

Le logiciel bibliométrique DATAVIEW et son application comme outil d'aide à l'évaluation de la concurrence

Rostaing Hervé*, Nivol William, Quoniam Luc, Latela Albert

*Centre de Recherches Rétrospectives de Marseille (CRRM)
Centre universitaire de Saint Jérôme
Avenue Escadrille Normandie Niemen, 13397 Marseille*

Pour assurer son développement dans un environnement concurrentiel de plus en plus menaçant, l'entreprise doit être, en permanence, informée sur son environnement extérieur.

Pour réussir cette démarche, il lui faut intégrer de plus en plus d'informations : scientifiques, techniques, technologiques, économiques, et les données dites complémentaires : juridiques, normatives, politiques etc.. Ces données évoluant sans cesse, il est primordial d'en évaluer les tendances, d'en déceler les indices de changement, d'essayer d'en deviner les synergies possibles afin de pouvoir anticiper et d'être toujours prêt à innover.

Pour cela, il est indispensable que l'entreprise développe une activité de Veille Technologique qui lui permette d'élaborer des indicateurs qui aideront les décideurs à diagnostiquer l'état des activités sur le plan concurrentiel (scientifique, technique, technologique...).

Nous présenterons, au cours de cette communication, la conception et l'utilisation d'un logiciel de traitements bibliométriques : "Dataview". Au cours du processus de veille technologique, cet outil va servir à l'élaboration des précédents indicateurs à partir d'une exploitation systématique et automatique de l'information issue des bases de données.

Nous exposerons, tout d'abord, les concepts sur lesquels reposent Dataview ainsi que les différents types de traitements qu'il permet de générer.

Dans un second temps, nous décrirons à partir de cas concrets d'analyses, comment cet outil peut aider à la construction d'informations élaborées en entreprise. Les traitements présentés ont été réalisés à partir d'information brevet téléchargé de la base de données WPIL produite par la société Derwent.

I. Dataview : concepts et développement

Un outil informatique bibliométrique ouvert

*

Membre du CRRM. Actuellement scientifique du contingent détaché au CEDOCAR (Centre de Documentation des Armées, 2 bis rue Bossoutrot 0460 Armée)

Lorsqu'on étudie les travaux menés dans le domaine de la bibliométrie, on constate que peu d'auteurs parlent de **l'automatisation des manipulations de données** qu'ils effectuent. Les interventions de logiciels informatiques se limitent essentiellement à deux étapes dans une étude bibliométrique :

- ➔ **au moment de la collecte**, par l'utilisation de logiciels de télécommunication et de systèmes d'interrogation des bases de données.
- ➔ **au moment de l'analyse statistique des données**, par l'exploitation d'outils informatiques et mathématiques pour représenter ces données sous une forme graphique : courbes, histogrammes, secteurs, réseaux, nuages de points, arborescences, regroupements, etc..

On peut remarquer que les données traitées par ces deux étapes ne sont pas du même type. En effet, au moment de la collecte, les données sont textuelles (notices bibliographiques) tandis qu'à la seconde étape elles sont numériques.

Pour que la chaîne informatique des traitements bibliométriques soit complète, il manque donc de logiciels offrant la liaison entre ces deux étapes. Or, cette étape intermédiaire, qui consiste à transformer des données textuelles en données numériques, paraît mieux répondre à l'appellation de *"traitement bibliométrique"* **puisque c'est celle va transformer l'information qualitative en information quantitative.**

Seuls quelques travaux de recherche ont été menés dans cet objectif : Brooks avec son logiciel **"Bibliometrics Toolbox"** [BROOKS87], la société Derwent avec le logiciel **"PatStat+"**⁽¹⁾, la société Battelle avec le logiciel **"Patent Trend Analysis"**⁽²⁾ et la collaboration CSI-CDST avec le logiciel **"Leximappe"** [MICHELET88].

Chacun de ces outils a été développé dans un axe bien précis de traitement bibliométrique. Le *Toolbox de Brooks* permet d'estimer les caractéristiques des lois bibliométriques (Bradford, Lotka, Zipf...). *PatStat+* et *Battelle* permettent uniquement d'analyser des fichiers comportant des références brevets (de Derwent pour *PatStat+*, de Derwent et d'US Patents pour *Battelle*). Toutefois, bien que ces outils aient le mérite d'exister, ils ne permettent pas d'entreprendre de façon fiable et approfondie l'exploitation de l'information brevet. Les raisons de ces restrictions s'expliquent pour l'essentiel par la description des limites liées à l'exploitation statistique des informations brevets issues des bases de données [NIVOL93]. *Leximappe* est conçu pour traiter qu'un champ à la fois, ce champ devant être le représentant du contenu scientifique ou technique des notices bibliographiques sous la forme de mots-clés (champ titre ou champs des descripteurs).

Le CRRM développe aussi depuis de longues années des applications informatiques à finalités bibliométriques. La gamme de ces outils est variée et se partage entre des outils à traitements spécifiques ou à traitements à façon (***Datacode, Dataget, Datage, Datastra, Datalink, Datrans***) [DOU89],[DOU90],[LATELA87],[QUONIAM88]. L'utilisation de cette première génération de logiciels a

(1) Derwent Publications LTD, Rochdale House, Theodablds Road, London WC1 X3RP, GB

(2) Battelle Europe, 7 route de Drize, CH-1227 Carrouge-Genève, Suisse

permis au CRRM d'acquérir l'expérience des contraintes qu'imposent les données bibliométriques et une connaissance de la variété des traitements utiles en bibliométrie. Le logiciel **Dataview** représente l'aboutissement d'une longue réflexion menée par le CRRM et s'inscrit dans la continuité des précédents outils qui y ont été développés.

Dataview a été conçu spécialement pour la transformation de l'information textuelle en données numériques [ROSTAIN93].

La variété de ces traitements lui octroie un aspect ouvert à tout type configuration d'analyse bibliométrique. Pour aboutir à ce résultat, nous nous sommes focaliser lors de sa conception sur le fait qu'il réponde parfaitement à trois contraintes.

⇒ ***Intégrer la diversité des formats des sources d'information :***

l'outil informatique doit être suffisamment flexible et maniable pour permettre de manipuler des données textuelles sous des formes et des structures diverses.

⇒ ***Offrir la diversité des éléments bibliométriques à manipuler :***

l'outil informatique doit permettre l'étude de toutes les différentes parties des données textuelles.

⇒ ***Permettre la diversité des traitements statistiques bibliométriques :***

L'outil informatique doit fournir en sortie toutes les catégories de données numériques employées en bibliométrie et en statistique.

La maîtrise de ces trois contraintes nous permet de constituer les données bibliométriques de base utiles à tout type d'études et de méthodes bibliométriques.

Concepts introduits lors du développement de Dataview

Pour respecter ces trois contraintes de diversité, lors de la conception de Dataview nous avons choisi d'adopter les solutions informatiques suivantes:

⇒ ***Solution à la diversité des formats des sources d'information :***

La diversité des formats proposés par les serveurs de base de données impose de définir quelques repères pour différencier les différentes parties des textes à analyser.

Dans le cas de signalement bibliographique on peut différencier trois parties dans le texte:

- les marques de séparations entre chaque notice
- les intitulés des champs
- les contenus des champs

Ces trois repères constituent la structure des références bibliographiques. Comme cette structure varie d'un serveur à l'autre, nous avons choisi de prendre comme repère du format des notices celui qui est le plus souvent rencontré en standard sur les serveurs professionnels.

Donc DATAVIEW reconnaît comme:

- **repère de séparation entre les notices :**
une ligne vide ou l'apparition d'un champ déjà répertorié comme appartenant à la notice précédente
- **repère d'intitulé d'un champ :**
une chaîne de caractère en début de ligne ne dépassant pas 10 caractères
- **repère du contenu d'un champ :**
ensemble des lignes alignés sur l'intitulé du champ (lignes commençant par autant blancs que de caractères dans l'intitulé du champ).

Grâce à ces caractéristiques de repérage, Dataview peut traiter sans aucun pré-traitement de reformatage les notices des données présentes sur les serveurs *Questel, Cedocar, Orbit, Dialog, STN...*

➡ **Solution à la diversité des éléments bibliométriques à manipuler**

Il faut indiquer à Dataview sur quel(s) champ(s) les traitements vont s'effectuer et comment il faut découper le(s) contenu(s) de ce(s) champ(s). En fonction du serveur, de la base et du champ étudié les caractères, qui séparent les éléments bibliographiques à traiter, sont différents.

Nous avons décidé de nommer ces éléments bibliographiques sous l'appellation de "forme graphique". Nous avons repris l'appellation que *Lebart* et *Salem* [LEBART88] donnent aux unités statistiques de comptages dans leurs traitements statistiques de réponses aux questions libres d'enquête. Par commodité, l'appellation s'est réduite à "**forme**". Donc, on nomme *forme* l'élément bibliométrique considéré lors des traitements dans DATAVIEW.

Dataview va constituer les formes à partir de deux renseignements:

- **l'intitulé du (des) champ(s) à traiter :**
on peut traiter jusqu'à dix champs simultanément
- **une liste de séparateurs pour chacun des champs :**
un séparateur est un caractère qui sert de repère de séparation entre chaque forme. Certains champs comportent plusieurs caractères différents pouvant jouer le rôle de séparateurs. Ainsi on peut préciser un ou plusieurs caractères séparateurs à considérer pour chaque champ traité.

Une forme, au sens de Dataview, est donc une suite de caractères, encadrée de part et d'autre par un caractère séparateur-de-forme. L'ensemble des formes représente l'ensemble des entités bibliométriques traitées.

