire à une quasi-sériation).

<sup>37</sup> Nous désignerons indifféremment sous le nom de *sériation*, ou *correspondance par blocs* la relation résultant du procédé de sériation. Le terme de *quasi-sériation* caractérise, quant à lui, une forme de résultat et non pas un mode de traitement.

Le schéma suivant présente une correspondance par blocs  $Z$  et ses différentes composantes :



### 7.2.2.1 Critères de Sériation

Le critère usuellement adopté pour mesurer l'adéquation d'une correspondance par blocs  $Z$  à une matrice relationnelle  $R$  croisant deux ensembles  $I$  et  $J$ , est le **critère général de sériation**. Sa formulation est la suivante :  $\sum \sum [r_{ij} z_{ij} + \bar{r}_{ij} (1 - z_{ij})]$ .

Il généralise le critère de Condorcet au **cas** du croisement de deux ensembles distincts et opère dans le même esprit que celui-ci, c'est-à-dire qu'il repose sur une notion de majorité et traite de la même façon les éléments lignes et les éléments colonnes.

Cependant, le fait d'être en présence de deux ensembles distincts peut, dans certains problèmes, s'avérer fondamental. Il faut alors envisager des critères qui feront jouer des rôles spécifiques aux éléments de deux ensembles.

Sans entrer dans le détail de la construction de tels critères nous pouvons toutefois en proposer trois relativement classiques :

- $\sum_i \sum_j \left( \frac{r_{ij}}{r_{i \bullet}} - \frac{1}{m} \right) z_{ij}$
- $\sum_i \sum_j \left[ r_{ij} - \frac{1}{2} \left( \frac{r_{i \bullet}}{m} + \frac{r_{\bullet j}}{n} \right) \right] z_{ij}$
- $\sum_i \sum_j \left[ \frac{r_{ij}}{r_{i \bullet} r_{\bullet j}} - \frac{1}{2} \left( \frac{1}{m r_{\bullet j}} + \frac{1}{n r_{i \bullet}} \right) \right] z_{ij}$

Ils prennent en compte, de façon croissante, le caractère non symétrique de la sériation et conduisent à une solution dans laquelle les éléments lignes et les éléments colonnes n'ont pas nécessairement des rôles **duaux**.

### 7.2.3 La Quadri-Décomposition

La procédure de **quadri-décomposition**, développée dans le cadre de la méthodologie relationnelle [Bede89a], permet de prendre en compte des informations sur les **données à différents niveaux** et de les **traiter simultanément**.

Avec les méthodes précédentes, on pouvait certes déjà aborder un grand nombre de problèmes de classification, mais chaque méthode, dotée de ses outils propres, couvrait un champ bien spécifique. La quadri-décomposition englobe, en une seule approche, ces différents domaines et s'ouvre sur des applications plus vastes, jusqu'alors inexploitées à l'aide des outils existants.

Le principe de base de la quadri-décomposition repose essentiellement sur la manière de prendre en compte les données dans une matrice "à entrées multiples".

Comme pour la majorité des problèmes de classification, on dispose, au départ, d'un ensemble  $I$  regroupant  $n$  individus et d'un ensemble  $J$  constitué de  $m$  variables qualitatives ou attributs descriptifs des individus.

Plus encore que dans les cas précédents, la dénomination "**individus/variables**" est à prendre ici dans une très large acception. En effet, le concept même de **quadri-décomposition** joue sur l'absence de distinction entre ces notions, lors de la prise en compte des données, pour ne la rétablir qu'au niveau de la solution.

Dans la configuration la plus générale, on considère que l'on possède des informations à quatre niveaux :

- sur les relations qu'entretiennent les éléments de  $I$ ,
- sur les relations qu'entretiennent les éléments de  $J$ ,
- sur les relations croisant les éléments de  $I$  avec ceux de  $J$ ,
- sur les relations croisant les éléments de  $J$  avec ceux de  $I$ .

Ces informations sont restituées par quatre matrices, que nous noterons respectivement:

$N$  **matrice de la relation sur  $I \times I$** ;  $N = (n_{i,i'})$  avec  $1 \leq i \leq n$  et  $1 \leq i' \leq n$

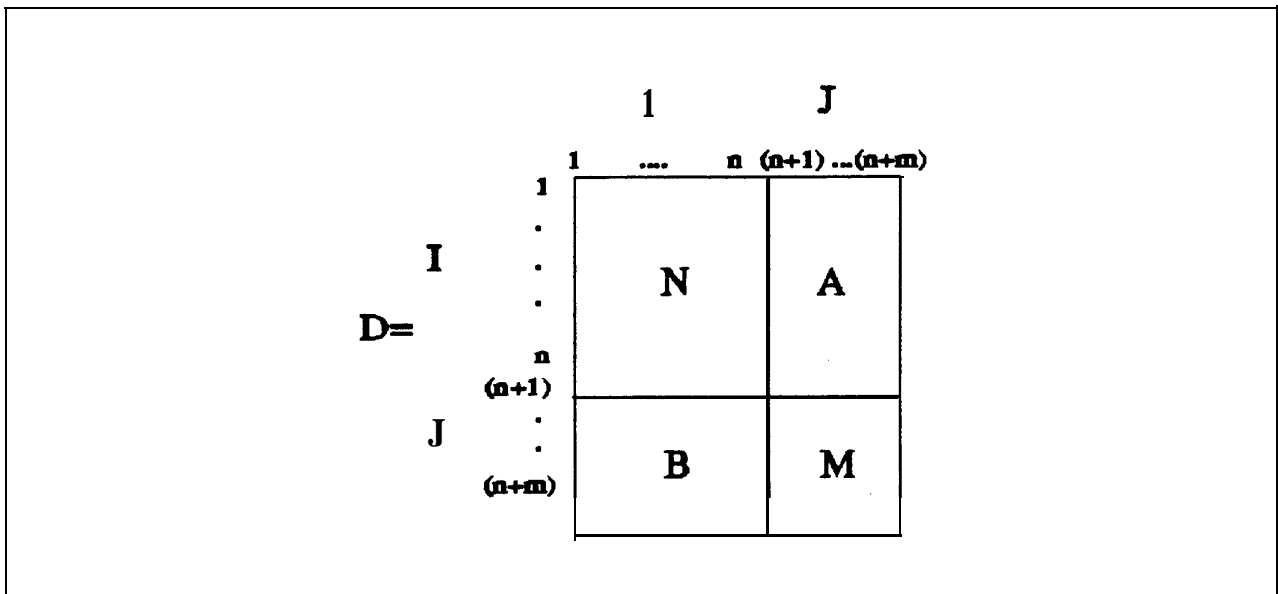
$M$  **matrice de la relation sur  $J \times J$** ;  $M = (m_{j,j'})$  avec  $1 \leq j \leq m$  et  $1 \leq j' \leq m$

$A$  **matrice de la relation sur  $I \times J$** ;  $A = (a_{ij})$  avec  $1 \leq i \leq n$  et  $1 \leq j \leq m$

$B$  **matrice de la relation sur  $J \times I$** ;  $B = (b_{ji})$  avec  $1 \leq j \leq m$  et  $1 \leq i \leq n$

On regroupe ces données à l'intérieur des quatre blocs d'une unique matrice  $D$ , dite **matrice de quadri-décomposition**, en les disposant suivant un schéma précis.

Cette matrice, sur laquelle reposent les fondements de la méthode, se présente sous la forme :



$D$  est, par construction, une matrice carrée de taille  $(n + m) \times (n + m)$ . Ses autres caractéristiques dépendent, quant à elles, des structures propres de ses composantes  $A$ ,  $B$ ,  $N$  et  $M$ .

L'ensemble des objets à classifier est, cette fois, l'ensemble  $L = I \cup J$ , réunion des deux ensembles de départ. Dans la matrice  $D = (d_{ij})$  individus et variables jouent désormais un rôle équivalent, c'est pourquoi nous les désignerons par le terme unique "objets". Nous rejoignons alors le domaine de l'agrégation des similarités, puisque le problème consiste encore à classifier un ensemble d'objets. Cette fois, cependant, l'ensemble en question n'est plus ni  $I$ , ni  $J$ , mais la réunion des deux, c'est-à-dire  $L$ . La solution du problème étant une partition de  $L$ , nous noterons  $Q$  la relation d'équivalence correspondante sur  $L \times L$ .

On cherche donc à regrouper, au sein de classes **homogènes**, les objets les plus ressemblants tant par les liaisons qu'ils entretiennent que par leurs oppositions communes aux autres objets. Il est alors possible, dans cette optique, de choisir de mesurer l'adéquation de la solution aux données par l'intermédiaire du désormais classique critère de Condorcet. Dans le contexte de la quadri-décomposition il prend la forme :

$$\delta(Q) = \sum_{i \in L} \sum_{i' \in L} d_{ii'} \cdot q_{ii'} + \sum_{i \in L} \sum_{i' \in L} \bar{d}_{ii'} \cdot (1 - q_{ii'})$$

Par analogie avec les modèles précédents, la matrice  $\bar{D}$  représente le complémentaire de  $D$  au maximum de relations possibles liant une paire d'objets sur le croisement considéré.

Précisons que la gamme de critères utilisables en quadri-décomposition est la même que celle de l'Analyse Relationnelle. C'est, comme nous allons le voir, sur le choix des structures de solutions que cette méthodologie se distingue.

Comme tous les problèmes relevant de l'analyse relationnelle, **définis** dans le chapitre précédent, le problème de **quadri-décomposition** se modélise sous la forme d'un programme linéaire. Celui-ci n'est autre que le programme standard de classification, mais appliqué, cette fois, à la matrice de données  $D$  et à la relation d'équivalence cherchée  $Q$ . Il s'écrit :

$$\text{Max } \delta(Q), \text{ avec } Q \text{ relation d'équivalence sur } L$$

A l'instar des données, le critère, linéaire en  $Q$ , se décompose en quatre parties :

$$\delta(Q) = \delta_I(Q) + \delta_J(Q) + \delta_{IJ}(Q) + \delta_{JI}(Q)$$

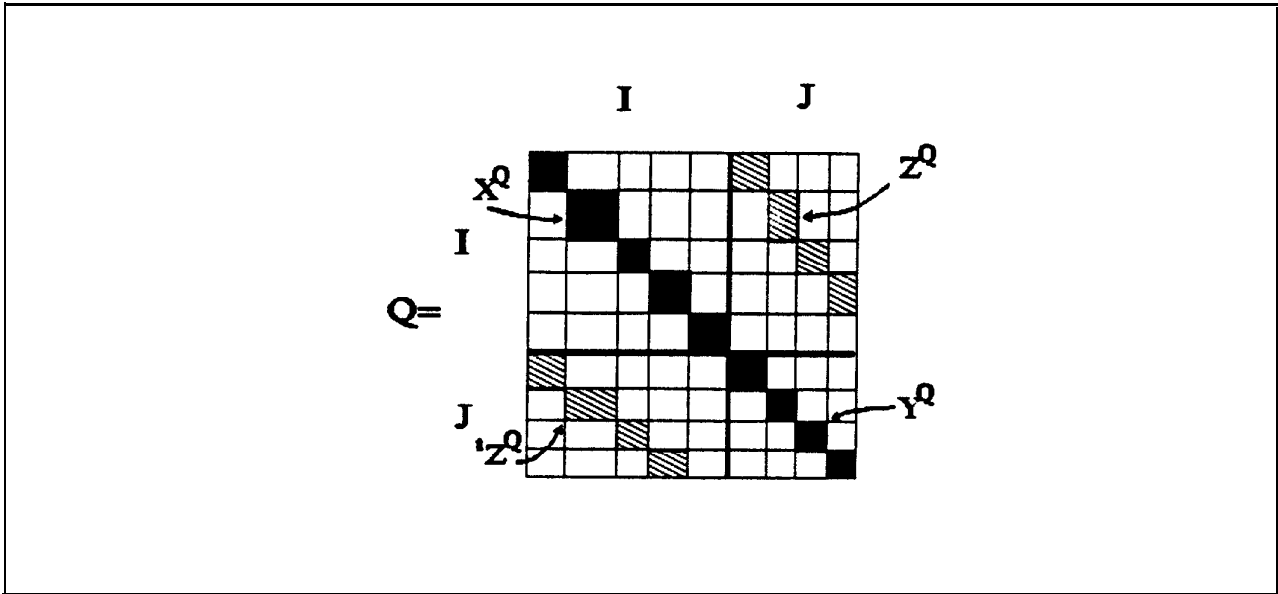
Cette décomposition sert, d'un point de vue formel, pour montrer l'équivalence entre le modèle de quadri-décomposition et les différents modèles de classification et, d'un point de vue pratique, pour "activer" ou "désactiver" certains blocs de la quadri-décomposition, selon les données dont on dispose **et/ou** les traitements que l'on désire effectuer.

