Infométrie mathématique et Infométrie statistique

Yves-François Le Coadic

Mathématique et statistique en science de l'information

Infométrie mathématique et Infométrie statistique

Yves F. LE COADIC

CNAM – ICST
2 rue Conté – 75141 PARIS Cedex 03
Téléphone/télécopie= (0)140272866
Courriel=lecoadic@cnam.fr

Résumé

L'application de la mathématique et de la statistique à l'étude des phénomènes informationnels a entraîné la naissance en science de l'information d'un nouvel axe de recherche et de développement, l'infométrie. Après avoir montré l'intérêt de cette application mais aussi avoir mis en garde contre certains abus et contre certains mauvais usages, nous présentons quelques exemples d'infométrie mathématique et d'infométrie statistique. Ils illustrent l'étendue et l'efficacité des analyses qui peuvent être faites sur une ou plusieurs variables informationnelles.

Abstract

Informetrics, the mathematical and statistical study of information processes, is a new promising field of research in information science. Advantages but also pitfalls and misuses of mathematics and statistics in social sciences are presented. A selection of applications (mono and multidimensionnal) coming from mathematical informetrics and statistical informetrics illustrate the efficiency of these methods.

Mots-clés

Mathématique, statistique, infométrie mathématique, infométrie statistique, bibliométrie, scientométrie, médiamétrie, muséométrie, webométrie, nombre, mots, documents, cartes, ZIPF, BOOLE.

Keywords

Mathematics, statistics, mathematical informetrics, statistical informetrics, bibliometrics, scientometrics, mediametrics, museometrics, webometrics, number, words, documents, maps, ZIPF, BOOLE.

« Or, je soutiens que dans toute théorie particulière de la nature, il n'y a de science proprement dite qu'autant qu'il s'y trouve de mathématique »

E. KANT – Premiers principes métaphysiques de la science de la nature

- INTRODUCTION

L'étude des phénomènes informationnels a révélé l'existence de régularités, de rapports mesurables, de distributions qui ne peuvent être mis à jour que par l'application de la mathématique et de la statistique.
Cela a donné naissance à un nouveau champ de recherches en science de l'information appelé INFOMÉTRIE. À l'intérieur de l'infométrie sont regroupés les sous-champs de recherches formés sur des secteurs informationnels spécialisés comme celui du livre, la bibliométrie (la première née), de la R&D (recherche-développement), la scientométrie, des mass-médias, la médiamétrie, des musées, la muséométrie et du WorldWideWeb, la webométrie (la dernière née).

Mathématique et statistique s'appliquent donc en science de l'information et ont, si l'on en juge par le panorama des applications que nous avons choisi de présenter, une incroyable efficacité. Mais elles peuvent aussi se révéler nocives si on n'en fait pas bon usage.

- I - LA MATHÉMATIQUE S'APPLIQUE

Traditionnellement, pour beaucoup, la mathématique s'applique pour construire des ponts, des machines; elle s'applique aussi en physique, discipline particulièrement mathématisée, en chimie, en biologie. De plus en plus aux sciences sociales comme l'économie, la psychologie, la sociologie et …la science de l'information. Mais dans l'esprit des professionnels de ce secteur, cela ne va pas forcément de soi.

Les succès de la physique classique, puis de la relativité et de la mécanique quantique ont mis en lumière sa pleine fécondité. Mais ce sont les beaux travaux de sociologie mathématique (R.BOUDON, J.S. COLEMAN) qui nous ont révélé son incroyable efficacité.

Qu'est-ce que cette efficacité ? Elle apparaît au travers de trois capacités : une capacité prédictive, une capacité rétrodictive et une capacité explicative.

Une capacité prédictive

La mathématique est efficace dans la mesure où elle suggère la réalisation d'observations ou d'expérimentations et fournit des résultats numériques qui, à une certaine marge d'erreur près, rejoignent les résultats empiriques issus de ces observations ou de ces expérimentations.