⇒ **Solution à la diversité des traitements statistiques bibliométriques**

La bibliométrie est une discipline qui fait appel à la mesure. Cette mesure est forcément établie sur des données numériques. Toutes les évaluations bibliométriques sont calculées à partir d'une mesure unique : **l'occurrence**. La cooccurrence de deux formes, autre mesure souvent pris en compte lors d'analyse bibliométrique, n'est en fait qu'une combinaison des occurrences des deux éléments. **L'unité de mesure dans les études bibliométriques est donc principalement l'occurrence des formes.**

En fait, pour la plupart de ces formes cette unité équivaut au recensement du nombre de références où elles sont présentes. Ceci est vrai pour les formes qui font partie d'un des **champs contrôlés**. Cette appellation signifie que le contenu du champ a été retraité par une personne: champs des mots-clés et des codes mais aussi pour certaines bases normalisées les champs des auteurs, des sources, des affiliations... Tous ces champs traités par une intervention humaine sont généralement épurés de toute information redondante, si bien que le nombre d'occurrences des formes qu'ils contiennent est réduit à l'unité dans chaque référence.

Par contre, les autres champs dont le contenu est rempli par un **langage libre** (titres, résumés), ont souvent des redondances de formes. Le nombre d'occurrences dans une référence varie pour chaque forme.

Pour DATAVIEW l'unité bibliométrique choisie pour le comptage des apparitions des formes est la référence. Ne connaissant pas le biais que pourrait introduire la prise en compte, trop systématique, de toutes les occurrences d'une forme, nous avons préféré uniquement comptabiliser la présence ou l'absence de la forme dans la référence, **c'est-à-dire faire l'inventaire des références où le concept qu'elle symbolise est présent.**

⇒ **La notion de fréquence d'une forme et la notion de l'occurrence d'une forme:**

DATAVIEW est tout de même conçu pour fournir ces deux types de dénombrements pour chaque forme: le nombre des références comportant la forme et le nombre d'occurrences de cette même forme.

Pour différencier ces deux comptages nous leur avons donné deux noms:

- **fréquence** : abréviation de ***fréquence de la forme***, c'est-à-dire selon notre unité le nombre de références
- **occurrence** : abréviation de ***fréquence des occurrences de la forme***, c'est-à-dire le nombre de fois que la forme apparaît dans l'ensemble des références

⇒ La notion de paire:

Comme pour le comptage des formes, la notion de cooccurrence va se traduire ainsi, lors des traitements par Dataview:

DATAVIEW comptabilise le nombre de références comportant la co-apparition de deux formes (deux concepts). Pour différencier cette notion de celle de la cooccurrence nous l'avons nommée la ***paire***. Le terme de *paire* est l'abréviation de ***paire de formes***. **La fréquence d'une paire comptabilise donc le nombre de références où les deux formes de la paire sont présents, c'est à dire le nombre de références où les deux concepts qu'elles symbolisent apparaît.**

⇒ La notion d'indice d'association:

Tous les comptages effectués par DATAVIEW se basent sur ces deux notions de fréquence : ***la fréquence de forme*** et ***la fréquence de paire***. Une nouvelle catégorie de mesures pour estimer la force de relation entre deux formes est calculée à partir de ces deux types de fréquence. Ces mesures sont nommées des ***indices d'association***.

La plus simple mesure de relation entre deux formes est le comptage de la fréquence de la paire. Mais elle est en fait biaisée puisqu'elle ne prend pas en compte l'importance des fréquences relatives aux deux formes.

Pour relativiser le poids des formes dans chaque association il est possible d'obtenir par DATAVIEW des indices statistiques de similitude ou de dissimilitude : indices d'associations.

Ces indices se calculent à partir d'un tableau qui résume les données des associations entre deux formes:

		Forme X	
		Présence	Absence
F o r m e Y	Présence	N_A	N_B
	Absence	N_C	N_D

on a

N	= Nombre de co-présences de X et Y
N^A	= Nombre de présences de Y en l'absence de X
N^B	= Nombre de présences de X en l'absence de Y
N^C	= Nombre de co-absences de X et Y
N^D	

plus la relation $N_A + N_B + N_C + N_D = M =$ Nombre de références bibliographiques

Les indices combinent ces quatre valeurs de façon à donner plus ou moins de poids à chacune d'elles dans la mesure de l'association entre X et Y. L'utilisateur dispose d'une liste d'indices statistiques binaires dans laquelle il peut choisir celui qui lui paraît le mieux retracer la mesure d'association entre les formes (voir liste en Annexe).

⇒ Edition des résultats:

Les analyses statistiques ou représentation graphique des données bibliométriques ne sont pas effectuées sous DATAVIEW. Ces outils existant déjà sur le marché avec souvent des traitements difficilement égalables (ex: Excel pour les tableurs et SAS pour les traitements statistiques), il paraît inutile d'y perdre son énergie.

Par contre, DATAVIEW est là pour dégager les caractéristiques bibliométriques d'un corpus de références et les éditer dans un format qui permettent des traitements ultérieurs. Ces édition sont de deux types: les listes de fréquences et les tableaux.

● Les différentes listes de fréquences sont :

- *profils de distribution par rang des fréquences de formes*
- *profils de distribution par rang des fréquences de paires*
- *répartition du nombre de formes par champ*
- *liste des fréquences de formes*
- *liste des fréquences de paires*

● Les différents tableaux sont :

- *matrice de présence-absence*
- *matrice disjonctive complète*
- *matrice symétrique*
- *matrice de contingence*
- *tableau de bord généralisé (tableaux de Burt)*

Ces éditions sont bien évidemment fournies dans un format qui puisse être directement importé dans les applications informatiques exploitées pour les post-traitements statistiques ou infographiques. Dataview propose plusieurs formats d'exportation vers des logiciels commerciaux:

☛ Pour les tableurs :

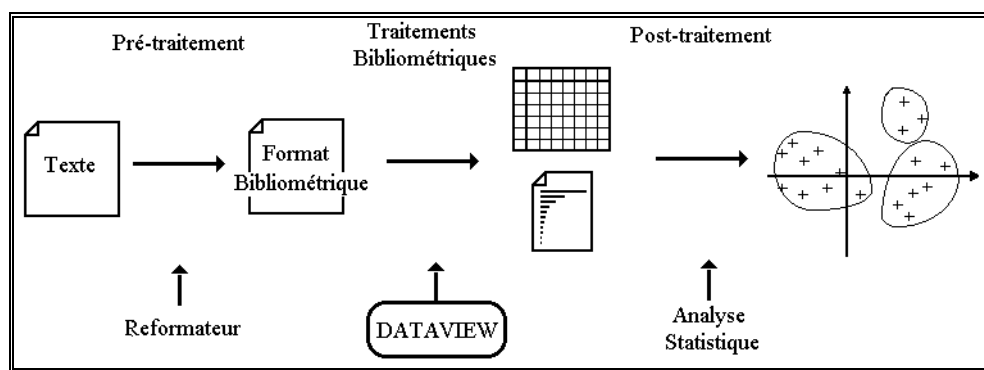
- ☐ Excel
- ☐ Lotus 123
- ☐ Multiplan

☛ Pour les logiciels d'analyse statistique :

- ☐ Statitcf
- ☐ Clustan
- ☐ Arcade (logiciel d'Analyse Relationnelle du CEMAP-IBM)
- ☐ Tétralogie

⇒ Dataview : passerelle entre les données textuelles et les données statistiques

On peut préciser la position centrale que joue Dataview dans les traitements bibliométriques grâce au schéma ci-dessous. Il est important d'insister sur le fait que l'ensemble des traitements bibliométriques réalisés par Dataview permet notamment de préparer les exploitations vers l'utilisation d'outils d'analyses statistiques [DOU90b] :



II. Exploitation bibliométrique de l'information brevet à partir de dataview

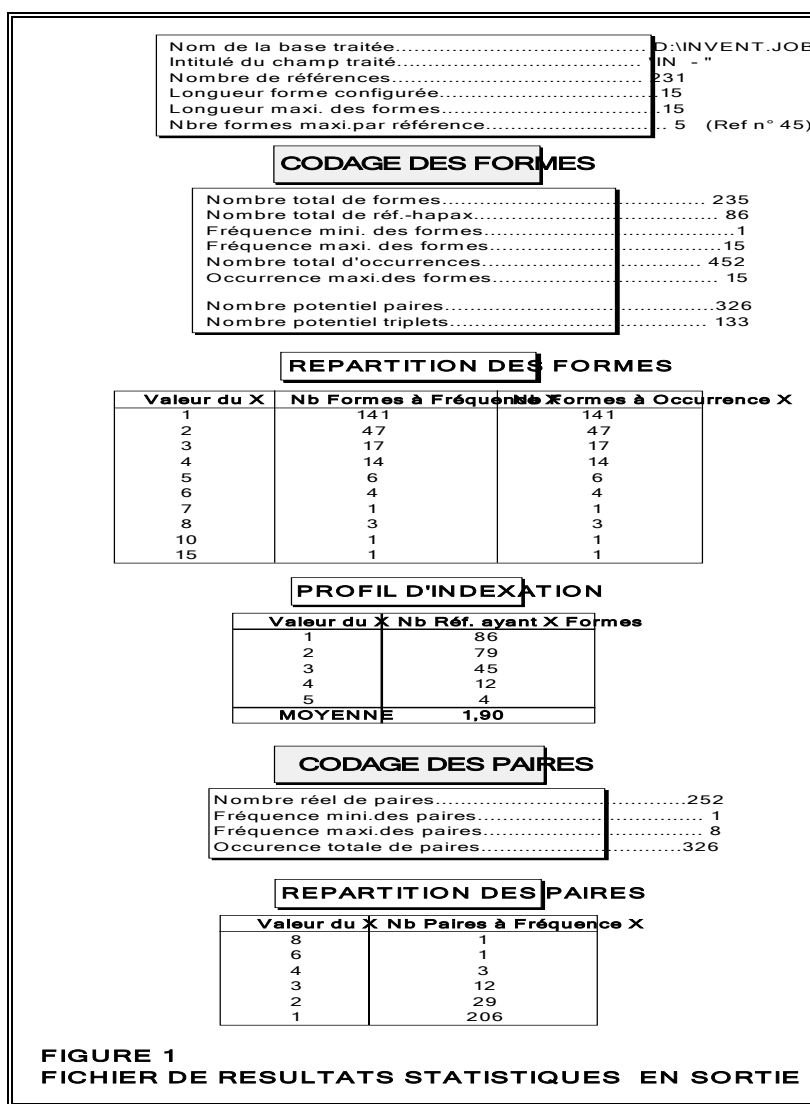
EXPLOITATION DES PROFILS DE DISTRIBUTIONS STATISTIQUES DE FORMES ET DE PAIRES OBTENUS A PARTIR DE DATAVIEW.