La solution  $Q$  est une partition de l'ensemble  $L$ . Ses classes sont donc constituées d'objets issus de  $L = I \cup J$ , c'est-à-dire d'individus **et/ou** de variables provenant des deux ensembles de base. Tout comme les données et le critère, la solution est soumise au processus de quadri-décomposition.

Sur chacun des blocs de la quadri-décomposition se calque une relation binaire, restriction de  $Q$  au croisement correspondant, et dont la structure est déterminée par les contraintes de relation d'équivalence portant sur  $Q$ . Les propriétés de ces quatre relations se déduisent directement de celles de  $Q$ ; on montre que :

- sur le croisement  $I \times I$  se calque une relation d'équivalence que nous noterons  $X^Q$
- sur le croisement  $J \times J$  se calque une relation d'équivalence que nous noterons  $Y^Q$
- sur le croisement  $I \times J$  se calque une relation de correspondance par blocs que nous noterons  $Z^Q$
- sur le croisement  $J \times I$  se calque une relation de correspondance par blocs qui est simplement la transposée de  $Z^Q$ .

Finalement, la solution  $Q$  **quadri-décomposée** peut se résumer à l'aide du schéma suivant



La structure de cette solution générale, liée aux contraintes de quadri-décomposition, subit des transformations en fonction de la configuration de données traitée. Le remplissage plus ou moins complet des quatre blocs de  $D$  et le type, symétrique ou non symétrique, des données exercent une action sur les différents constituants de  $Q$ .

C'est précisément cette **adaptabilité de la solution** qui permet de traiter des problèmes de natures très différentes à l'aide d'une méthodologie unique.

Notons que grâce au modèle général, qui combine approche relationnelle et programmation linéaire, **rien n'est fixé a priori** : ni le nombre de classes des partitions  $X^Q$  et  $Y^Q$ , ni le nombre de blocs en correspondance dans  $Z^Q$ . Seule l'information contenue dans les données influe sur ces paramètres caractéristiques de la solution. C'est, rappelons-le, l'une des idées maîtresses de l'analyse relationnelle, et la quadri-décomposition, **affiliée** à cette méthodologie, hérite de cette propriété.

La quadri-décomposition, comme les autres méthodes relevant de l'analyse relationnelle, s'avère bien adaptée aux données de type binaire. C'est dans cette catégorie qu'entrent, en particulier, les données lexicographiques [Bede89b] (**présence/absence**) ou les traditionnels tableaux **disjonctifs**. Mais elle peut être étendue à des types de données plus générales, au même titre que toutes les méthodes d'analyse relationnelle, car ses principes fondamentaux sont indépendants de la nature des données.

C'est en fait le choix du critère de classification qui va permettre d'exploiter des données de différents types: binaire, comptages, fréquences, intensités, etc., la décomposition du critère sur les quatre blocs autorisant même la prise en compte simultanée de plusieurs de ces types.

Nous allons maintenant passer brièvement en revue les structures des différents problèmes que la méthode de **quadri-décomposition** permet de résoudre. Partant des plus simples, qui se ramènent à des problèmes déjà connus en analyse relationnelle, nous finirons par les plus complets à travers lesquels la **quadri-décomposition** trouve ses applications les plus originales.

Nous définissons sous le nom de **configurations en données simples** les cas où l'on ne dispose d'aucune information sur les relations intra ensembles, c'est-à-dire sur  $I \times I$  et  $J \times J$ . En ce qui concerne les croisements inter ensembles, seules les données sur  $I \times J$  sont connues. On ne peut ainsi traiter qu'une seule matrice  $A$  contenant l'information sur les relations qu'entretiennent les objets à classifier. Les matrices  $N$  et  $M$ , qui portent respectivement sur  $I \times I$  et  $J \times J$  dans la structure de quadri-décomposition, ont donc ici une fonction "neutre".

Le **principe de neutralisation** est l'un des éléments fondamentaux de la méthodologie. Cette neutralisation dérive de la structure du critère. S'il l'on veut neutraliser, ou désactiver, l'un des 4 blocs, il suffit de poser pour chacune des paires d'objets  $(i, i')$  de ce bloc  $d_{ii'} - \bar{d}_{ii'} = 0$ . On s'assure ainsi que le critère prend une valeur constante sur ce bloc, à savoir  $\sum \sum (1 - d_{ii'})$ . Ceci a pour conséquence que la comparaison des objets de ce bloc n'intervient pas dans le processus d'optimisation, soit encore, que la solution est construite sans prendre en compte les informations concernant ce bloc.

Ainsi, on peut n'exploiter dans la matrice **quadri-décomposée**  $D$  qu'un seul des 4 blocs que l'on remplit avec la matrice de données  $A$ . On a alors deux configurations possibles :

- si  $A$  est une matrice carrée symétrique, le traitement de  $A$  par quadri-décomposition conduit simplement à la classification de ses éléments,
- si  $A$  est carrée non symétrique ou bien rectangulaire, la traiter par le processus de quadri-décomposition est équivalent à la soumettre à la sériation (sans les contraintes d'affectation).

La première configuration correspond donc au modèle de classification simple, la seconde, au modèle de sériation simple. Ainsi, pour ces configurations en données simples, la procédure de **quadri-décomposition** permet d'intégrer, dans un "**moule**" unique, des structures de données, qui relevaient auparavant de modèles distincts.

Nous allons évoquer, maintenant, les configurations de données que seule la structure de quadri-décomposition permet de gérer. Il s'agit de prendre en compte, simultanément, plusieurs natures d'informations. Ainsi, par opposition aux configurations en données simples, nous désignerons sous le label de **données multiples** les cas où l'on dispose de plus d'un type d'information sur les différents croisements des ensembles  $I$  et  $J$ . C'est-à-dire que l'on connaît 2, 3, voire 4 des matrices  $A$ ,  $B$ ,  $N$  et  $M$  à intégrer à la matrice de **quadri-décomposition**  $D$ . Les autres blocs de  $D$ , non occupés par une relation donnée, seront neutralisés.

Selon le nombre et la position des relations traitées, on définit quatre problèmes auxquels correspondent quatre modèles spécifiques. Ces modèles, caractéristiques de l'approche quadri-décomposée, sont les suivants :

1. le modèle de classifications enrichie,
2. le modèle de sériation enrichie,
3. le modèle de classification conditionnée,
4. le modèle de sériation conditionnée.

Chacune de ces méthodes répond à une préoccupation bien spécifique. Avant d'opter pour un modèle, il est important de définir clairement le problème que l'on cherche à résoudre et la structure que l'on veut voir émerger des données.

La classification sur  $I$  fournit une partition optimale des éléments en ligne, les colonnes n'apparaissant pas directement dans la solution. Inversement, si l'on partitionne  $J$ , ce sont les lignes qui sont exclues de la solution. Ce n'est qu'à *posteriori*, par une analyse détaillée du résultat, que l'on peut quantifier la contribution de chacun des éléments de l'autre ensemble aux **différentes** classes de la partition obtenue. Mais il est plus **difficile** de déceler, au vu de ces indicateurs, quels sont, par exemple, les groupes de modalités qui ont **entraîné** le partitionnement des individus.

A l'inverse, la méthode de sériation simple conduit à la classification croisée optimale des deux ensembles  $I$  et  $J$ ; on dispose là, véritablement, de toute l'information sur la meilleure correspondance entre groupes d'éléments de deux ensembles. Cette sériation (ou **quasi-sériation** si l'on ne tient pas compte des contraintes d'affectation) fournit, de par sa structure, une partition de  $I$  et une partition de  $J$ , mais elles n'ont malheureusement pas, dans la plupart des cas, un caractère optimal.

Les 4 modèles de quadri-décomposition combinent, en quelque sorte, les préoccupations des méthodes précédentes, en permettant l'**optimisation conjointe des trois formes d'ajustements** et la restitution directe de ces différentes informations au niveau de la solution :

- la classification des éléments lignes, à travers  $X^Q$ ,
- la classification des éléments colonnes, à travers  $YQ$ ,
- la classification croisée des deux ensembles, à travers  $ZQ$ .

Ils privilégient chacun l'un des 3 problèmes tout en accordant une place, plus ou moins importante, aux autres.

Les modèles enrichis constituent le premier pas vers la combinaison de ces trois problématiques. Par rapport aux modèles simples, ils garantissent que la solution tient compte des informations sur les relations croisées dans l'optique classification, et des informations liant les éléments à l'intérieur de leur propre ensemble dans l'optique sériation.

Les modèles conditionnés sont, quant à eux, les plus "exigeants" sur cette combinaison des problèmes. Ainsi, en classification conditionnée, on veut que les partitions des deux ensembles soient les meilleures possibles et que, simultanément, la correspondance entre les classes de ces deux partitions ait un caractère optimal.

Parallèlement, pour le modèle de sèriation enrichie, l'accent est mis sur la recherche de la meilleure classification croisée, qui garantisse également un très bon partitionnement à l'intérieur de chacun des deux ensembles.

### **7.3 Conclusion**

---

Nous venons de donner, dans ce chapitre, un résumé des possibilités des techniques de **l'Analyse Relationnelle** applicables à la **Veille Technologique**, c'est-à-dire aux fichiers avec information non standardisée et se ramenant essentiellement à des tableaux de "présence-absence" ou à des tableaux fréquentiels. La haute souplesse de prise en compte des données qu'offre **l'Analyse Relationnelle** n'a été que partiellement présentée ici, mais est sans nul doute un atout supplémentaire important par rapport au point sur lequel nous avons beaucoup insisté la non **fixation** d'hypothèses de départ permettant l'application non triviale d'algorithmes.



# *Seconde annexe technique*

**Développement d'indicateurs pour  
l'Analyse Factorielle-Relationnelle**

*Chantai Bédécarrax et Charles Huot*

Etude du CEMAP, n° MAP-005, mai 1992

## 8.1 Introduction

---

Cet article fait suite aux recherches de **F.Marcotorchino** publiées en 1991 sous le titre de **L'Analyse Factorielle-Relationnelle : Parties I et II**. Dans cette publication [Marc91a], l'auteur poursuit ses travaux de 1988 sur les Liaisons Analyse **Factorielle-Analyse Relationnelle (1): "Dualité Burt-Condorcet"** [Marc89], en consacrant une partie importante à la recherche des expressions relationnelles des inerties factorielles. Cette réécriture sous une forme relationnelle lui permet de déduire un certain nombre de propriétés sur les inerties ainsi que sur la majoration du critère de Condorcet pondéré. **Enfin** F.Marcotorchino présente un certain nombre de propriétés de la solution du **critère** de Condorcet pondéré, en montrant que cette solution est souvent la même que celle qui optimise le **critère** de la **différence** inertielle.