Une capacité rétrodictive

La mathématique est efficace parce qu'elle reproduit des résultats déjà connus en les organisant dans un formalisme concis. La mathématique fournit ici des outils servant seulement à « sauver les phénomènes ». Par exemple, grâce à la méthode des moindres carrés, on recherche des courbes passant au plus près des points expérimentaux.

Une capacité explicative

Pour qu'une théorie mathématique soit vraiment efficace en science, il faut qu'elle rende manifeste une explication des phénomènes, c'est-à-dire une suite d'inférences reliant leurs descriptions à des principes reconnus comme fondamentaux. Cette capacité explicative va de pair avec une capacité unificatrice (expliquer, c'est ramener la diversité des phénomènes à un très petit nombre de principes) et une capacité générative (suggérer des concepts nouveaux, des stratégies nouvelles).

En résumé, une mathématique efficace est un formalisme doué de capacités prédictives, rétrodictives et explicatives; autrement dit un langage permettant de décrire, d'expliquer et de maîtriser les phénomènes.

ATTENTION !

Si nous avons l'espoir que cette incroyable efficacité, que nos qualités de logique, de clarté devraient aider la science de l'information, il peut aussi avoir une contamination en sens inverse. Dans la mesure où la culture mathématique est imposée de façon artificielle, de l'extérieur, sans qu'il y ait – comme ce fut le cas en physique – de véritable exigence interne, les mathématiques perdent de leur caractère de sûreté puisqu'elles s'appliquent en définitive sur n'importe quoi et n'importe comment[1]. L'exigence en physique impose de repérer des régularités qu'on représente par des fonctions analytiques simples et d'exiger de bons ajustements. Alors qu'en

---

[1] **XIRDAL Zéphirin – Mathématiques et sciences humaines – Union libre ou mariage forcé – *Impascience*, 4/5, printemps 1976.**

bibliologie, discipline avatar de la bibliométrie, la tendance est plutôt la recherche de la corrélation même faible en s'en tenant au minimum de maths nécessaires.

Plus que partout ailleurs peuvent jouer l'esbroufe, la manière de faire croire que l'on comprend mieux que l'autre, les connivences entre initiés (les matheux) qui comprennent par-dessus la tête de ceux qui ne comprennent pas (les non-matheux).

Quelles sont alors les mathématiques efficaces pour décrire, expliquer et maîtriser les phénomènes informationnels ? Que représente la branche mathématique de l'infométrie et quelles sont les principales applications de mathématique infométrique ? Ce sera l'objet de notre première partie.

- II - L'INFOMÉTRIE MATHÉMATIQUE

Quelles sont les premières applications des mathématiques à l'étude des phénomènes informationnels? Elles vont constituer la branche mathématique de l'infométrie, branche que nous appelons infométrie mathématique. Ferons partie de cette branche les applications de ces mêmes mathématiques aux bibliothèques (bibliométrie mathématique), aux médias (médiamétrie mathématique), au WEB (webométrie mathématique), à la recherche-développement (scientométrie mathématique) et aux musées (muséométrie mathématique).

Les applications mathématiques peuvent prendre en compte une information ou un ensemble d'informations.

2.1. - une information :

- ***La fonction puissance et la mesure de la fréquence des mots dans un texte (loi de Zipf)***

Les fonctions polynomiales simples sont bien connues :

$$y = x^m$$

où l'exposant *m* est un nombre entier positif ou négatif.

x^m signifie que l'on fait :

- *m* fois le produit de *x* si *m* est un entier positif : c'est la fonction puissance,
- *m* fois l'inverse de ce produit si *m* est un entier négatif : c'est la fonction hyperbolique[2]. Quel que soit *m* entier positif, on a :

$$y = x^{-m} = \frac{1}{x^m}$$

[2] **G.K. Zipf, *Human behavior and the principle of least effort*, Cambridge, Addison-Wesley, 1949 (Reprinted Hafner, New York, 1965).**

Application :