Chaque traitement, sous DATAVIEW, génère un fichier de résultats statistiques permettant d'observer les profils de distributions de formes et de paires relatifs au corpus de documents analysés.

Chacune de ces distributions, dans ce fichier, est complétée par un ensemble de statistiques élémentaires. Nous avons traité, sur un échantillon de documents brevets issus d'un corpus téléchargé du fichier WPIL, le champ "IN" (*Inventeur*). Cet échantillon concerne précisément l'ensemble des dépôts enregistré par une société d'envergure internationale, dans le secteur de la cosmétique depuis l'année de priorité 1982 jusqu'à nos jours.

La figure 1 illustre, à partir de cet exemple, le type de résultats statistiques généré en sortie de DATAVIEW. Ce fichier de résultats nous apporte de nombreux renseignements sur la structure des informations analysées.

Ainsi, on constate qu'il existe 235 formes différentes (c'est à dire, dans notre cas, 235 noms d'inventeurs différents) pour 231 documents. D'autres informations relatives aux codages⁽³⁾ des formes et des paires nous renseignent sur leurs fréquences et occurrences extrêmes (minimum et maximum).



■ Signalons que la notion de "*références hapax*" permet de caractériser l'ensemble des références ne comportant **QU'UNE SEULE FORME DANS LE CHAMP D'ANALYSE**. La valeur "86" dans notre exemple signifie qu'il existe 86 références, parmi les 231 traitées, ne comportant qu'un seul nom d'inventeur dans le champ "IN".

(3) Au cours de l'opération de "*codage*", le logiciel établit dans des fichiers, à partir d'un algorithme de H-Coding [LATELA87], l'ensemble des renseignements qui vont permettre d'obtenir, très rapidement, le comptage de la diversité de chacune des formes du fichier traité, ainsi que leurs localisations les unes par rapport aux autres dans l'échantillon analysé. Ces Fichiers, similaires à des index, comportent en fait de nombreuses informations complémentaires renseignant sur les liens existant entre chaque forme recensée. Ces liens établissent la relation existante entre l'ensemble des formes prises deux à deux. ; il s'agit de la notion de "PAIRE".

⇒ A partir des tableaux de répartitions de formes et de paires (figure 1), plusieurs types de profils peuvent être tracés. Les figures 2 à 4 illustrent, pour chacun, un exemple de représentation.

● **La figure 2 représente le profil de collaboration entre inventeurs (co-dépôts par inventeur).**

Elle permet **d'observer la fréquence des collaborations entre chaque individu**. Ainsi, on relève que seul un quart des collaborations entre inventeurs est répétée plus d'une fois. D'autre part, on constate que les fréquences maximales des collaborations entre inventeurs sont relativement faibles ; sur 252 collaborations (paires d'inventeurs) seulement deux ont été reconduites à plus de 5 occasions!

● **La figure 3 représente le profil du nombre d'inventeurs par document.**

Elle permet **de mesurer la taille des groupes de collaboration**. Ainsi, on observe que 75% des brevets étudiés ne comportent pas plus de deux noms d'inventeurs et que seulement 5% de brevets restants en contiennent plus de 3.

● **La figure 4 représente le profil du nombre de brevets déposés par inventeur.**

Elle permet **de mesurer le "monopole de recherche" de chaque individu**. Ainsi, on s'aperçoit qu'il n'y a pas dans notre exemple d'inventeur qui centralise de forts taux de dépôts ; seul 1% des inventeurs a déposé plus de 10 brevets. Par ailleurs, on relève que 70% d'entre eux ne possèdent qu'un seul brevet.

⇒ Ainsi à partir de ces premiers résultats, on en déduit facilement qu'il existe une grande diversité d'inventeurs par rapport aux nombres de documents traités.

⇒ De plus, ces inventeurs ne collaborent pas souvent ensemble et lorsqu'ils le font, il est rare que cette collaboration soit répétée. Par ailleurs, il semble probable que ce faible niveau de collaboration soit en fait le résultat d'un nombre de brevets relativement faible par inventeur.

*Cette première étape, à partir de ces quelques résultats statistiques, constitue une phase importante dans l'analyse de l'information traitée. Elle permet non seulement de mieux percevoir **sa structure** et **sa répartition** dans le corpus de documents analysés, mais aussi d'apporter des premiers résultats d'informations générales importants et **discriminants** pour le choix des méthodes d'analyses ultérieures.*

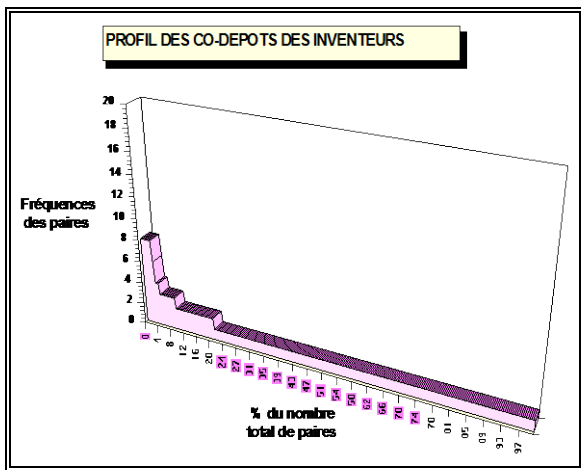


FIGURE 1

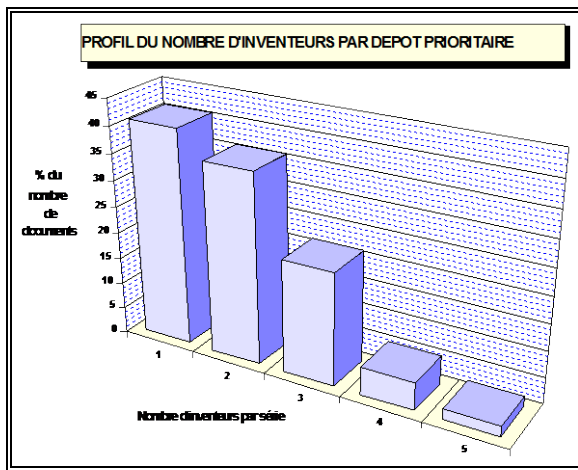


FIGURE 3

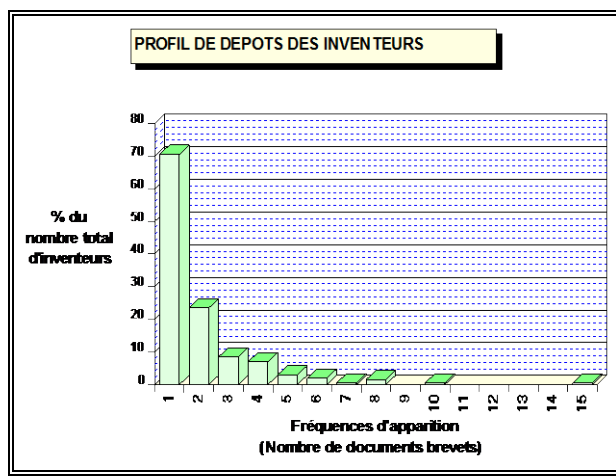


FIGURE 4

EXPLOITATION DES LISTES DE FORMES

DATAVIEW permet de visualiser la liste des formes relatives au corpus de documents traités. Il existe différentes possibilités de visualisation de cette liste. L'édition la plus simple permet d'obtenir, pour chaque forme recensée, sa fréquence d'apparition, son occurrence et sa valeur d'hapax. Le tableau 1 représente un extrait du listing de cette édition, obtenu à partir du traitement de l'information *Inventeur* de l'exemple précédent. Etant donné que chaque inventeur est enregistré une seule fois par document, il est normal de constater que le nombre de fréquence de chaque forme est égal ici à son nombre d'occurrence.

La notion de "*références-hapax*" est très intéressante pour ce type d'étude, car elle permet directement de savoir, parmi les inventeurs, quels sont ceux qui travaillent souvent tout seul. **Ainsi, dans notre exemple, on relève que l'inventeur ayant le plus de brevets à son actif est aussi celui pour lequel on enregistre le plus de dépôts sans collaboration.** On s'aperçoit par ailleurs, qu'il existe plusieurs inventeurs, titulaires d'un certain nombre de brevets, pour lesquels la totalité de leurs dépôts est réalisée sans aucune collaboration (lignes grisées sur le tableau 1).

EDITION DES FORMES DANS LES FREQUENCES : 1 - 15

FICHER TRAITE: D:\INVENT.JOB

FREQUENCE	OCCURRENCE	HAPAX	FORME
15	15	5	BOLICH R E
10	10	0	PARRAN J J
8	8	0	SAKKAB N Y
8	8	0	RUSSELL G D
8	8	1	LUEBBE J P
7	7	0	SUNBERG R J
6	6	3	WALLEY D R
6	6	0	TANNER P R
6	6	0	FARRIS R D
6	6	0	BISSETT D L
5	5	3	SNYDER W E
5	5	0	SCHWARTZ J R
5	5	2	JUNEJA P S
5	5	0	GROTE M B
5	5	0	BUSH R D
4	4	0	WOLFF A M
4	4	1	TORGERSON P M
4	4	1	STEURI C
4	4	2	SMITH G D
4	4	4	SAMPATHKUM P
4	4	2	SABATELLI A D
4	4	2	OH Y S
4	4	3	MCCALL P C
4	4	0	HUTCHINSON N K
4	4	0	GUAY C B
4	4	0	CHATTERJEE R
4	4	0	CASSIDY W A
4	4	0	CARPENTER R S
4	4	0	BENEDICT J J
3	3	0	WILLIAMS T B
3	3	3	WETZEL T A

TABLEAU 1

Dans ce type d'analyse, DATAVIEW présente la particularité, quelle que soit la forme considérée, **de pouvoir établir le lien entre cette forme et les références qui la contiennent**. Cette relation **FORME-REFERENCE**, établie à partir d'une édition spécifique, permet très facilement, lorsqu'un résultat remarquable est relevé, de "remonter" au document d'origine pour restituer la forme dans son contexte bibliographique.