L'objectif de notre article est de déterminer des indicateurs permettant d'obtenir un résultat détaillé de la solution d'une analyse factorielle-relationnelle. Dans leur grande majorité, ces indicateurs sont la réécriture sous une forme relationnelle d'indicateurs inertiels connus.

Nous présenterons un exemple d'application de ces indicateurs lors d'une classification factorielle relationnelle de l'ensemble des **félidés**, dont la classification relationnelle fut présentée à de nombreuses reprises dans les travaux de F.Marcotorchino ou P.Michaud [Marc81, Mich85].

## 8.2 Développements d'Indicateurs pour l'interprétation des résultats d'une Analyse Factorielle Relationnelle

---

Rappelons que nous nous situons dans le contexte général de l'AFCM d'un nuage de N individus décrits par M variables à modalités. On note  $K = (k, \dots)$  le tableau disjonctif complet décrivant les données. Les autres notations, que nous ne détaillerons pas systématiquement, sont empruntées aux méthodologies factorielles et relationnelles.

### 8.2.1 Inertie totale du nuage des Individus

---

- globalement

$$I_T = \sum_{i \in I} f_i \|O_i - G\|^2 = \sum_{i \in I} f_i d_{iG}^2(O_i, G)$$

- par classe

$$I_T = \sum_{k=1}^K I_k \text{ avec } I_k = I_b + I_w \forall k = 1, \dots, K$$

où  $I_b$  représente l'inertie between de la classe  $k$  et  $I_w$  l'inertie within de la classe  $k$ .

### 8.2.2 Inertie Inter-classe

Dans son article F. Marcotorchino déduit l'écriture relationnelle de  $I_B$  de la différence entre  $I_T$  et  $I_W$ . Nous allons détailler ici les étapes de son obtention, en définissant au passage, l'inertie inter d'une classe  $I_b$ .

$$I_B = \sum_{k=1}^K I_b$$

$$\text{avec } I_b = f_k \|G^* - G\|^2 = f_k \sum_{j=1}^J \frac{(f_j^* - f_{.j})^2}{f_{.j}}$$

où  $f_{.j}$  représente la coordonnées du centre de gravité  $G$  du nuage:

$$f_{.j} = \frac{k_{.j}}{k_{..}} = \frac{k_{.j}}{NM}$$

$f_k$  représente la masse de la classe  $k$ :

$$f_k = \sum_{i \in k} f_{i.} = \sum_{i \in k} \frac{k_{i.}}{k_{..}} = \frac{|k|M}{NM} = \frac{|k|}{N}$$

et  $f_j^*$  représente les coordonnées du centre de gravité  $G^*$  de la classe  $k$ :

$$f_j^* = \frac{\sum_{i \in k} f_{i.} f_{ij}}{\sum_{i \in k} f_{i.}} = \frac{\sum_{i \in k} \frac{k_{i.}}{k_{..}} \frac{k_{ij}}{k_{i.}}}{\sum_{i \in k} \frac{k_{i.}}{k_{..}}} = \frac{\sum_{i \in k} k_{ij}}{\sum_{i \in k} k_{i.}} = \frac{\sum_{i \in k} k_{ij}}{|k|M}$$

$$\begin{aligned} \text{d'où } I_b &= \frac{|k|}{N} \sum_{j=1}^J \frac{NM}{k_{.j}} \left[ \frac{\sum_{i \in k} k_{ij}}{|k|M} - \frac{k_{.j}}{NM} \right]^2 \\ &= |k| \sum_{j=1}^J \frac{M}{k_{.j}} \left[ \frac{(\sum_{i \in k} k_{ij})^2}{|k|^2 M^2} + \frac{k_{.j}^2}{N^2 M^2} - \frac{2k_{.j}(\sum_{i \in k} k_{ij})}{N|k|M^2} \right] \\ &= \sum_{j=1}^J \frac{1}{k_{.j}} \left[ \frac{(\sum_{i \in k} k_{ij})^2}{|k|M} + \frac{k_{.j}^2 |k|}{N^2 M} - \frac{2k_{.j}(\sum_{i \in k} k_{ij})}{NM} \right] \end{aligned}$$

$$\begin{aligned}
&= \sum_{j \in J} \frac{(\sum_{i \in k} k_{ij})(\sum_{r \in k} k_{rj})}{k_j |k| M} + \frac{|k| \sum_{j \in J} k_j}{N^2 M} - \frac{2 \sum_{i \in k} \sum_{j \in J} k_{ij}}{NM} \\
&= \frac{1}{M|k|} \sum_{i \in k} \sum_{r \in k} \sum_{j \in J} \frac{k_{ij} k_{rj}}{k_j} + \frac{|k|}{N} - \frac{2|k|M}{NM} \\
&= \sum_{i \in k} \sum_{r \in k} \frac{\hat{c}_{ir}}{M|k|} - \frac{|k|}{N}
\end{aligned}$$

et finalement, en notations relationnelles :

$$I_B^k = \sum_{i \in k} \sum_{r \in k} \frac{\hat{c}_{ir}}{M} \frac{x_{ir}}{x_{i \cdot}} - \frac{|k|}{N}$$

Remarque:

$$\text{On peut également écrire } I_B = \sum_{i \in k} \sum_{r \in k} \left( \frac{\hat{c}_{ir}}{M} - \frac{1}{N} \right) \frac{x_{ir}}{x_{i \cdot}}$$

où l'on voit mieux apparaître le **coût** de comparaison d'une paire d'individus.

On vérifie qu'au niveau global on retrouve bien la formule relationnelle de l'inertie totale **définie** dans [Marc91a].

$$I_B = \sum_{k=1}^K I_B^k = \sum_{k=1}^K \sum_{i \in k} \sum_{r \in k} \frac{\hat{c}_{ir}}{M|k|} - \sum_{k=1}^K \frac{|k|}{N} = \sum_{k=1}^K \sum_{i \in k} \sum_{r \in k} \frac{\hat{c}_{ir}}{M|k|} - 1$$

soit en notations relationnelles:

$$I_B = \sum_i^N \sum_{i'}^N \frac{\hat{c}_{ii'}}{M} \frac{x_{ii'}}{x_i} - 1$$

### 8.2.3 Inertie between entre classes

Nous définissons l'inertie between entre les classes  $k$  et  $k'$  par:

$$I_{kk'} = \frac{f_k f_{k'}}{f_k + f_{k'}} \|G_k - G_{k'}\|^2$$

$$\text{soit } I_{kk'} = \frac{|k| |k'|}{N(|k| + |k'|)} d_{\frac{1}{2}}^2(G_k, G_{k'})$$

Traditionnellement, cette quantité mesure la perte d'inertie provoquée par la réunion des deux classes  $k$  et  $k'$ . En analyse factorielle-relationnelle, elle conserve naturellement cette signification mais apparaît plus encore comme indicateur de "désaccords"<sup>38</sup> entre classes pour rejoindre la terminologie relationnelle.

Développons  $d_{\frac{1}{2}}^2(G_k, G_{k'})$  :

$$\begin{aligned} d_{\frac{1}{2}}^2(G_k, G_{k'}) &= \sum_{j \in J} \frac{k_{.j}}{k_j} \left( \frac{\sum_{i \in k} k_{ij}}{M|k|} - \frac{\sum_{i \in k'} k_{i'j}}{M|k'|} \right)^2 \\ &= \frac{NM}{M^2} \sum_{j \in J} \frac{1}{k_j} \left[ \frac{(\sum_{i \in k} k_{ij})^2}{|k|^2} + \frac{(\sum_{i \in k'} k_{i'j})^2}{|k'|^2} - 2 \frac{\sum_{i \in k} k_{ij} \sum_{i \in k'} k_{i'j}}{|k| |k'|} \right] \\ &= \frac{N}{M} \sum_{j \in J} \frac{1}{k_j} \left[ \frac{\sum_{i \in k} k_{ij} \sum_{i \in k'} k_{i'j}}{|k|^2} + \frac{\sum_{i \in k'} k_{i'j} \sum_{i \in k} k_{ij}}{|k'|^2} - 2 \frac{\sum_{i \in k} \sum_{i \in k'} k_{ij} k_{i'j}}{|k| |k'|} \right] \end{aligned}$$

<sup>38</sup> voir [Mich85] pour une définition de la notion d'accords-désaccords en Analyse Relationnelle.

$$\begin{aligned}
&= \frac{N}{M} \left[ \frac{1}{|k|^2} \sum_{i \in k} \sum_{r \in k} \left( \sum_{j \in J} \frac{k_{ij} k_{i'j}}{k_j} \right) + \frac{1}{|k'|^2} \sum_{i \in k'} \sum_{r \in k'} \left( \sum_{j \in J} \frac{k_{ij} k_{i'j}}{k_j} \right) - \frac{2}{|k||k'|} \sum_{i \in k} \sum_{r \in k'} \left( \sum_{j \in J} \frac{k_{ij} k_{i'j}}{k_j} \right) \right] \\
&= \frac{N}{M} \left[ \frac{\sum_{i \in k} \sum_{r \in k} \hat{c}_{ii'}}{|k|^2} + \frac{\sum_{i \in k'} \sum_{r \in k'} \hat{c}_{ii'}}{|k'|^2} - \frac{2 \sum_{i \in k} \sum_{r \in k'} \hat{c}_{ii'}}{|k||k'|} \right] \\
&= N \left[ \frac{1}{|k|} \frac{\sum_{i \in k} \sum_{r \in k} \hat{c}_{ii'}}{|k|M} + \frac{1}{|k'|} \frac{\sum_{i \in k'} \sum_{r \in k'} \hat{c}_{ii'}}{|k'|M} - \frac{2 \sum_{i \in k} \sum_{r \in k'} \hat{c}_{ii'}}{|k||k'|M} \right]
\end{aligned}$$

L'inertie **between** entre deux les classes  $k$  et  $k'$  peut donc s'écrire de la façon suivante:

$$I_{\mathcal{B}}^{k'} = \left[ \frac{|k'|}{|k| + |k'|} \frac{\sum_{i \in k} \sum_{r \in k} \hat{c}_{ii'}}{|k|M} + \frac{|k|}{|k| + |k'|} \frac{\sum_{i \in k'} \sum_{r \in k'} \hat{c}_{ii'}}{|k'|M} - \frac{2 \sum_{i \in k} \sum_{r \in k'} \hat{c}_{ii'}}{M(|k| + |k'|)} \right]$$

Or nous savons que  $I_{\mathcal{B}} = \frac{\sum_{i \in k} \sum_{r \in k} \hat{c}_{ii'}}{|k|M} - \frac{|k|}{N}$

d'où

$$I_{\mathcal{B}}^{k'} = \left[ \frac{|k'|}{|k| + |k'|} \left( I_{\mathcal{B}} + \frac{|k|}{N} \right) + \frac{|k|}{|k| + |k'|} \left( I_{\mathcal{B}} + \frac{|k'|}{N} \right) - \frac{2 \sum_{i \in k} \sum_{r \in k'} \hat{c}_{ii'}}{M(|k| + |k'|)} \right]$$

$$I_{\mathcal{B}}^{k'} = \left[ \frac{|k'|}{|k| + |k'|} I_{\mathcal{B}} + \frac{|k|}{|k| + |k'|} I_{\mathcal{B}} + \frac{2|k||k'|}{N(|k| + |k'|)} - \frac{2 \sum_{i \in k} \sum_{r \in k'} \hat{c}_{ii'}}{M(|k| + |k'|)} \right]$$

soit finalement:

$$I_B^{kk'} = \frac{1}{|k| + |k'|} \left[ |k'| I_B^k + |k| I_B^{k'} - 2 \frac{\sum_{i \in k} \sum_{i' \in k'} \hat{G}_{ii'}}{M} + \frac{2|k||k'|}{N} \right]$$

Remarque:

Sachant que  $I_B$  peut également s'écrire  $I_B = \frac{1}{2} \sum_{k=1}^K \sum_{k'=1}^K f_k f_{k'} \|G_k - G_{k'}\|^2$ , on a la relation suivante entre  $I_B$  et les  $I_B^{kk'}$

$$I_B = \frac{1}{2} \sum_{k=1}^K \sum_{k'=1}^K (f_k + f_{k'}) I_B^{kk'}$$

Cette formule sera utilisée plus loin pour construire des indicateurs normalisés de liaison inter classes.

#### 8.2.4 Inertie Intra classe

L'inertie intra classe de la classe  $k$  se définit comme la moyenne des distances des points de la classe au centre de gravité de la classe ou encore comme la moyenne des distances des points de la classe deux à deux. Si l'on développe cette seconde définition il vient que:

$$I_W = \frac{1}{2} \sum_{i \in k} \sum_{i' \in k} f_i f_{i'} d_{ii'}^2(O_i, O_{i'})$$

soit:

i

$$I_w = \sum_{i \in k} \sum_{r \in k} \frac{\bar{\hat{c}}_{ir}}{M|k|}$$

Car, rappelons-le,  $d^2(O_i, O_r) = \frac{2N}{M} \bar{\hat{c}}_{ir}$  (cf (55) de [Marc91a]).

### 8.2.5 Développement d'indicateurs d'analyse du résultat

A la classification du nuage d'individus on peut associer toute une famille d'indicateurs permettant d'analyser le résultat obtenu tant au niveau global qu'au niveau des classes ou au niveau individuel.

Une fois obtenue la partition optimale des individus en termes de critère de Condorcet pondéré, donc également optimale en termes de critère de la **différence** inertielle, il est nécessaire de retourner aux données afin de détailler le résultat obtenu.

Cette étape est fondamentale aussi bien pour la validation que pour l'explication de la partition trouvée.

#### 8.2.5.1 Indicateurs globaux

Les indicateurs globaux fournissent, comme leur nom l'indique, une évaluation générale de la qualité de la partition.

$$\tau(I_B/I_T) = \frac{I_B}{I_T} \text{ et } \tau(I_w/I_T) = \frac{I_w}{I_T}$$

$\tau(I_B/I_T)$  est à comparer à  $\frac{\sum_{i=1}^r \lambda_i}{I_T}$ , part d'inertie expliquée par les  $r$  premiers axes factoriels retenus pour l'analyse.

#### 8.2.5.2 Indicateurs par classe

##### 1. Par classe

Les indicateurs permettent d'évaluer la cohérence (homogénéité, compacité, cohésion) et la distinction (éloignement, séparation) des **différentes** classes de la partition solution.

Pour chacune des classes, les indicateurs peuvent être construits en normalisant par  $I_k$  si l'on s'intéresse à la qualité absolue de la classe ou par  $I_T$  si l'on mesure la part relative de la classe sur l'ensemble de la partition.

$$\text{absolu: } \tau(I_b^k/I_k) = \frac{I_b^k}{I_k} \text{ et } \tau(I_w^k/I_k) = \frac{I_w^k}{I_k}$$

$$\text{relatif: } \tau(I_b^k/I_b) = \frac{I_b^k}{I_b}, \tau(I_w^k/I_w) = \frac{I_w^k}{I_w} \text{ et } \tau(I_k/I_T) = \frac{I_k}{I_T}$$

## 2. Inter classes

Les indicateurs inter classes mesurent les phénomènes d'attraction qu'exercent les classes les unes sur les autres ● Cette fois encore, on peut évaluer la part absolue d'une liaison inter classe ou bien sa part relative rapportée à l'ensemble de la partition.

$$\tau(I_b^{k'}/I_b^k) = \frac{I_b^{k'}}{\sum_{k'=1}^r I_b^{k'}}$$

Nous pouvons proposer un nouvel indicateur qui mesure le rapport entre l'inertie between entre deux classes et l'inertie between globale ●

$$\tau(I_b^{k'}/I_b) = \frac{\frac{|k| + |k'|}{2N} I_b^{k'}}{I_b}$$

### 8.2.5.3 Indicateurs par individu

#### 8.2.5.3.1 Contribution d'un individu à l'inertie totale

Nous trouvons dans la thèse de N. El Ayoubi [Ayoubi90] la formule relationnelle de la contribution absolue d'un individu à l'inertie totale (formule 133).

$$Ca(i) = f_i \cdot \|O^i - G\|^2$$

soit après développements:

$$Ca(i) = \frac{\hat{c}_{ii}}{M} - \frac{1}{N}$$

De la même façon nous pouvons calculer la contribution d'un individu  $i$  à une classe  $k$ , comme nous allons le montrer .

### 8.2.5.3.2 Contribution d'un individu à une classe

A partir de la formule de l'inertie intra-classe

$$I_w = \sum_{i \in k} \sum_{r \in k} \frac{\hat{c}_{ir}^2}{M|k|}$$

on définit la contribution interne d'un individu  $i$  à sa classe par:

$$C^-(i) = \sum_{r \in k} \frac{\hat{c}_{ir}^2}{M|k|}$$

qui après développement de  $\hat{c}_{ir}$ , devient:

$$(1) \quad C^-(i) = \frac{\hat{c}_{ii}^2}{2M} + \sum_{r \in k} \frac{\hat{c}_{i,r'}^2}{2M|k|} - \sum_{r \in k} \frac{\hat{c}_{ir}^2}{M|k|}$$

Nous parlons ici de contribution interne de l'individu à sa classe d'appartenance. Cette quantité mesure en quelques sortes les "désaccords" que l'individu  $i$  génère au sein de sa classe; elle peut être comprise comme la part de l'individu  $i$  à l'inertie *within* de la classe.

Remarque:

La quantité  $C^-(i)$ , dans l'approche factorielle, se définit par:

$$C^-(i) = \frac{1}{2} f_i \cdot \sum_{r \in k} f_{r'} \cdot \|O_i - O_r\|^2$$

Nous aurons pu choisir d'adopter l'autre version de décomposition en posant:

$$C^*(i) = f_i \cdot \|O_i - G_k\|^2$$

qui nous aurait conduit à:

$$C_{\star}^{-}(i) = \frac{1}{M} \left[ \hat{c}_{ii} + \frac{\sum_{r \in k} \sum_{r' \in k} \hat{c}_{r r'}}{|k|^2} - 2 \frac{\sum_{r \in k} \hat{c}_{i r'}}{|k|} \right]$$

Les deux quantités  $C^{-}(i)$  et  $C_{\star}^{-}(i)$  sont différentes au niveau individuel, mais leurs sommes sur  $i$ , dans un cas comme dans l'autre, nous ramène à  $I_{\star}$  ●  
 C'est pour des raisons de cohérence de notations que nous avons adopté la définition  $C^{-}(i)$ .

Nous allons maintenant définir la contribution externe de l'individu  $i$  à sa classe,  $C^{+}(i)$ .

Puisqu'au niveau d'une classe  $k$

$$I_k = \sum_{i \in k} \sum_{r \in k} \frac{\hat{c}_{i r'}}{M|k|} - \frac{|k|}{N}$$

nous posons:

$$(2) \quad C^{+}(i) = \sum_{r \in k} \frac{\hat{c}_{i r'}}{M|k|}$$

Cette quantité mesure les "accords" que l'individu  $i$  génère au sein de sa classe; elle peut être comprise comme la part de l'individu  $i$  à l'inertie between de sa classe.

### 8.2.5.3.3 Contribution d'un individu selon le critère de la différence d'inertie

Rappelons au préalable quelques notations utiles dans cette partie.

Soit  $i$  un individu quelconque de  $I$  et  $k$  une classe quelconque de  $P^*$ ,

- en termes de critère de Condorcet pondéré on a :

$$C(i/k) = \sum_{r \in k} \frac{\hat{c}_{ir} - \hat{c}_{i'r}}{M} = \text{contribution de l'individu } i \text{ à la classe } k.$$

$$C(k) = \sum_{i \in k} C(i/k) = \sum_{i \in k} \sum_{r \in k} \frac{(\hat{c}_{ir} - \hat{c}_{i'r})}{M} = \text{contribution de la classe } k.$$

$$\text{Par construction, } \hat{C}(P^*) = \sum_{k \in P^*} C(k) + \text{cte est maximum.}$$

- en termes de critère inertiel on a :

$$CI(i) = \frac{C(i/k)}{|k|} = \sum_{r \in k} \frac{\hat{c}_{ir} - \hat{c}_{i'r}}{M|k|}$$

$$CI(k) = \frac{C(k)}{|k|} = \sum_{i \in k} \sum_{r \in k} \frac{\hat{c}_{ir} - \hat{c}_{i'r}}{M|k|}$$

$$CI(P) = \sum_{k \in P^*} CI(k)$$

Nous retrouvons, avec l'aide des formules (1) et (2), la formule de la contribution moyenne des éléments d'une classe au sens du critère de la différence **d'Inertie**, soit:

$$CI(i) = C^+(i) - C^-(i)$$

Les contributions individuelles, telles que nous venons de les définir, s'inscrivent dans la logique globale du critère de la différence inertielle qui a présidé au partitionnement des individus. Ainsi, si l'on remonte la chaîne "individu-classe-global", on retrouve le critère de départ.

En effet:

$$1. CI(i) = C^+(i) - C^-(i) = \sum_{i' \in k} \frac{\hat{c}_{ii'} - \bar{\hat{c}}_{ii'}}{M|k|}$$

$$2. CI(k) = \sum_{i \in k} CI(i)$$

$$= \sum_{i \in k} \sum_{i' \in k} \frac{\hat{c}_{ii'} - \bar{\hat{c}}_{ii'}}{M|k|}$$

$$3. CI(P) = \sum_{k=1}^{\kappa} CI(k)$$

$$= \sum_{k=1}^{\kappa} \sum_{i \in k} \sum_{i' \in k} \frac{\hat{c}_{ii'} - \bar{\hat{c}}_{ii'}}{M|k|}$$

$$= \sum_{i \in I} \sum_{i' \in I} \left( \frac{\hat{c}_{ii'} - \bar{\hat{c}}_{ii'}}{M|k|} \right) \frac{x_{ii'}}{x_i}$$

$$= I_B - I_W + 1$$

### 8.3 Application à la classification factorielle-relationnelle des félidés

Afin d'illustrer les indicateurs présentés dans le chapitre précédent, nous montrerons leur application dans le cadre d'une analyse factorielle-relationnelle. Cette analyse porte sur un ensemble de  $N = 30$  félidés décrits par  $M = 14$  variables (c.f. tableau 1 en annexe) et  $P = 36$  modalités. On trouve dans le tableau 2 (en annexe) la correspondance entre le codage des modalités et leurs significations.