Toutefois, pour mieux étudier chaque forme dans son environnement initial, sans pour cela retourner au document d'origine, DATAVIEW propose une édition particulière qui **permet de visualiser, pour chaque forme éditée, toutes les formes environnantes dans un contexte de proximité variable**. Nous avons représenté au tableau 2, un extrait de ce type de représentation concernant les 4 premiers inventeurs du tableau 1.

De la même façon qu'au tableau précédent (tableau 1), on retrouve, pour chaque forme, sa fréquence d'apparition dans le corpus, son nombre d'occurrences et sa valeur d'hapax mais aussi les numéros des références dans lesquelles chacune est enregistrée. De plus, on relève, pour chaque référence concernée, toutes les formes qui, dans le champ d'analyse, sont situées à proximité de celle considérée. Bien que DATAVIEW offre une grande souplesse dans le choix des valeurs de proximités de gauche et de droite, nous n'avons considéré ici que les formes adjacentes des deux côtés de celles examinées. Sur cette représentation, l'environnement de gauche est représenté à gauche du symbole "« »" et celui de droite, à droite du même symbole.

➡ Ainsi, il est facile **d'évaluer pour chaque inventeur quels sont ceux qui collaborent le plus avec lui**.

EXPLOITATION DES LISTES DE PAIRES

➡ Rappelons que la notion de "paire" détermine la relation d'appartenance de deux formes différentes à la même référence.

Ainsi, à partir de l'analyse du champ *INVENTEUR*, chaque fois que l'inventeur A sera présent dans le même document que l'inventeur B, la paire "A-B" sera formée. Dans ce type de traitement, il n'y a pas de différence entre une paire particulière et sa paire "bijective" ; c'est à dire que la paire A-B est identique à la paire B-A. Le tableau 3 représente un extrait de la liste des paires correspondant au champ *INVENTEUR* précédent.

EDITION DES PAIRES DANS LES FREQUENCES : 1 - 8

FICHER TRAITE: D:\INVENT.JOB
Intervalle de fréquences des formes constitutives: 1 - 15

FREQUENCE	CORRELATION	PAIRE
8	0.8	SAKKAB N Y — PARRAN J J
6	0.8	LUEBBE J P — TANNER P R
4	0.8	CASSIDY W A — SCHWARTZ J R
4	1	WOLFF A M — CARPENTER R S
4	0.7	BENEDICT J J — SUNBERG R J
3	0.5	FARRIS R D — SCHWARTZ J R
3	0.6	FARRIS R D — CASSIDY W A
3	0.8	LAD P J — CARPENTER R S
3	0.8	LAD P J — WOLFF A M
3	0.4	BOLICH R E — NORTON M J
3	0.6	RUSSELL G D — NORTON M J
3	0.2	RUSSELL G D — BOLICH R E
3	0.3	BOLICH R E — TORGERSON P M
3	0.5	BISSETT D L — BUSH R D
3	0.6	BISSETT D L — CHATTERJEE R
3	0.4	LUEBBE J P — FARRIS R D
3	0.4	RUSSELL G D — GROTE M B
2	0.6	SCHWARTZ J R — GEHRING T A
2	0.7	CASSIDY W A — GEHRING T A
2	0.4	HUTCHINSON N K — SCHWARTZ J R
2	0.4	HUTCHINSON N K — CASSIDY W A
2	0.3	HUTCHINSON N K — FARRIS R D
2	0.7	GUAY C B — HINTON J P

TABEAU 3

On s'aperçoit que la notion "**D'OCCURRENCE**" a laissé la place à celle de "**CORRELATION**". Cette dernière représente une mesure de la liaison entre les deux formes définissant chaque paire. Il s'agit d'un indice d'association permettant, au même titre que la fréquence, de mesurer l'intensité de la relation définissant la paire.

DATAVIEW propose, à cet effet, une vingtaine d'indices d'association différents, autres que le coefficient de *corrélation*, tels que la distance *Euclidienne*, le coefficient de *Jaccard*, de *Hamman* etc. **Selon le type de relation à mettre en évidence (similitude ou dissimilitude), l'utilisateur devra judicieusement choisir l'indice à utiliser.** La valeur du coefficient de *corrélation* utilisé ici varie entre 1 et -1. Une valeur proche de 1 signifie que les formes constituant la paire sont fortement liées alors qu'une valeur proche de -1 signifie le cas contraire. Quel que soit l'indice choisi, cette mesure est essentielle car **la valeur de fréquence ne suffit pas, à elle seule, à rendre compte de la nature de l'association entre les deux formes.**

- *En effet, en imaginant que la paire A-B ait une fréquence de 10 et que chaque forme A et B ait aussi une fréquence de 10, la nature de la relation A - B est différente du cas où, pour **une même fréquence de paire**, les formes A et B aurait une fréquence respective de 100. En associant ces formes à des inventeurs, cela signifie, dans le premier cas, que tous les brevets de A et de B sont issus d'une collaboration commune, alors que dans le deuxième cas, cette collaboration ne représente que 10% de leur dépôts respectifs. **Et pourtant, la fréquence de la paire A - B est identique dans les deux cas !***

Comme pour les éditions de formes, le logiciel permet d'éditer la totalité ou une liste restreinte de paires répondant à un intervalle de fréquences particulier ou répondant à des formes constitutives, elles-mêmes définies dans un intervalle de fréquences spécifique. Là encore, l'édition des références contenant les paires extraites est permise, ainsi que la possibilité d'effectuer différents types de tris (alphabétique, par fréquence, par indice de d'association).

⇒ **Une des caractéristiques essentielles de DATAVIEW, à propos de l'analyse des paires, est l'édition des paires obtenues par blocage d'une liste de formes.**

Cette fonction permet d'obtenir uniquement celles qui contiennent **simultanément, dans la même référence**, chacune des formes définissant cette liste. Cette particularité trouve des avantages dans différents types d'analyse. Concernant celle du champ *INVENTEUR*, cela permet de connaître les collaborations à trois, à quatre, ou à X inventeurs. Pour étudier l'émergence de nouveaux concepts d'études il suffira, à partir de l'analyse des informations Années de priorité et Descripteurs, d'éditer pour chaque année "bloquée", la liste des paires de Descripteurs associés et de comparer ces listes entre elles.

De façon générale, cette fonction d'édition sera utilisée lorsqu'il s'agira d'étudier la liaison entre des concepts dont le nombre est supérieur à 2.

Ainsi, dans le cadre d'une analyse par sujet, nous pourrions obtenir l'ensemble des paires de "mots-clés" associées à une liste pré-établie de "mots-clés" différents des premiers. Le même type d'analyse peut être réalisé pour détecter, globalement par société, les tendances de dépôts par pays de priorité et d'extension et ceci sans devoir traiter chaque société de façon isolée.

⇒ **La notion de paires permet de dégager des réseaux d'association entre formes.**

Par exemple, pour la liste de paires représentée au tableau 4, le réseau de collaborations correspondant est illustré à la figure 5. Dans ce type de représentation [PETERS89],[DOU89a], l'épaisseur des traits reliant deux formes est proportionnelle à la valeur de l'indice d'association qui les caractérise (ici la fréquence).

FREQUENCE	CORRELATION	PAIRE
10	1	A - B
10	1	A - C
5	1	B - C
5	1	J - K
5	1	K - L
1	1	B - D
1	1	B - H
1	1	B - I
1	1	C - D
1	1	C - H
1	1	C - I
1	1	E - F
1	1	E - G
1	1	F - G

TABLEAU 4

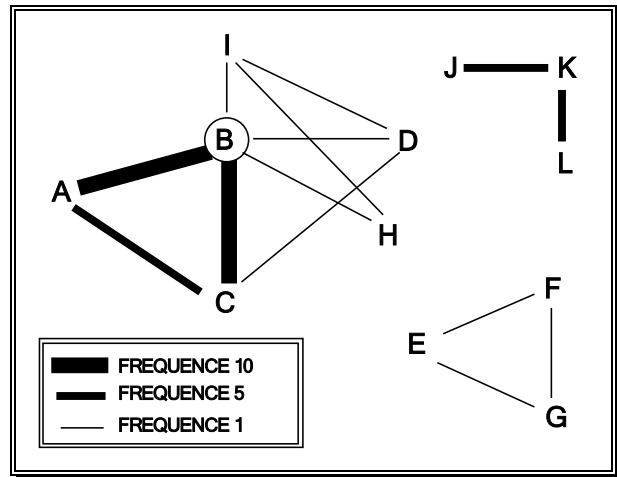


FIGURE 5

Cette représentation permet d'obtenir, graphiquement, la structure des relations entre chaque forme. On observe, au travers de cet exemple, différents types d'agrégats. Le plus ramifié définit le réseau de relation principal. On relève deux autres réseaux secondaires non reliés au premier. Ces réseaux **INDEPENDANTS** définissent des "**CLIQUES**". Lorsque tous les éléments d'une clique sont reliés entre eux, on parlera de "**clique fermée**" : c'est le cas des formes E, F et G. Dans le cas contraire on parlera de "**clique ouverte**" : c'est le cas des formes J, K et L.