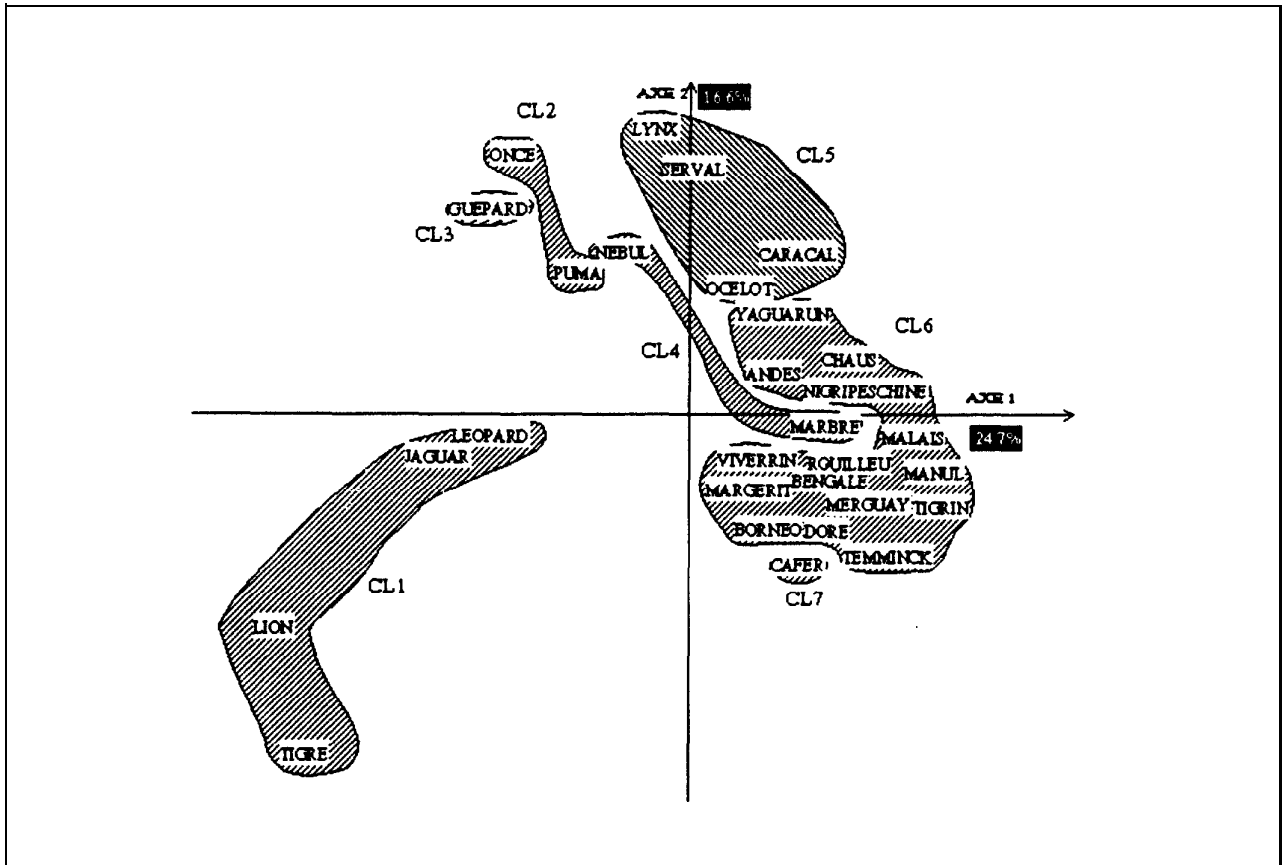
La partition du critère de Condorcet pondéré possède  $\kappa = 7$  classes pour une valeur du critère :

$$CP = 54.7$$

L'application des conditions de transferts et **d'échanges**, définies dans [Marc91a], montre que cette partition est également optimale vis-à-vis du critère de la **différence** inertielle. Ce critère atteint la valeur  $CI = .2435$

C'est donc cette partition que nous allons représenter sur les diagrammes factoriels et que nous allons analyser à l'aide des indicateurs définis dans l'article.

### 8.3.1 Diagrammes factoriels et représentation des classes

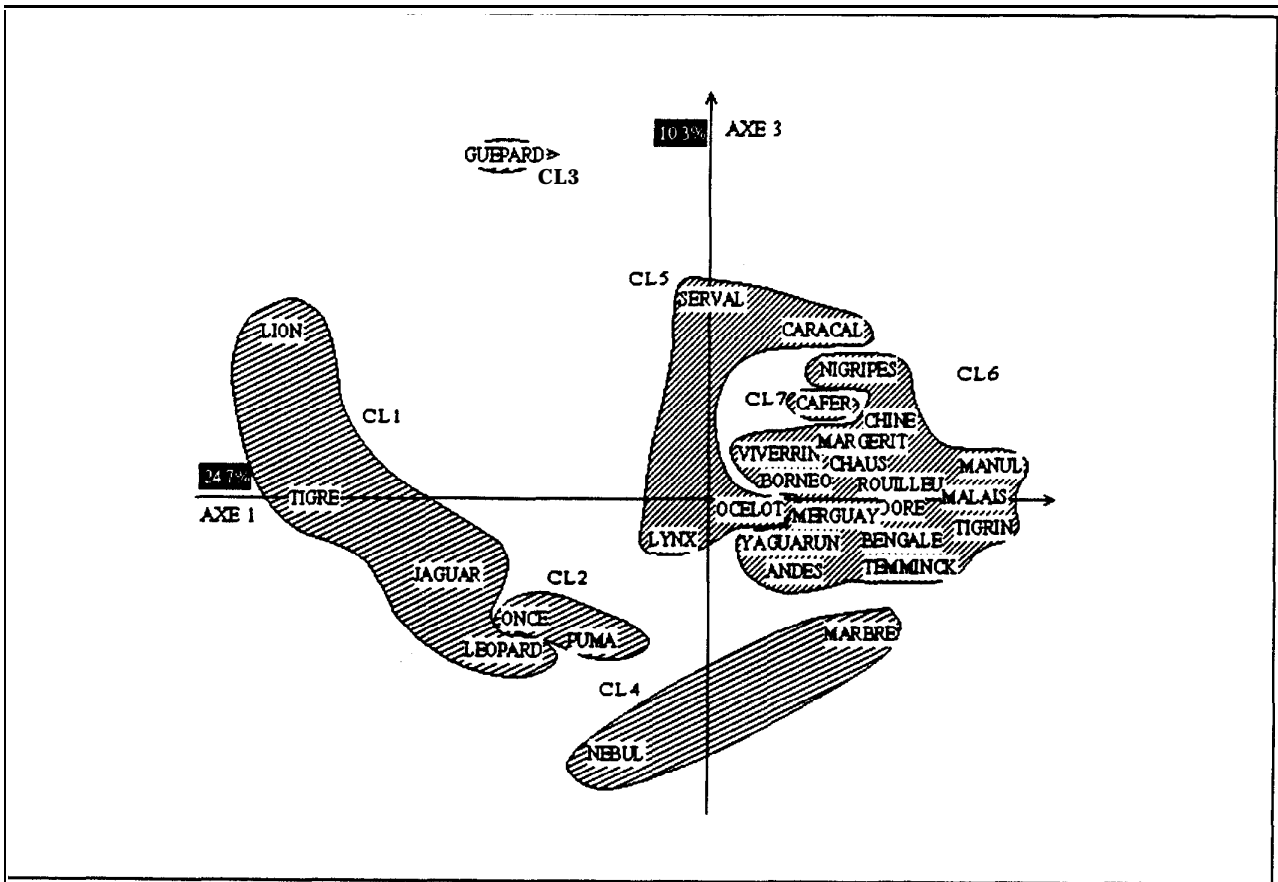


Analyse Factorielle Relationnelle sur l'exemple des félidés, projection sur les axes 1 et 2. (41,3 % d'inertie projetée)

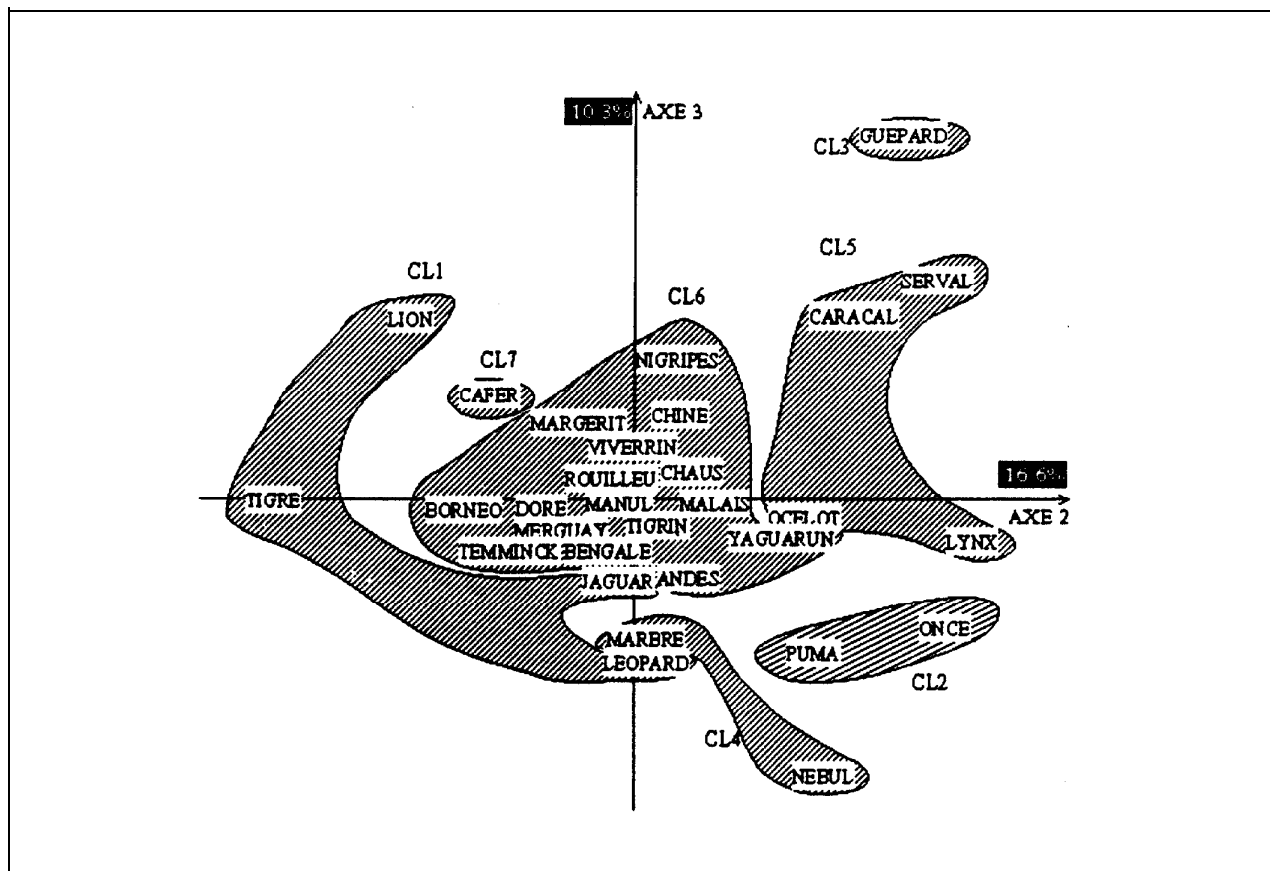
Nous avons retenu les trois premiers axes de l'AFCM du nuage des individus car à eux seuls ils expliquent près de 52 % de l'inertie (on trouvera en annexe toutes les valeurs propres de l'analyse et les coordonnées des individus sur les 3 premiers axes).

Les classes de la partition sont représentées sur les trois diagrammes factoriels.

L'examen successif des 3 figures permet de comprendre la formation de certaines classes à travers le jeu de la rotation dans l'espace.



Projection sur les axes 1 et 3. (35 % d'inertie projetée)



Projection sur les axes 2 et 3. (26,9 % d'inertie projetée)

### 8.3.2 Indicateurs globaux

Inertie Between =  $1.9075 - 1 = 0.9075$   
 Inertie Within =  $0.6640$   
 --> Inertie Totale =  $1.5715$

$\tau(I_B/I_T) = 57,7\%$  c'est la part d'inertie expliquée par la partition. Rappelons que les trois premiers axes de l'analyse factorielle restituent 51,6 % de l'inertie totale. Il faudrait analyser un axe supplémentaire pour dépasser le potentiel explicatif de la partition. On arriverait alors à 61,2% mais le détail du résultat de l'analyse factorielle s'en trouverait d'autant compliqué.

L'une des avancées majeures qu'apporte l'analyse factorielle relationnelle, c'est précisément de faciliter le travail d'interprétation des résultats, dans un environnement

méthodologique cohérent. Les indicateurs construits dans le cadre relationnel sont entièrement compatibles avec la vision factorieliste du résultat.

### **8.3.3 Indicateurs par classe**

Intéressons-nous maintenant au détail du résultat par classe.

Tableau des  $J_b$ ,  $J_w$  et  $J_t$  pour les 7 classes de la partition

	<b>Between</b>	<b>Within</b>	<b>Totale</b>	<b>Taille</b>
cl 1	<b>8.2617</b>	<b>8.1662</b>	<b>0.4278</b>	<b>4</b>
cl 2	<b>8.0964</b>	<b>0.8429</b>	<b>8.1393</b>	<b>2</b>
cl 3	<b>0.1514</b>	<b>0.0000</b>	<b>0.1514</b>	<b>1</b>
cl 4	<b>0.1055</b>	<b>0.0320</b>	<b>0.1375</b>	<b>2</b>
cl 5	<b>8.1176</b>	<b>0.1027</b>	<b>8.2284</b>	<b>4</b>
cl 6	<b>0.1176</b>	<b>0.3202</b>	<b>0.4378</b>	<b>16</b>
cl 7	<b>0.0573</b>	<b>0.0000</b>	<b>0.0573</b>	<b>1</b>
	<b>8.9075</b>	<b>0.6648</b>	<b>1.5715</b>	

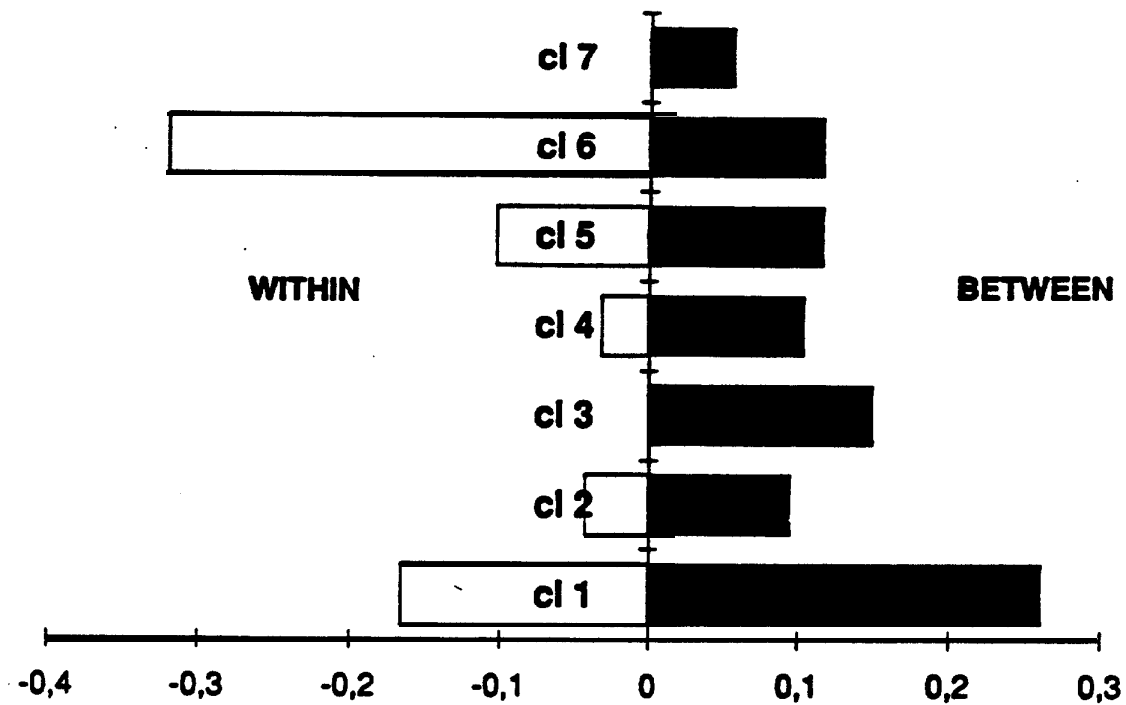


Schéma des inerties par classe

Ce schéma restitue l'information contenue dans le tableau des  $I_b$ ,  $I_w$  et  $I_f$ . Dans le cas où le nombre de classes est peu élevé, il synthétise quasiment à lui seul les différents indicateurs d'analyse par classe. En balayage horizontal il permet de mesurer la cohérence et la distinction absolue de chaque classe alors qu'un balayage vertical il fournit les parts relatives des classes les unes par rapport aux autres.

Lorsque la partition atteint un nombre important de classes, il est toutefois indispensable de construire les indicateurs pour faciliter l'analyse.

## 1. indicateurs absolus

	IBk / ITk	IWk / ITk
cl 1	0,6117	0,3885
cl 2	0,6920	0,3079
cl 3	1	0
cl 4	0,7672	0,2327
cl 5	0,5335	0,4659
cl 6	0,2686	0,7313
cl 7	1	0

Pour une classe donnée, on mesure la répartition entre son inertie between et son inertie within, autrement dit, on évalue sa cohésion et sa distinction. Si l'on excepte les 2 classes à 1 élément, qui par construction, atteignent les valeurs maximales, on constate que ce sont les classes 2 et 4 qui ont les meilleurs indicateurs. La classe 6, qui est la plus importante en taille, est aussi la moins cohérente.

## 2. Indicateurs relatifs

	IBk / IB	IWk / IW	ITk / IT
cl 1	0,2883	0.2583	0,2722
cl 2	0,1062	0,0646	0,0886
cl 3	0,1668	0	0,0963
cl 4	8.1163	0,0482	0,0875
cl 5	0,1296	0,1547	0.1402
cl 6	0,1296	0,4822	0,2786
cl 7	0,0631	0	0,0365

On évalue ici la part de chacune des classes dans la répartition de l'inertie totale. Si ce sont naturellement les plus grosses classes qui contribuent le plus à l'inertie totale, la répartition between / within, est en revanche plus subtile. Ainsi, pour une contribution à l'inertie totale identique, les classes 1 et 6 ont des parts très sensiblement **différentes** dans l'inertie between et l'inertie within de la partition.

A ce stade de l'analyse, il est important de mettre en évidence les phénomènes d'attraction qui lient les classes les unes aux autres.

Tableau des  $I_{jk}^k$

Inerties Between entre classes

	cl 1	cl 2	cl 3	cl 4	cl 5	cl 6	cl 7
cl 1	0.0000	0.1242	0.1575	0.1761	0.2355	0.3235	0.1075
cl 2	0.1242	0.0000	0.1094	0.0819	0.0787	0.1408	0.0998
cl 3	0.1575	0.1094	0.0000	0.1433	0.1371	0.1760	0.1136
cl 4	0.1761	0.0819	0.1433	0.0000	0.1235	0.1215	0.0791
cl 5	0.2355	0.0787	0.1371	0.1235	0.0000	0.1479	0.0835
cl 6	0.3235	0.1408	0.1760	0.1215	0.1479	0.0000	0.0534
cl 7	0.1075	0.0998	0.1136	0.0791	0.0835	0.0534	0.0000

Si ces valeurs brutes sont déjà informatives, il est toutefois utile de les normaliser pour faire apparaître plus clairement les parts respectives de chacune des liaisons au sein d'une classe d'abord, puis au niveau global ensuite.

	cl 1	cl 2	cl 3	cl 4	cl 5	cl 6	cl 7
cl 1	0,0000	0,1105	0,1401	0,1566	0,2095	0,2877	0,0956
cl 2	0,1957	0,0000	0,1723	0,1290	0,1240	0,2218	0,1572
cl 3	0,2481	0,1723	0,0000	0,1712	0,1638	0,2103	0,1357
cl 4	0,2428	0,1129	0,1975	0,0000	0,1703	0,1675	0,1090
cl 5	0,2921	0,0976	0,1701	0,1532	0,0000	0,1835	0,1036
cl 6	0,3359	0,1462	0,1827	0,1262	0,1536	0,0000	0,0554
cl 7	0,2002	0,1859	0,2116	0,1473	0,1555	0,0995	0,0000

Ce tableau est à voir comme le tableau des profils de distinction ou d'éloignement entre les classes. Il n'est naturellement pas symétrique puisque par construction il est relatif à chaque classe.

Ainsi pour chacune d'entre elles, on peut évaluer les classes avec lesquelles elle entretient les répulsions les plus fortes ou, au contraire, celles vers lesquelles elle est attirée.

Par exemple, la classe 1 s'oppose de façon très nette aux classes 5 et 6. Ce phénomène est clairement restitué sur le premier plan factoriel alors qu'il s'estompe progressivement sur les deux suivants. Proportionnellement, cette même classe 1 est peu opposée à la classe 7. Il faut aller jusqu'à l'examen du plan formé des axes 2 et 3 pour visualiser cette proximité qui n'apparaît pas sur les 2 premiers plans. Cette remarque est confirmée par l'examen du

**profil** de liaison de la classe 7, celle qu'elle entretient avec la classe 1 est relativement importante.

	cl 1	cl 2	cl 3	cl 4	cl 5	cl 6	cl 7
cl 1	0,0000	0,0137	0,0145	0,0194	0,0346	0,1188	0,0099
cl 2	0,0137	0,0000	0,0060	0,0060	0,0087	0,0465	0,0055
cl 3	0,0145	0,0060	0,0000	0,0079	0,0126	0,0549	0,0042
cl 4	0,0194	0,0060	0,0079	0,0000	0,0136	0,0402	0,0044
cl 5	0,0346	0,0087	0,0126	0,0136	0,0000	0,0543	0,0077
cl 6	0,1188	0,0465	0,0549	0,0402	0,0543	0,0000	0,0167
cl 7	0,0099	0,0055	0,0042	0,0044	0,0077	0,0167	0,0000

On s'intéresse cette-fois à la part de chaque "liaison" inter classe vis-à-vis de l'inertie between globale, autrement dit, on décompose l'inertie between globale non plus par classe simple mais par paire de classe. On peut ainsi mesurer les éloignements ou les rapprochements les plus significatifs,

Il apparaît clairement que la distinction entre la classe 6 et la classe 1 prend une part très importante dans le découpage du nuage. C'est d'ailleurs bien cette opposition que caractérise le premier axe factorielle et dans une moindre mesure, l'opposition de cette classe 6 avec toutes les autres classe, à l'exception de la classe 7. Vient ensuite la distinction entre la classe 1 et la classe 5 qui atteint une valeur relativement importante. L'axe factoriel 2 restitue cette opposition entre la classe 1 et les autres classes de la partition à l'exception toujours de la classe 7. Cette classe 7 d'ailleurs ne se démarque jamais **véritablement**. On pouvait déjà le constater sur le tableau précédent **où** son profil de liaison absolu n'était pas véritablement typé.

### 8.3.4 Indicateurs par individu

Après l'examen du résultat par classe, on passe à la dernière étape de l'analyse qui nous **amène** aux individus eux-mêmes.