Dans l'analyse du champ INVENTEUR, cette représentation permet :

- **d'identifier les différents groupes de collaborations**
- **de définir la structure et la taille de ces groupes**
- **d'évaluer les liens inter et intra-groupes**
- **d'identifier les principaux points nodaux de chaque groupe ; c'est à dire, les inventeurs qui possèdent un niveau de ramification élevé.**

La figure 6 représente l'image de ce réseau obtenu à partir de l'édition complète des paires d'inventeurs dont un extrait a été présenté au tableau 3. Dans cette représentation, chaque forme est suivie de sa fréquence d'apparition. On s'aperçoit que la structure générale de ce réseau est très éclatée, puisqu'on relève 12 cliques différentes dont 7 correspondent à des collaborations entre deux inventeurs.

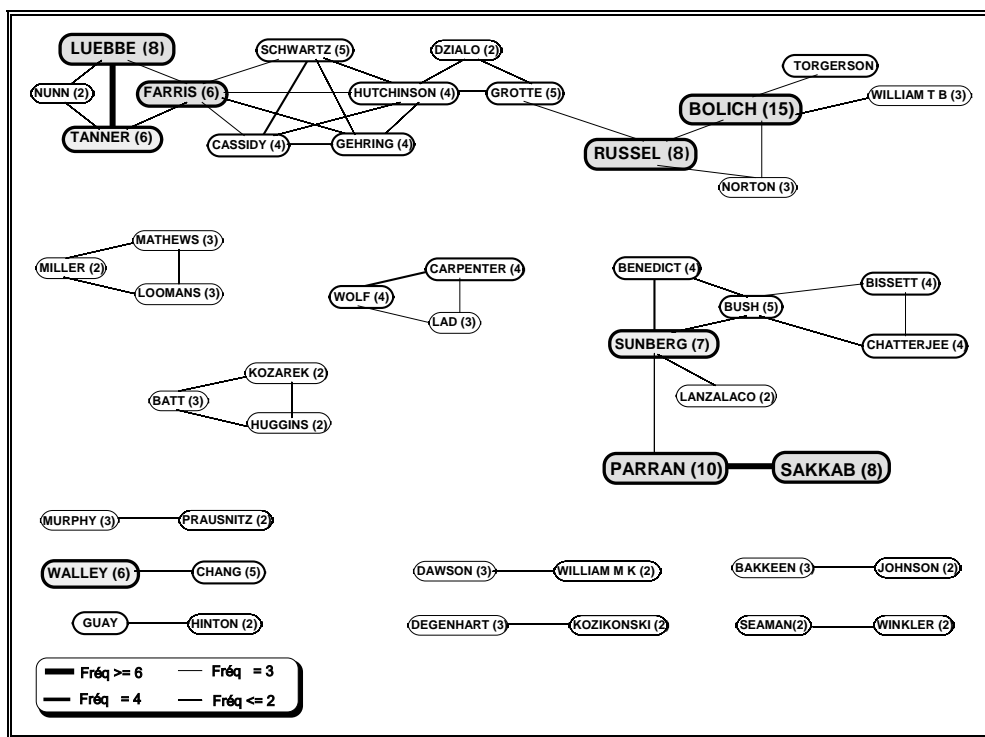


FIGURE 6

UNE REPRESENTATION SYNTHETIQUE PARTICULIERE : LES GRAPHES B.C.G. APPLIQUES AUX BREVETS.

L'évaluation des politiques de dépôts et de recherches à partir d'un corpus de documents brevets dépend d'un nombre important de variables d'observations qui se rattachent à différentes informations [NIVOL93]:

- ▣ **Priorité (date et pays de dépôts)**
- ▣ **Société**
- ▣ **Domaine d'activité principal (ou thème d'activité)**
- ▣ **Famille de brevet**
- ▣ **Extension (nombre de pays d'extension et liste des pays d'extension)**

Généralement, les différents types de traitements bibliométriques utilisés pour traiter l'information brevet conduisent à des résultats dont les représentations graphiques ne représentent que des **VUES PARCELLAIRES** des politiques générales de dépôts et de recherches développées par société. Il semble nécessaire d'obtenir, à partir d'une représentation **SYNTHETIQUE**, un résumé des principales caractéristiques de chacune des politiques.

➡ Pour cela, nous avons appliqué au domaine de l'information brevets, un type de représentation particulière : **LES GRAPHES B.G.C.**

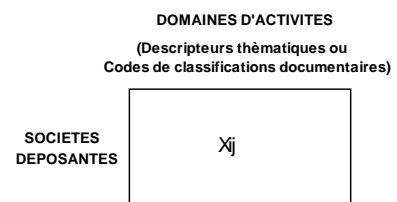
Ces graphes sont issus du modèle d'analyse de portefeuilles d'activités développé par le cabinet américain **Boston Consulting Group**⁽⁴⁾ (B.C.G.). Ce modèle, initialement conçu pour le **domaine de la stratégie d'entreprise**, a pour but d'établir des diagnostics stratégiques de sociétés. Il a été largement développé dans le cadre de travaux relatifs aux méthodes de "**diagnostic stratégique**" [MARTINET90] [DUR689]. Jacques MORIN [MORIN88], dans son ouvrage, y fait d'ailleurs largement référence. A partir de ce modèle, il est possible de déterminer, pour des domaines d'activités concurrentiels, les zones d'activités spécifiques pour lesquelles l'entreprise doit réagir pour ne pas se laisser dépasser.

Sur le plan de l'information brevet cette méthode permet de représenter pour chaque société, **LES PORTEFEUILLES DE DEPOTS BREVETS** selon différents critères d'observation : **domaine d'activité (représenté par différents codes de classifications documentaires), années de priorité, pays de d'extension, etc..**

➡ Par exemple, dans le cas où le critère d'observation est l'information "domaine d'activité", les graphes B.C.G. vont permettre d'évaluer, pour chaque société, la répartition de ses activités au travers de ces domaines et la position de chacune de ses activités par rapport à sa concurrente immédiate.

Ces graphes peuvent être obtenus, directement par **DATAVIEW**, à partir de la génération de différentes matrices de fréquences. Ces matrices ont pour principale caractéristique de mesurer l'intensité de la relation existante entre chacune des formes analysées prises deux à deux. C'est à dire, qu'elles vont permettre de comptabiliser le nombre de documents communs à chacune des formes.

Dans l'exemple précédemment cité, la matrice correspondante, servant de support à la construction des graphes, croise l'information "société" et l'information "domaine d'activité". En fait, on mesure, grâce à cette matrice, **la diversité de la recherche de chaque société au travers des différents domaines d'activités relevés dans le corpus de documents étudiés** (notons que ces domaines peuvent correspondre à différents descripteurs thématiques : index termes, etc., ou à des codes de classifications documentaires existant dans la base de données [ROSTAIN92] : Derwent Codes, Manual Codes, Codes de classification internationale des brevets, etc.).



(4) Le **Boston Consulting Group** est un cabinet américain de consulting en management et stratégie.

Dans ce cas, l'indice résultant " X_{ij} " correspond au nombre de documents brevets publiés par la société déposantes i dans le domaine d'activité j .

Pour observer les éventuelles discontinuités de ces répartitions d'activités au cours du temps, des traitements identiques peuvent être réalisés à partir du "**blocage**" d'une information permettant de repérer les documents au cours du temps (en ci nous concerne, cette information se rapporte à "l'année de priorité" de chaque famille de brevets).

Cette notion de "blocage", propre au logiciel DATAVIEW va permettre, dans notre cas, de générer des matrices dont chacune des formes lignes et colonnes correspondent à des documents qui se réfèrent à une année de priorité pré-définie.

Une fois les matrices de fréquences créées, les graphes sont **automatiquement** générés sous EXCEL, à partir d'un ensemble de marco-commandes.

Ces graphes permettent d'obtenir différents types informations :

🕒 **La répartition de ses activités.**

- ❑ En ordonnée, on définit le **TAUX DE RECHERCHE** comme étant le pourcentage de brevets prioritaires déposés dans un domaine considéré par rapport au total des brevets prioritaires déposés tous domaines confondus.
- ❑ L'ordonnée de la ligne horizontale est arbitraire. On a pris ici le cas d'une valeur moyenne établie sur l'ensemble des taux présents. Elle pourrait être placée différemment en fonction de préoccupations stratégiques particulières.

🕒 **La position de l'activité de la société par rapport à sa concurrente immédiate.**

- ❑ La ligne verticale d'abscisse 1, dite "**barre du leader**", délimite la zone **de gauche** où **la société considérée est LEADER dans le domaine en question**. Notons que, pour des raisons de représentations, l'échelle horizontale est de type logarithmique.
- ❑ Pour chaque domaine situé **à gauche** de cette barre (**abscisse > 1**), la valeur de l'abscisse lue sur le graphe indique **le coefficient** par lequel il faut multiplier le nombre de brevets prioritaires déposés (par la société étudiée) pour **atteindre celui du concurrent le plus proche**.
- ❑ Pour chaque domaine situé **à droite** de cette barre (**abscisse < 1**), la valeur de l'abscisse lue sur le graphe indique **le coefficient** par lequel il faut multiplier le nombre de brevets prioritaires déposés (par la société étudiée) pour **atteindre celui du concurrent leader**.
- ❑ Pour chaque domaine situé **sur la barre du leader (abscisse =1)**, la société étudiée est **leader ex aequo** avec un autre concurrent.

● **La mesure du monopole d'exploitation de chaque domaine d'activité considéré.**

- Les surfaces des cercles sont proportionnelles au nombre de brevets d'extension déposés. C'est à dire qu'ils représentent **la mesure du monopole d'exploitation pour chaque domaine considéré.**
- Ces cercles induisent *des gradients* qui permettent, à la lecture des graphes, de relever très facilement différentes anomalies dans les politiques d'extension par société.

Le diagramme se trouve ainsi divisé en quatre parties.

⇒ **La partie située en haut à gauche correspond à la zone VEDETTE.**

Dans cette zone, on travaille beaucoup et on est payé en retour puisque l'on est leader.

⇒ **La partie située en haut à droite correspond à la zone DILEMME.**

Les efforts dans cette zone sont aussi intenses que ceux relevés dans la précédente mais ici on n'est pas leader. Faut-il continuer ?