### Contribution des individus aux classes

		C+(i)	C-(i)	CI(i)	CI(i)/CI(k)
<b>Classe 1</b>					
1	IION	0.1130	0.0431	0.0699	.3055
2	TIGRE	0.1176	0.0423	0.0752	.3287
3	JAGUAR	0.0887	0.0384	0.0503	.2198
4	LEOPARD	0.0757	0.0423	0.0334	.1460
<b>Classe 2</b>					
5	ONCE	0.0866	0.0215	0.0651	.5420
7	PUMA	0.0765	0.0215	0.0550	.4580
<b>Classe 3</b>					
6	GUEPARD	0.1847	0.0000	0.1847	1.
<b>Classe 4</b>					
8	NEBUL	0.0925	0.0160	0.0765	.5460
27	MARBRE	0.0796	0.0160	0.0636	.4540
<b>Classe 5</b>					
9	SERVAL	0.0699	0.0245	0.0454	.3063
10	ORSTOT	0.0502	0.0260	0.0242	.1633
11	LYNX	0.0663	0.0286	0.0377	.2544
12	CARACAL	0.0646	0.0237	0.0409	.2760
<b>Classe 6</b>					
13	VIVERRIN	0.0378	0.0254	0.0124	.0375
14	YAGUARUN	0.0344	0.0290	0.0054	.0163
15	CHAUS	0.0401	0.0272	0.0129	.0390
16	DORE	0.0427	0.0136	0.0291	.0880
17	MERGUAY	0.0416	0.0143	0.0273	.0825
18	MARGERIT	0.0425	0.0240	0.0185	.0559
20	CHINE	0.0406	0.0245	0.0161	.0487
21	BENGALE	0.0416	0.0143	0.0273	.0825
22	ROUILLEU	0.0410	0.0157	0.0253	.0765
23	MALAIS	0.0429	0.0192	0.0237	.0716
24	BORNEO	0.0427	0.0136	0.0291	.0880
25	NIGRIPES	0.0382	0.0253	0.0128	.0387
26	MANUL	0.0429	0.0192	0.0237	.0716
28	TIGRIN	0.0416	0.0143	0.0273	.0825
29	TEMMINCK	0.0427	0.0136	0.0291	.0880
30	ANDES	0.0377	0.0268	0.0109	.0330
<b>Classe 7</b>					
19	CAFER	0.0906	0.0000	0.0906	1.

Les contributions des individus aux classes permettant, lorsqu'on étudie la composition de celles-ci, de déterminer les individus qui s'intègrent le mieux dans leur classe ou ceux qui jouent un rôle prédominant dans leur séparation.

Une contribution interne faible caractérise un individu bien positionné dans sa classe, et une contribution externe faible caractérise un individu bien distinct du nuage des classes.

Parmi les exemples significatifs, notons dans la classe **1** le cas du tigre et du léopard. Si leurs contributions internes sont identiques, ils se distinguent très nettement par leur contribution externes. La place du léopard est en effet moins affirmée que celle du tigre au sein de la classe. Cette remarque est confirmée par l'examen des 2 premiers plans factoriels **où** l'on voit le léopard "**s'approcher**" d'autres classes, alors que le tigre, sur les 3 plans, est toujours bien distinct.

#### **8.4 Conclusion**

---

La contribution des approches relationnelles et factorialistes s'avère être un outil très performant d'analyse des données. Elle apporte à l'analyse relationnelle une dimension graphique très utile et **parallèlement** dote l'**AFCM** d'un potentiel **classificateur** tout aussi important.

D'autant, rappelons-le, que les deux approches **opèrent** sur les mêmes quantités de base, traitent les mêmes tableaux et optimisent de la même façon un **critère** inertiel. On est donc garanti de la validité de la superposition des deux analyses.

Naturellement, l'analyse factorielle-relationnelle ne se substitue pas aux experts des domaines d'applications qui devront interpréter les résultats mais elle fournit une batterie d'outils de mesure et surtout une visualisation qui peuvent faciliter grandement son travail.

## **8.5 Annexes**

**Les annexes comportent:**

- 1. le tableau des données**
- 2. le tableau de correspondance entre le codage et les modalités**
- o ■ **le tableau des valeurs propres**
- 4. le tableau des coordonnées des individus sur les trois premiers axes factoriels ainsi que leur classe d'appartenance**

Tableau 2. Tableau des données														
FELIDE	TYPPEL	LONGPOIL	GRIFFES	COMPORT	OREILLES	LARYNX	TAILLE	POIDS	LON- GUEUR	QUEUE	DENTS	TYPPROIE	ARBRES	CHASSE
LION	1	1	2	1	1	2	3	3	3	2	2	1	1	2
TIGRE	3	1	2	3	1	2	3	3	3	2	2	1	1	1
JAGUAR	2	1	2	2	1	2	3	3	2	1	2	1	2	1
LEOPARD	2	1	2	3	1	2	3	3	2	2	2	2	2	1
ONCE	2	2	2	1	1	2	2	2	2	3	2	2	2	1
GUEPARD	2	1	1	1	1	1	3	2	2	3	1	2	1	2
PUMA	1	1	2	2	1	1	2	3	2	3	2	2	2	1
NEBUL	4	1	2	3	1	2	2	2	2	3	2	3	2	1
SERVAL	2	1	2	1	2	1	2	2	2	1	1	3	2	2
OCELOT	2	1	2	2	1	1	2	2	2	2	1	3	2	1
LYNX	2	2	2	2	2	1	2	2	2	1	2	2	2	1
CARACAL	1	1	2	2	2	1	2	2	1	1	1	3	2	2
VIVERRIN	2	1	2	2	1	1	1	1	2	2	1	3	1	1
YAGUARDUN	1	1	2	2	1	1	1	2	2	3	1	3	2	1
CHAUS	1	2	2	3	2	1	1	2	1	2	1	3	2	1
DORE	1	1	2	3	1	1	1	1	1	2	1	3	2	1
MERGUAY	2	1	2	3	1	1	1	1	1	2	1	3	2	1
MARGERIT	1	2	2	2	1	1	1	1	1	2	1	3	1	1
CAFER	3	1	2	3	1	1	1	1	1	2	1	3	2	2
CHINE	1	1	2	2	2	1	1	1	1	1	1	3	2	1
BENGAL	2	1	2	3	1	1	1	1	1	2	1	3	2	1
ROUILLEU	2	1	2	2	1	1	1	1	1	2	1	3	2	1
MALAIS	1	2	2	3	1	1	1	1	1	1	1	3	2	1
BORNEO	1	1	2	3	1	1	1	1	1	2	1	3	2	1
NIGRIPES	2	1	2	2	1	1	1	1	1	1	1	3	2	2
MANUL	1	2	2	3	1	1	1	1	1	1	1	3	2	1
MARBRE	4	1	2	3	1	1	1	1	1	3	1	3	2	1
TIGRIN	2	1	2	3	1	1	1	1	1	2	1	3	2	1
TEMMINCK	1	1	2	3	1	1	1	1	1	2	1	3	2	1
ANDES	2	2	2	3	1	1	1	1	2	2	1	2	2	1

**Tableau 3. Tableau de correspondance des modalités**

Variable	modalité 1	modalité 2	modalité 3	modalité 4
TYPPEL	UNI	TACHETE	RAYE	MARBRE
LONGPOIL	RAS	LONG		
GRIFFES	NONRETR	RETRAC		
COMPORT	DIURNE	DIU-NOCT	NOCTURNE	
OREILLES	RONDE	POINTUE		
LARYNX	SANS-OS	AVEC-OS		
TAILLE	TPETITE	TMOYENNE	TGRANDE	
POIDS	FAIBLE	MOYEN	FORT	
LONGUEUR	LPETITE	LMOYENNE	LGRANDE	
QUEUE	LONGUE	MOYENNE	COURTE	
DENTS	CAN-NDEV	CAN-DEV		
TYPPROIE	GROSSE	GRO-PET	PETITE	
ARBRES	NGRIMPE	GRIMPE		
CHASSE	NCHASSE	CHASSE		

**(3) Tableau des valeurs propres**

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE
1	.3881	24.70	24.70
2	.2607	16.59	41.29
3	.1625	10.34	51.63
4	.1503	9.56	61.19
5	.1133	7.21	68.40
6	.0921	5.86	74.26
7	.0788	5.01	79.27
8	.0649	4.13	83.40
9	.0540	3.44	86.84
10	.0494	3.14	89.98
11	.0418	2.66	92.64
12	.0285	1.81	94.45
13	.0191	1.22	95.67
14	.0172	1.09	96.77
15	.0158	1.01	97.77
16	.0135	.86	98.63
17	.0078	.50	99.13
18	.0060	.38	99.50
19	.0041	.26	99.77
20	.0024	.15	99.92
21	.0011	.07	100.00
22	.0001	.00	100.00

(4) Tableau des coordonnées des individus sur les trois premiers axes factoriels

FELIDE	Axe1	Axe 2	Axe 3	Classe
LION	-1.57625	-0.79979	0.55625	1
TIGRE	-1.47563	-1.25592	-0.02743	<b>1</b>
JAGUAR	-1.02703	-0.17279	-0.29105	1
LEOPARD	-0.85582	-0.09915	-0.54928	1
ONCE	-0.72419	0.91925	-0.43898	2
PUMA	-0.49238	0.49130	-0.51876	2
GUEPARD	-0.87123	0.72413	1.14382	3
NEBUL	-0.41787	0.51526	-0.91210	4
MARBRE	0.39869	-0.08854	-0.49003	4
SERVAL	-0.10762	0.86326	0.67665	5
OCELOT	0.06231	0.43144	-0.08279	5
LYNX	-0.20547	1.00982	-0.16156	<b>5</b>
CARACAL	0.25265	0.56266	0.56881	5
VIVERRIN	0.13666	-0.15039	0.14929	6
YAGUARUN	0.16310	0.33906	-0.10765	6
<b>CHAUS</b>	0.48844	0.08362	0.06336	6
DORE	0.50827	-0.38766	-0.05999	6
MERGUAY	0.45550	-0.31512	-0.07477	6
<b>MARGERIT</b>	0.37892	-0.31110	0.17867	6
CHINE	0.53959	0.05491	0.25229	6
BENGALE	0.45550	-0.31512	-0.07477	6
ROUILLEU	0.43336	-0.19495	0.01152	6
MALAIS	0.56230	-0.16109	-0.04031	6
BORNEO	0.50827	-0.38766	-0.05999	6
NIGRIPES	0.38200	0.00206	0.42744	6
<b>MANUL</b>	0.56230	-0.16109	-0.04031	6
TIGRIN	0.45550	-0.31512	-0.07477	6
TEMMINCK	0.50827	-0.38766	-0.05999	6
ANDES	0.20060	0.07584	-0.27726	6
CAFER	0.30126	-0.56855	0.31366	7

## 9.0 Bibliographie

---

- [Ayou90] N ● El Ayoubi.  
*Liaison Analyse Factorielle-Analyse Relationnelle: Extensions.*  
PhD thesis, Thèse de l'Université Paris VI, 1990.
- [Barr91] R. Barré and coll.  
Science et Technologie Indicateurs 1992.  
Economica, p<sup>gds</sup> ● Octobre 1991.
- [Bede89a] ε ● Bedecarrax.  
*Classification Automatique en Analyse Relationnelle: la Quadri-Décomposition et ses applications* ●  
PhD thesis, Thèse de l'Université Paris VI, Décembre 1989.
- [Bede91] ] C. Bedecarrax and C. Huot.  
Application de l'Analyse Relationnelle à la Veille Technologique: des outils d'analyse de l'information documentaire.  
*Revue Française de Bibliométrie*, pages 66-80, Congrès S.F.B.A., Ile Rousse, 1991.
- [Bede89b] C. Bedecarrax and I. Warnesson.  
Relational Analysis and εγλομνλγs ●  
*Applied Stochastic Models and Data Analysis*, 5:131-151, 1989.
- [Benz80] J.P. Benzécri .  
*L'Analyse des Données: la Taxinomie (Tome I)* ●  
Dunod, Paris, 1980.
- [Broc92] K.K. Brockhoff.  
Instruments for Patent Data Analysis in Business Firms.  
*Technovation*, 12(1):41-59, Elsevier Science Publishers Ltd, February 1992.
- [Broo84] B.C. Brookes .  
Ranking techniques and the Empirical Log Law.  
*Information Processing and Management*, (20):37-46, 1984.
- [Cele89] G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy.  
*Classification automatique des données.*  
Dunod, Paris, 1989.