⇒ **La partie située en bas à gauche correspond à la zone VACHE A LAIT.**

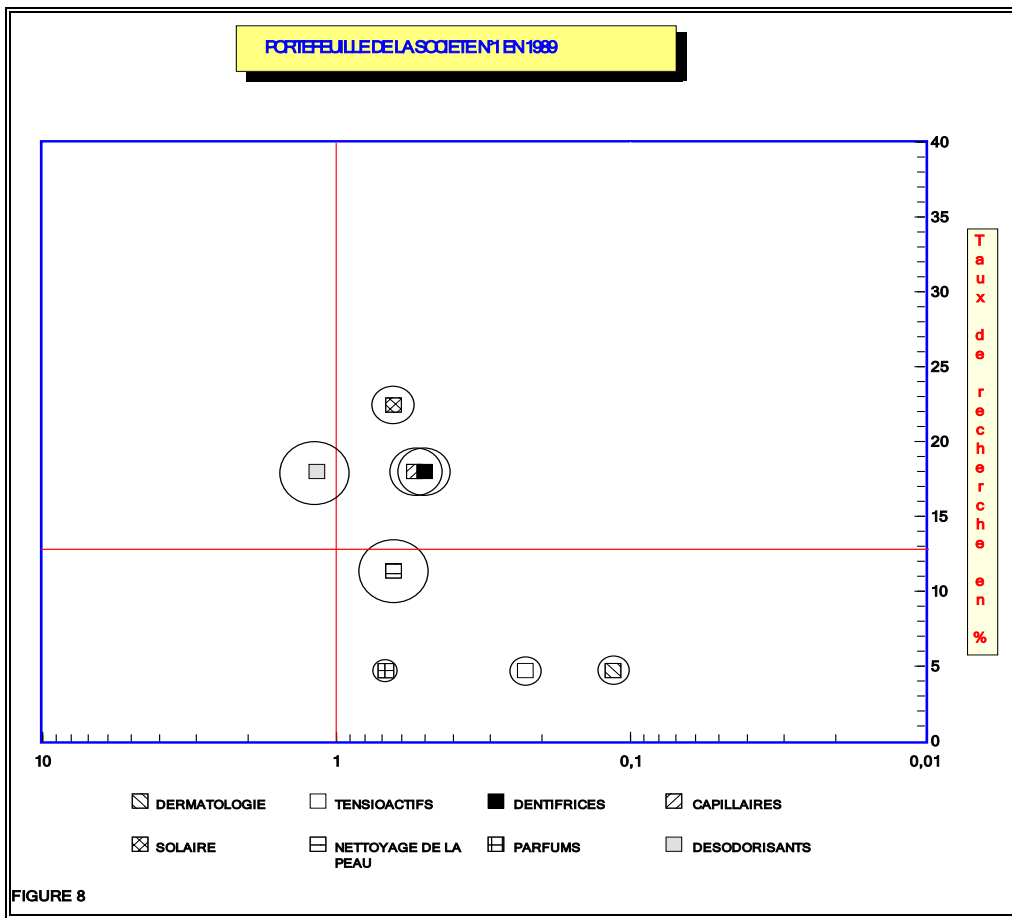
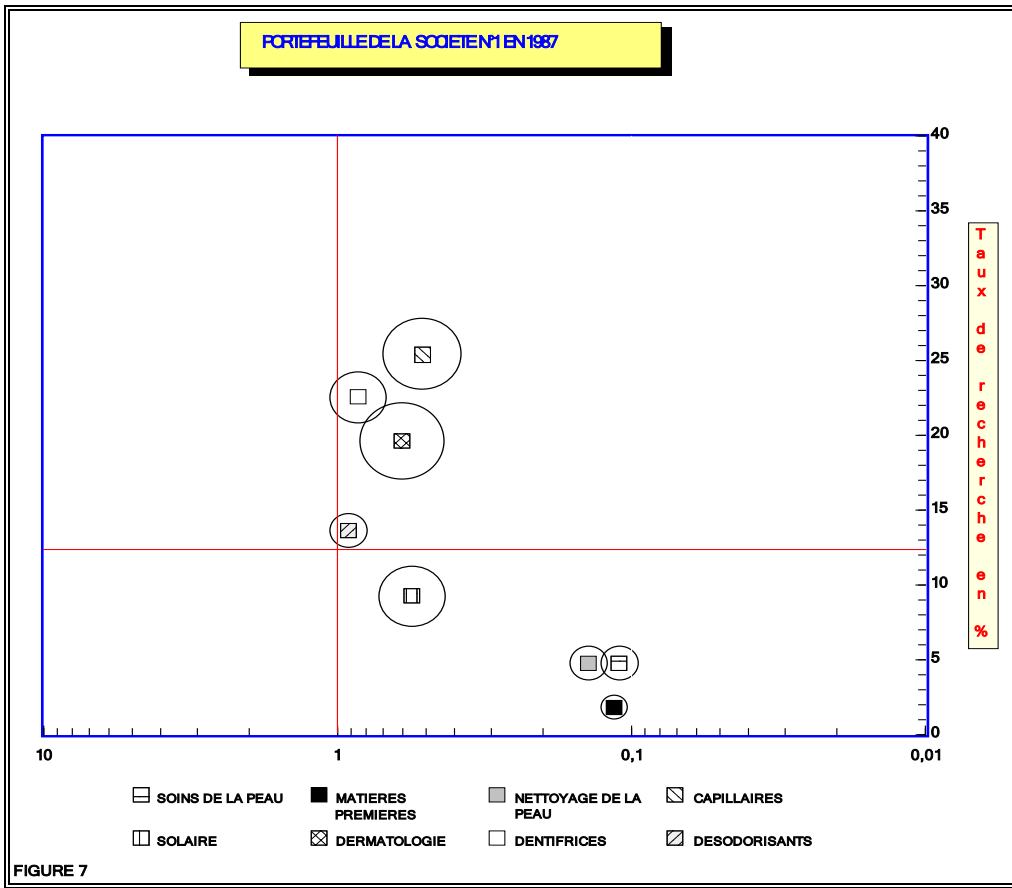
On domine sans effort.

⇒ **La partie située en bas à droite correspond à la zone POIDS MORT.**

On ne fait pas grand chose et on est dominé.

C'est bien sûr avec beaucoup de précautions que l'on manipulera ces notions. Un bon brevet situé dans la dernière zone peut avoir un impact commercial sans commune mesure avec son poids quantitatif.

Pour illustrer ces graphes, nous avons choisi de présenter deux exemples de sociétés. Pour chacune, deux graphes sont établis pour les années de priorité 1987 et 1989 (figure 7 à 10). Ces exemples sont extraits de cas concrets d'analyses ayant conduit à l'élaboration de dossiers d'informations stratégiques destinés à informer les décideurs dans l'entreprise. Les traitements ont été réalisés à partir d'un corpus de documents brevets téléchargés de la base de données brevets WPIL de Derwent. Le secteur analysé concerne **la cosmétique** en général. Il a été découpé en différents domaines d'activités secondaires. Ces domaines résultent d'une indexation documentaire thématique réalisée localement avec l'aide de spécialistes du secteur de la cosmétique, sur les documents du corpus téléchargé.



PORTFEUILLE DE LA SOCIÉTÉ EN 1987

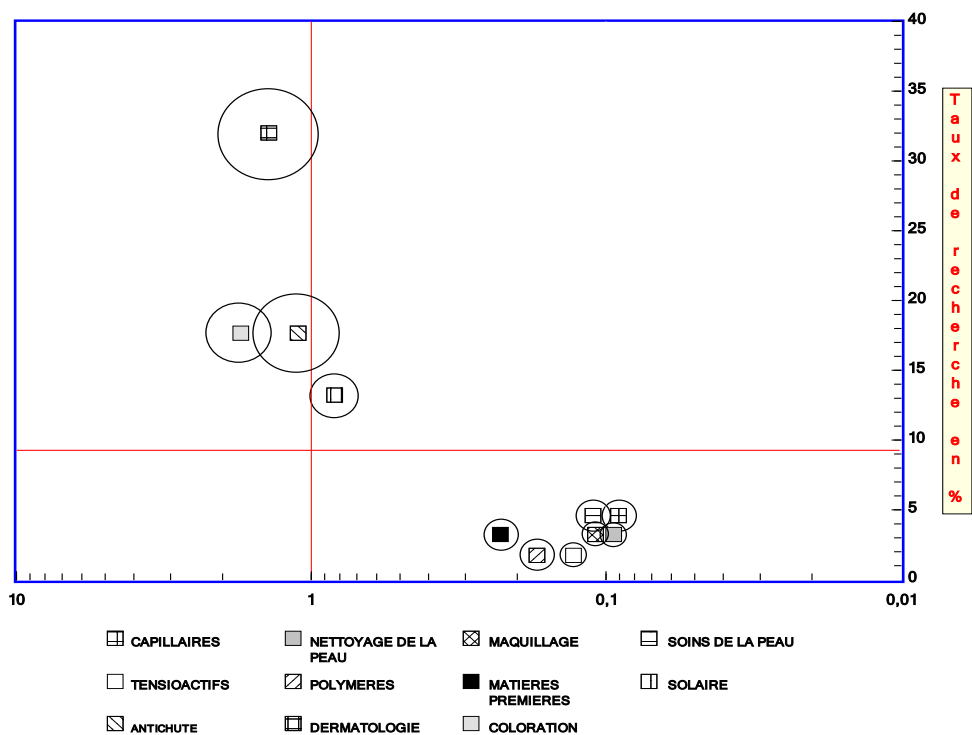


FIGURE 9

PORTFEUILLE DE LA SOCIÉTÉ EN 1989

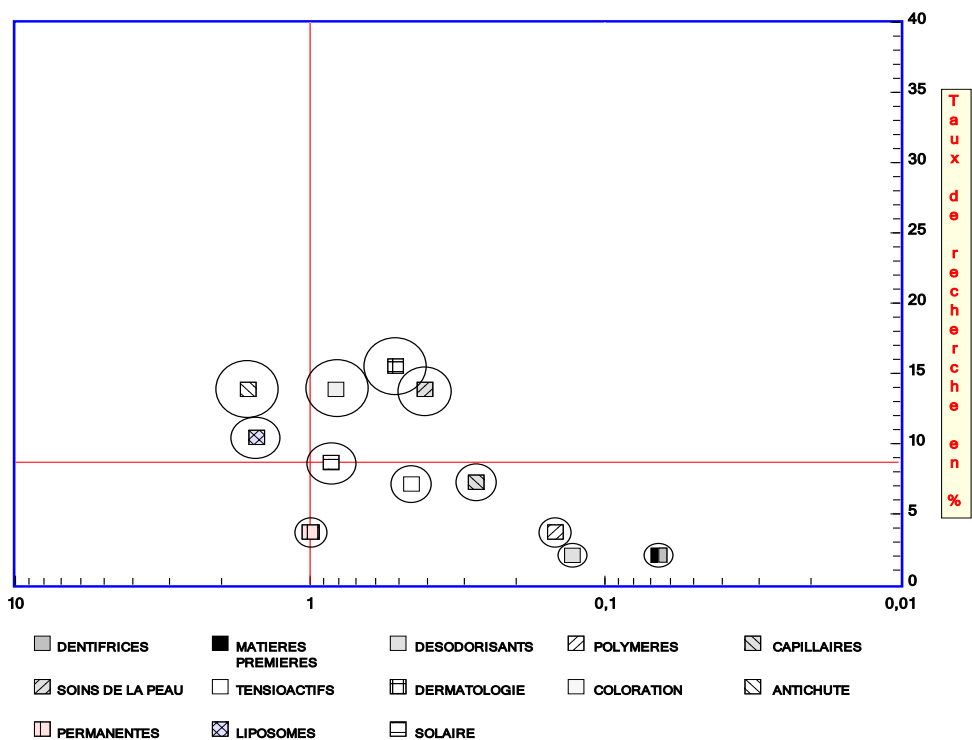


FIGURE 10

A partir de ces représentations, on constate très facilement **l'apparition et la disparition** de nombreux **thèmes de recherches** entre les deux années.

- ➡ C'est le cas, aux figures 7 et 8, des *Soins de la peau* qui ont complètement disparu en 1989 ou encore des *Parfums* qui émergent en 1989 alors qu'ils n'étaient pas présents en 1987.

Ainsi, il est possible de constater la régularité des dépôts dans chaque domaine de recherche.

D'autre part, on observe que des thèmes très faiblement situés sur l'échelle des taux de recherches en 1987, ont vu leur activité s'accroître en 1989.

- ➡ C'est le cas du *Solaire*, des *Désodorisants* et du *Nettoyage de la peau*.
- ➡ D'autres au contraire, comme le domaine de la *Dermatologie*, ont vu leur activité fortement diminuée.

Entre les deux années, on s'aperçoit que la société est devenue leader dans un seul domaine, celui des *Désodorisants*.

Toujours à partir des figures 7 et 8, on remarque, quelle que soit l'année, que les domaines d'activité se situent en général à droite de l'axe vertical mais proche de celui-ci. **Autrement dit, même si cette société ne domine pas souvent, elle se situe, tout de même, non loin de chaque leader.**

Les surfaces des cercles permettent de visualiser que les domaines les plus importants en taux de recherche ne sont pas forcément ceux pour lesquels les inventions brevetés sont les plus étendues à l'étranger. C'est le cas du *Solaire* à la figure 8 ou des *Désodorisants* à la figure 1.

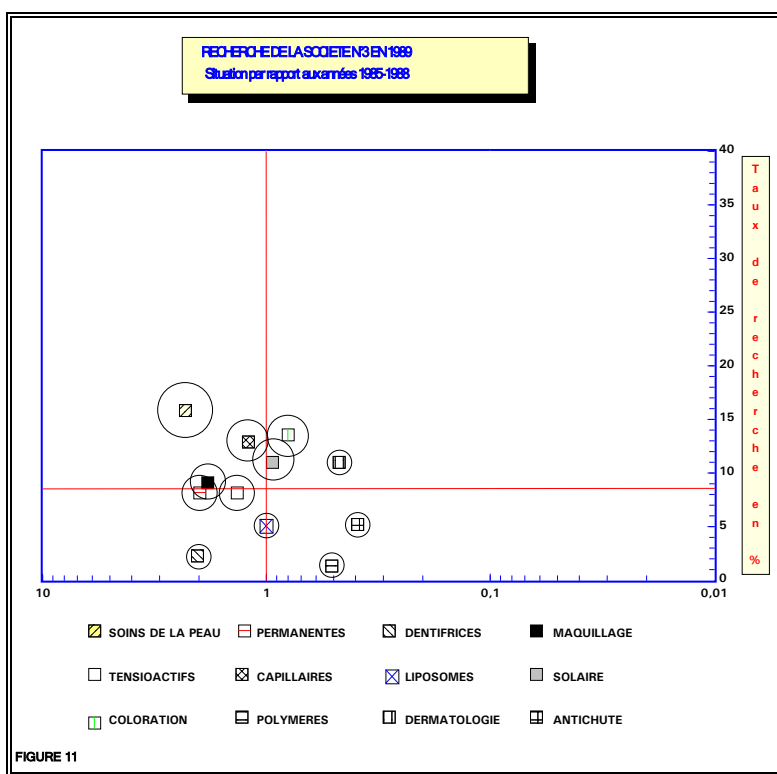
Enfin, la comparaison par société permet de révéler d'importantes dissemblances concernant la position géographiques des domaines dans le graphes. En effet, on constate que la société représentée aux figures 9 et 10 possède en 1987 un profil "**plus à gauche**" que la précédente.

- ➡ **Cette dernière est arrivée leader dans 5 domaines différents au cours des années analysées (*Antichute, Coloration, Dermatologie en 1987 et Permanente, Liposomes, Antichutes en 1989*) alors que la société N°1 n'a dominée qu'une seule fois en 1989 dans le domaine des *Désodorisants*.**

Ce type de représentation peut être appliqué de différentes façons. Nous avons illustré à la figure 11, un autre exemple d'application. Il ne s'agit plus ici de comparer les scores d'une société par rapport à ceux de ses concurrents mais d'établir une comparaison, pour une société donnée, entre les scores obtenus dans une année précise et ceux de chacune des autres années considérées.

- ➡ **Ce graphe permet de répondre à la question : est-ce que la société concurrente "X" a fait mieux en 1989 par rapport à chacune des autres années allant de 1985 à 1988 ?**

Ainsi, on voit pour la société analysée à la figure 11, qu'elle a obtenu, en 1989, les plus importants taux de recherche dans 5 domaines par rapport à la période allant de 1985 à 1988. Dans un domaine particulier, ce taux a été égalé une fois au cours des 4 années précédentes. Pour les 5 domaines restants, la valeur du taux de recherche a baissé. Cette baisse est considérable pour trois d'entre eux et relativement faible pour les deux autres.



Ces graphes permettent d'obtenir, dans une certaine mesure, des schémas de synthèse concernant un nombre important d'informations à partir d'une même représentation. Cette méthode, très utilisée en stratégie d'entreprise trouve au niveau de l'information brevets un nouveau domaine d'application.

Nous avons montré ici son applicabilité au niveau de la recherche mais elle peut **s'étendre à différents types de données.**

➔ Un autre exemple, que nous ne présenterons pas ici, concerne **la répartition géographique des portefeuilles brevets.** En effet, n'est-il pas intéressant de connaître, sur le plan brevet pour chaque concurrent, l'importance relative de ses marchés nationaux ? Dans quels pays chacun est-il leader ? Est-ce que les marchés de tel ou tel pays régressent ou au contraire augmentent ? Autant de questions auxquelles il est facile de répondre à partir de ce type de représentation, **dans la mesure où le nombre de variables à visualiser permet d'envisager ce type de construction graphique!**

III CONCLUSION

- ❶ **Quels sont les avantages apportés par l'outil d'analyse Dataview lorsqu'il est COUPLE aux techniques de surveillances classiques en entreprise ?**

La réponse à cette question se trouve principalement dans la description de ses principales caractéristiques.

- ❑ **Capacité à s'adapter à la diversité des sources et des types d'information à traiter** : pour une surveillance plus large sur le plan concurrentiel (scientifique, technique, technologique, etc.).
- ❑ **Capacité à traiter des volumes d'informations de dimensions illimitées** : pour obtenir une vision plus **exhaustive** et plus **globale** des différentes stratégies existantes.
 - ⇒ L'exploitation bibliométrique des informations à partir d'outils informatiques sur d'importants volumes de données permet d'entrevoir des faits saillants et des corrélations d'idées que l'analyste seul aurait du mal à observer en raison de ces volumes.
 - ⇒ Ces importantes quantités d'informations ne permettent pas facilement d'obtenir les résultats les plus élémentaires. Leur exploitation statistique nécessite l'utilisation d'outils appropriés.
- ❑ **Automatisme** : pour une exploitation plus **systématique** de l'information dans les surveillances menées.
 - ⇒ Une même analyse peut être réalisée plusieurs fois à différents moments afin de déceler l'apparition de phénomènes nouveaux au cours du temps.
- ❑ **Rapidité** : afin d'anticiper et de réagir suffisamment tôt face aux phénomènes relevés.
 - ⇒ Quelques jours suffisent, aujourd'hui, pour réaliser un examen complet et détaillé des différentes stratégies mises en jeu par la concurrence à partir de plusieurs milliers de documents brevets.