- [Chah86] S. Chah.  
Nouvelles *techniques de codage d'association et de classification*.  
PhD thesis, Thèse de Doctorat d'état Université Paris VI, 1986.
- [Chal90] M. Chalet and T. Wolton.  
*Les visiteurs de l'ombre*.  
Grasset, Paris, 1990.
- [Cour90] J.P. Courtial.  
*Introduction à la Scientomktrie*.  
Anthropos, Paris, 1990.
- [Dehe90] P. Deheuvels,  
*La recherche scientifique*.  
in Collection "Que-sais-je". PUF, Paris, 1990.
- [Doré86] J.C. Doré, J. Gilbert, J.F. **Miquel**, A. Deroulede, and C. Dutheuil,  
Banques de Données et Analyses Multivariables.  
*Actes*, Congrès CNIC, 1986.
- [Dou92] H. Dou and all.  
*La Veille Technologique*.  
Dunod, Paris, 1992.
- [Dou83] H. Dou and P. Hassanaly.  
How to use **online** databases as a tool for forecasting fundamental and applied research.  
*Learned Information*, pages 175- 183, 7th International **Online** Information Meeting (London), Oxford, 1983,
- [Dou89] H. DOU, P. Hassanaly, and L. Quoniam.  
Easy Mapping **Classification** of Patent **References** with Microcomputers.  
*Proceedings, The* Montreux International Chemical Information **Conference**, 1989.
- [Dou90a] H. **Dou**, P. Hassanaly, L. Quoniam, and M. Gilbert.  
Analyse des Données et Traitement Automatique de l'Information.  
*Analisis, 1990*.
- [Dou90b] H. Dou, P. Hassanaly, L. Quoniam, and A. La **Tela**.  
Post processing of **online** search.  
*Bladen voor de documentatie, 3:51-70*, Bruxelles, 1990.

- [Dou90c] H. Dou, P. Hassanaly, L. Quoniam, and A. La Tela.  
Veille technologique et information documentaire.  
*Documentaliste*, 3(27):132-141, Paris, mai-juin 1990.
- [Dou90d] H. Dou, L. Quoniam, H. Rostaing, and W. Nivol.  
L'analyse des données au service de la bibliométrie: Outils de Veille Technologique à la dimension des moyennes en (rapports)  
*Revue française de Bibliométrie*, (8):27-67, Paris, décembre 1990.
- [Dous87] B. Dousset and T. Benjamaa.  
Une approche interactive de la manipulation, l'analyse et la classification de données bibliométriques par représentation tridimensionnelle sur micro-ordinateur graphique couleur.  
*Les systèmes d'informations élaborées*, pages 79-86, Congrès S.F.B.A, Ile Rousse, 1987.
- [Duth87] C. Dutheuil.  
Analyse factorielle et classification automatique appliquées aux données bibliométriques  
*Colloque CNIC*, pages 20, Paris, 1987.
- [Duth90] C. Dutheuil.  
Du Corpus Documentaire à l'Interprétation des Résultats de l'Analyse des Données Bibliométriques  
*Revue Française de Bibliométrie Appliquée*, (6):90-104, Février 1990.
- [Duth91] C. Dutheuil.  
L'état de l'art de la bibliométrie et de la scientométrie en France et à l'étranger.  
SGDN, Décembre 1991.
- [Eggh89] L. Egghe.  
The duality of informetric systems with applications to the empirical laws.  
*Journal of Information Science*, (16):17-27, Elsevier, 1989.
- [Esco88] B. Escoffier and J. Pages.  
*Analyses factorielles simples et multiples*, pages 241-250  
Dunod, Paris, 1988.
- [Ghas90] S. Ghashghaie.  
*Analyse des données ordinales: généralisation des critères d'association.*  
PhD thesis, Thèse de l' Université Paris VI, Juin 1990.

- [Gira88] A. Girard and M. Moureau.  
Etude statistique des brevets: un nouvel outil d'aide à la décision.  
*Revue de l'Institut Français du Pétrole*, 43(1), Jan-Fev 1988.
- [Jako88] F. Jakobiak.  
*Maîtriser l'Information Critique*.  
Editions d'Organisation, Paris, 1988.
- [Jako89] F. Jakobiak.  
Rapport Veille Technologique du Xième Plan Commission "Europe Technologique Industrielle et Commerciale, Innovation et Recherche".  
Groupe " Veille Technologique et Politique de Propriété Industrielle", 1989.
- [Jako90] F. Jakobiak.  
*Pratique de la Veille Technologique*.  
Les Editions d'Organisation, Paris, 1990.
- [Jako92] F. Jakobiak.  
*Exemples commentés de veille technologique*.  
Les Editions d'Organisation, Paris, 1992.
- [Laf90] T. Lafouge.  
Une autre approche de la circulation de l'information.,  
*Revue Française de Bibliométrie*, (7): I-13, Paris, 1990.
- [Lasf89] Y. Lasfargue.  
*Technojolies, technofolies*.  
Editions d'Organisation, Paris, 1989.
- [Leyd87] L. Leydesdorff and R. Zaal.  
Co-word and citations relations between document sets and environments.,  
*Proceedings*, pages 31, aspects of information retrieval, Luc (Belgium), 1987.
- [Marc84] F. Marcotorchino.  
Utilisation des comparaisons par paires en statistique des contingences, Partie II.  
*Etude du Centre Scientifique, F069*, IBM France, 1984.
- [Marc85] F. Marcotorchino.  
Utilisation des comparaisons par paires en statistique des contingences, Partie III.  
*Etude du Centre Scientifique, F081*, IBM France, 1985.

- [Marc87] F. Marcotorchino.  
 $\wedge$  Unified Approach of the Block-Serialisation Problems.  
*Applied Stochastic Models and Data Analysis*, 3(2), J. Wiley, New York, 1987.
- [Marc89] F. Marcotorchino.  
 Liaison Analyse Factorielle-Analyse Relationnelle (I) "Dualité Burt-Condorcet".  
*Etude du Centre Scientifique*, F142, IBM France, 1989.
- [Marc91a] F. Marcotorchino.  
 L'Analyse Factorielle-Relationnelle: Parties I et II.  
*Etude du CEMAP*, MAP-03, IBM France, Décembre 1991.
- [Marc91b] F. Marcotorchino.  
 La classification automatique aujourd'hui: bref aperçu historique et calculatoire.  
*Publications Scientifiques et Techniques*, (2):35-93, IBM France, Novembre 1991.
- [Marc78] F. Marcotorchino and P. Michaud.  
*Optimisation en Analyse Ordinale des Données*.  
 Masson, Paris, 1978.
- [Marc79] F. Marcotorchino and P. Michaud.  
 Modèles d'Optimisation en Analyse des Données Relationnelles.  
*Math. et Sciences Humaines*, (67), Gauthier Villars, Paris, 1979.
- [Marc81] F. Marcotorchino and P. Michaud.  
 Agrégation des Similarités en Classification Automatique.  
*Revue de Statistique Appliqués*, 30(2), Dunod, 1981.
- [Mart89] B. Martinet and J.M. Ribault.  
*La Veille Technologique Concurrentielle et Commerciale*.  
 Editions d'Organisation, Paris, 1989.
- [Mess90] H. Messatfa.  
*Unification Relationnelle des critères et structures de contingences*.  
 PhD thesis, Thèse de l'Université Paris VI, Mai 1990.
- [Meye90] H. Meyer.  
*L'Information*.  
 Rivages/Les Echos, Paris, 1990.

- [Mich82] P. Michaud.  
Agrégation à la Majorité 1: Hommage à Condorcet.  
*Etude du Centre Scientifique, F05 1*, IBM France, 1982.
- [Mich85] P. Michaud.  
Agrégation à la Majorité II: Analyse du Résultat d'un Vote.  
*Etude du Centre Scientifique, F052*, IBM France, 1985.
- [Mour90] M. Moureau and A. Girard.  
L'utilisation des données bibliométrique à l'institut français du pétrole.  
*Revue Française de Bibliométrie, (7)*, Juin 1990.
- [Paol87] C. Paoli, P. Billard, P. **Blanchet**, and C. Longuevialle.  
Apport de l'analyse factorielle et de la **classification** ascendante hiérarchique dans l'analyse des banques de données bibliographique.  
*Acres*, pages **65-75**, Congrès S.F.B.A: Les systèmes d'informations élaborées, Ile Rousse, 1987.
- [Quon88] L. Quoniam.  
*Bibliométrie Informatisée et Information Stratégique*.  
PhD thesis, Thèse de l'**Université** de Droit et des Sciences d'Aix-Marseille, 1988.
- [Quon90a] L. Quoniam, H. Dou, P. Hassanaly, and M. Gilbert.  
Analyse des Données et Traitement Automatique de l'Information.  
*Analisis, 1990*.
- [Quon90b] L. Quoniam, H. **Dou**, and C. Huot.  
Les Méthodes **d'Analyse** des Données face à l'Information Stratégique et l'innovation.  
*Acres, ENSAIS*, Strasbourg, 1990.
- [Reyn90] M. Reyne.  
*Le Développement de l'entreprise par la veille technologique*.  
Hermès, Paris, 1990.
- [Rip88] A. Rip.  
Mapping of Science, **Possibilities** and Limitation.  
*Handbook of Quantitative Studies of Science and Technology*, pages 253-273, North-Holland, Amsterdam, 1988.

- [Rock79] J.F. Rockart.  
Chief Executive Define their Own Data Needs.  
*Harvard Business Review*, March-April 1979.
- [Rous90] R. Rousseau.  
Classification Ascendant Hiérarchique et décomposition en blocs diagonaux.  
*Colloque International de Sériation par Bloc*, Strasbourg, 1990.
- [Tela87] A. La Tela.  
*Systèmes Interactifs d'aide à la décision (S.I.A.D)*, pages 96.  
PhD thesis, Thèse de l' Université de Marseille III, 1987.
- [Tijs88] R.J.W. Tijssen, J. De Leew, and A.F.J Van Raan.  
A method for mapping bibliometric relations based on field-classifications and citations of articles.  
in L. Egghe R. Rousseau, editor, *informetrics 87/88*, Elsevier Science Publishers, Amsterdam, 1988.
- [Todo90] R. Todorov and M. Winterhager.  
Mapping Australian geophysics: a co-heading analysis.  
*Scientometrics*, 19(1-2):35-56, 1990.
- [Turn89] W.A. Turner.  
De la bibliométrie à l'infométrie: des axes de recherche nouveaux pour la veille scientifique et technologique.  
*Actes*, pages 161-179, Congrès S.F.B.A: Les systèmes d'informations élaborées, Mai-Juin 1989.
- [Van 89] A.F.J Van Raan and H.P.F. Peters.  
Dynamics of a scientific field analysed by co-subfield structures.  
*Scientometrics*, 15(5-6):607-620, 1989.
- [Vill90] J. Villain.  
*L'entreprise aux aguets*.  
Masson, Paris, 1990.
- [Warn90] I. Warnesson.  
Traitement des données lexicographiques basé sur une approche relationnelle des problèmes de sériation par blocs.  
*Actes*, ENSAIS, Strasbourg, 1990.

[Warn91] I. Wamesson.

Structuration du lexique et documentation automatique.

Actes, pages 324-351, SFBA, Ile Rousse, 1991.

[Whit89] H.D. White and K.W. Mc Cain.

*Bibliometrics*, pages 119- 186.

in Williams M.E. Ed., *Annual Review of Information Science and Technology*.

Elsevier, 1989.