- ❶ **Nous avons montré le rôle essentiel du traitement de l'information en bibliométrie. Cependant, il ne faut pas sous-estimer dans cette tâche la partie relative à la représentation de l'information qui en résulte.**

- ❑ Très souvent, en bibliométrie, les méthodes de représentations font défaut. Certaines, issues des méthodes d'analyses statistiques multidimensionnelles, sont très performantes mais leurs interprétations nécessitent des compétences telles que seuls des spécialistes, coutumiers de ces techniques, peuvent réaliser. Par ailleurs, bien que de plus en plus de travaux concernent ces méthodes, elles demeurent tout de même encore peu utilisées en milieu industriel.
- ❑ D'autres, souvent trop sommaires, ne permettent d'obtenir que **des vues partielles de phénomènes qui se révèlent souvent complexes**. Dans ce cas, il faut multiplier ces vues pour arriver à dégager **une synthèse cohérente** des résultats qu'elles sous-tendent. Et

pourtant, il parait indispensable de livrer l'information élaborée sous une forme qui soit très **facilement et rapidement interprétable**.

Face à cette situation paradoxale, la méthode des graphes B.C.G. constitue un élément de réponse. En effet, ces représentations, largement appliquées et utilisées en stratégie par nombre de décideurs, **permettent facilement et rapidement d'obtenir des représentations simples de phénomènes multivariés complexes**.

- **Aujourd'hui, nous pouvons affirmer que le rôle de la bibliométrie dans le processus de veille technologique, et plus précisément celui au niveau de l'exploitation de l'information, est primordial.**

La bibliométrie ne se limite pas à la stricte élaboration et détermination d'indicateurs synthétiques spécifiques à un sujet traité. **Son apport dans la constitution du dossier d'information stratégique est essentiel.**

Elle permet :

- ⇒ d'élaborer des grilles de lecture de documents primaires,
- ⇒ de rendre un suivi systématique des sciences, des techniques et des technologies de la concurrence,
- ⇒ de déterminer l'évolution des tendances à partir de recherches rétrospectives pour mieux prévoir les comportements futurs,
- ⇒ l'exploitation du suivi thématique de documents et d'élaborer des vues panoramiques, de "cartographier" l'état d'un sujet pour mieux le maîtriser, de "photographier" au mieux son environnement technologique,
- ⇒ de mettre en place des systèmes d'alertes, des clignotants technologiques qui permettent de détecter le plus tôt possible, à partir de signaux faibles, les mutations technologiques en cours.

Cependant, elle ne peut être réellement efficace que si elle est couplée à d'autres méthodes d'analyse et de représentation de l'information, il s'agit notamment d'outils statistiques et infographiques.

- ⇒ **C'est par cette synergie que l'on pourra obtenir une information qualifiée d'élaborée, voire de stratégique lorsque que celle-ci aura été validée par l'avis d'experts compétents.**

L'opération de transformation d'une information brute en une information élaborée requiert des compétences particulières. Elle nécessite l'utilisation et le développement d'outils de traitements **spécifiques, automatiques et interactifs**.

Il est indispensable, pour assurer une gestion dynamique de l'information, que l'ensemble de ces outils soit intégré dans des systèmes **informatiques automatisés** afin d'obtenir, de façon rapide, une exploitation plus systématique de l'information recueillie.

Annexes

Liste des indices d'association statistique calculés dans DATAVIEW

Nom	Formule	Indications
Bray & Curtis	$(N_B + N_C) / (2 * N_A + (N_B + N_C))$	dissimilitude [0, 1]
Concordance	$(N_A + N_D) / M$	similitude [0, 1] (Sokal & Mich. 85)
Corrélation de Pearson	$((N_A * N_D) - (N_B * N_C)) / \sqrt{((N_A + N_B) * (N_A + N_C) * (N_B + N_D) * (N_C + N_D))}$	similitude [-1, 1]
Czekanowski	$(2 * N_A) / (2 * N_A + N_B + N_C)$	similitude [0, 1]
Différence de Forme	$(M * (N_B + N_C) - (N_B - N_C)^2) / M^2$	dissimilitude [0, 1]
Différence de Modèle	$N_B * N_C / M^2$	dissimilitude [0, 1]
Différence de Taille	$(N_B + N_C)^2 / M^2$	dissimilitude [0, 1]
Dispersion	$((N_A * N_D) - (N_B * N_C)) / M^2$	similitude [-1, 1]
Euclidienne	$(N_B + N_C) / M$	dissimilitude [0, 1]
Equivalence	$N_A^2 / (N_A + N_B) * (N_A + N_C)$	similitude [0, 1]
Faith	$(N_A + N_D / 2) / M$	(Faith 83)
Hamman	$((N_A + N_D) - (N_B + N_C)) / M$	similitude [-1, 1]
Inclusion	$N_A / \min \{(N_A + N_B), (N_A + N_C)\}$	
Jaccard	$N_A / (N_A + N_B + N_C)$	similitude [0, 1] (Jaccard 1900)
Kulczynski 1	$N_A / (N_B + N_C)$	similitude [0, ∞] (Kulczynski 28)
Kulczynski 2	$(N_A / (N_A + N_B) + N_A / (N_A + N_C)) / 2$	similitude [0, 1] (Sokal Sneath 63)
Moyenne ²	$(N_B + N_C) / M$	dissimilitude [0, 1]
Ochiai 1	$N_A / \sqrt{((N_A + N_B) * (N_A + N_C))}$	similitude [0, 1] (Ochiai 1957)
Ochiai 2	$N_A / \sqrt{((N_A + N_B) * (N_C + N_D) * (N_A + N_C) * (N_B + N_D))}$	
Q de Yule	$((N_A * N_D) - (N_B * N_C)) / ((N_A * N_D) + (N_B * N_C))$	similitude [-1, 1]
Rogers & Tanimoto	$(N_A + N_D) / ((N_A + N_D) + 2 * (N_B + N_C))$	similitude [0, 1] (Rogers Tanim. 0)
Russel & Rao	N_A / M	similitude [0, 1] (Russel Rao 40)
Shannon	$2 * (N_B + N_C) * \text{Log}(2)$	dissimilitude [0, ∞]
Sokal & Sneath 1	$2 * (N_A + N_D) / (2 * (N_A + N_D) + (N_B + N_C))$	similitude [0, 1] (Sokal Sneath 63)
Sokal et Sneath 2	$N_A / (N_A + 2 * (N_B + N_C))$	similitude [0, 1] (Sokal Sneath 63)
Sokal et Sneath 3	$(N_A + N_D) / (N_B + N_C)$	similitude [0, ∞]
Sokal et Sneath 4	$(N_A / (N_A + N_B) + N_A / (N_A + N_C) + N_D / (N_B + N_D) + N_D / (N_C + N_D)) / 4$	similitude [0, 1]
Sokal et Sneath 5	$N_A * N_D / \sqrt{((N_A + N_B) * (N_A + N_C) * (N_B + N_D) * (N_C + N_D))}$	similitude [0, 1]
Variance	$(N_B + N_C) / (4 * M)$	similitude [0, 1]

BIBLIOGRAPHIE

1. **[BROOKS87]**
BROOKS T.A.
 Bibliometrics Toolbox.
Software and documentation available from North City Bibliometrics 15825 6th Ave. NE. Seattle WA 98155, 1987
2. **[DESVAL92]**
Sous la direction de DESVALS Hélène et de Dou Henri
 La Veille technologique
Editions Dunod, 1992, 436 p
3. **[DURÔ89]**
DURÔ Robert
 L'atout concurrentiel
Les Editions D'Organisation, 1989, 162 p
4. **[DOU89]**
DOU Henri, Hassanaly Parina, QUONIAM Luc
 Infographic analytical tools for decision makers -
 Analysis of the research production in sciences,
 application to chemistry, comparaison between
 Marseille and Montpellier (France)
Scientometrics, 1989, vol 17, N° 1-2, 61-70
5. **[DOU90a]**
DOU Henri, HASSANALY Parina, QUONIAM Luc, LATELA Albert
 Veille technologique et information documentaire
Documentaliste, 1990, N°27 132-141
6. **[DOU90b]**
DOU Henri, QUONIAM Luc, Rostaing Hervé, NIVOL William
 L'analyse des données au service de la bibliométrie.
 Outils de veille technologique à la dimension des
 moyennes entreprises
Revue Française de Bibliométrie, 1990, vol 8, 27-67
7. **[LATELA87]**
LATELA Albert
 Système interactif d'aide à la décision (S.I.A.D.)
Thèse, Université d'Aix-Marseille III, 1987, 250 p
8. **[LEBART88]**
LEBART Ludovic, SALEM André
 Analyse statistique des données textuelle
Edition Dunod, 1988, 209 p
9. **[MARTINET90]**
MARTINET Alain-Charles
 Diagnostic stratégique
Edition VUIBERT. Collection Entreprise, 1990, 157 p
10. **[MORIN88]**
MORIN Jacques
 L'excellence technologique
Edition Publi Union, 1988, 253 p
11. **[MICHELET88]**
MICHELET B.
 L'analyse des associations
Thèse de Doctorat. Université de Paris VII, 26 octobre 1988
12. **[NIVOL93]**
NIVOL William
 Système de surveillance systématique pour le
 management stratégique de l'entreprise .
 Le traitement automatique de l'information brevet
Thèse, Université d'Aix-Marseille III, 10 mai 1993, 333 p
13. **[QUONIAM88]**
QUONIAM Luc
 Bibliométrie informatisée et information stratégique.
 Système automatique d'analyse des fichiers
 télédéchargés sur micro-ordinateur
Thèse, Université d'Aix-Marseille III, juillet 1988
14. **[PETERS89]**
PETERS H.P.F., VAN RAAN A.F.J.
 Structuring scientific activities by co-author analysis -
 An exercise a university faculty level
Scientometrics, 1991, vol 20, N°1 235-225
15. **[ROSTAING92]**
ROSTAING Hervé, NIVOL William, QUONIAM Luc, BEDECARRAX Chantal, HUOT Charles
 Exploitation systématique des bases de données -
 Des analyses stratégiques pour l'entreprise.
Congrès de l'ADEST, Paris 1 et 2 juin 1992
16. **[ROSTAING93]**
ROSTAING Hervé
 Veille technologique. Concepts - Outils - Applications
Thèse, Université d'Aix-Marseille III, 10 janvier 1993, 353 p