Ce qui caractérise un certain nombre de phénomènes informationnels, ce sont des comportements de nature hyperbolique[3], c'est-à-dire que le produit de puissances fixes des variables est constant :

$$F(x).x^n = constante$$

Dans leurs manifestations discrètes, cela se traduit par le fait qu'à une cause croissant de façon géométrique correspond un effet croissant de façon arithmétique.
Ainsi, le nombre d'occurrences de tout objet dans un ensemble, par exemple un livre dans une collection ou un mot dans un texte, obtenu par comptage, est appelé fréquence. Si on ordonne les objets en fonction de leur fréquence décroissante, on peut leur attribuer un rang. Plusieurs objets ayant la même fréquence auront des numéros d'ordre consécutifs. Les propriétés des courbes (rang/fréquence) ont été observées et étudiées dans des domaines très variés. Dans les années 50, George Zipf s'est intéressé à la fréquence des mots dans les textes. Il a observé une relation constante, de type hyperbolique, entre la fréquence et le rang des mots :

Rang. Fréquence = constante (notée k)

La relation entre rang et fréquence est de type puissance inverse d'exposant $b \geq 0$:

$$U(r) = \frac{k}{r^b}$$

où *U* représente la fréquence et *r* le rang.

- ***La fonction exponentielle et l'obsolescence de l'information :***

La fonction exponentielle est parfois appelée « fonction de croissance naturelle » car de nombreux processus naturels, comme la croissance d'une forêt, d'une population ou du nombre des publications scientifiques, varient de façon exponentielle.
La fonction exponentielle dite de base *e*(*e*=2,72828…, constante d'Euler) est notée :

$$exp(x) = e^x$$

[3] En science de l'information, on a l'habitude d'appeler fonction hyperbolique toute fonction puissance ayant un exposant négatif, qu'il soit entier ou non.

Application :

Corollaire de la croissance rapide du nombre de publications, il existe une obsolescence également rapide du stock d'informations disponibles. Ce qui veut dire que si les références à la littérature passée sont distribuées de façon aléatoire, sans rapport avec la date de publication, une majorité d'entre elles renvoie à des travaux récents, puisqu'il y a plus d'articles disponibles pouvant être cités :

$$C(t) = C(0)e^{-at}$$

où *a* est un nombre positif supérieur à 1 (figure 1).

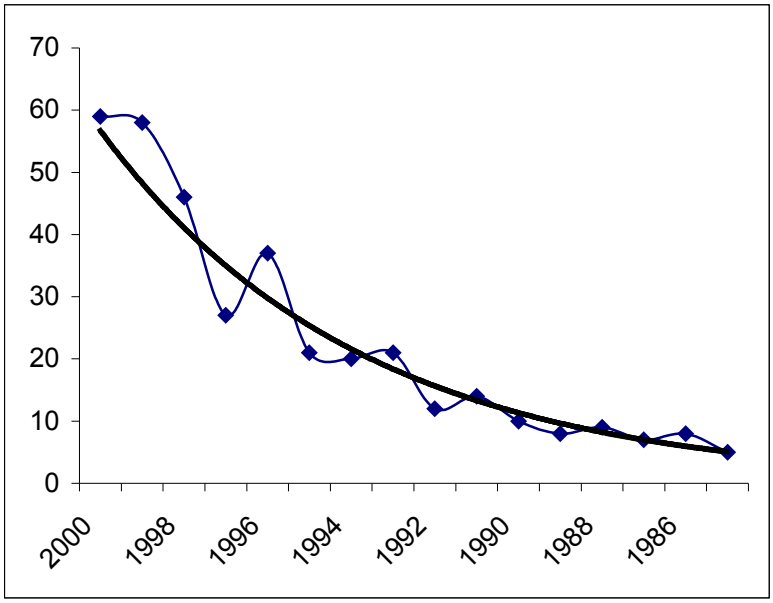

Figure 1 –Obsolescence de l'information

Les recherches sur la demi-vie des littératures scientifiques fournissent des éléments permettant d'éclairer ce type d'interrogation. La demi-vie d'une littérature est le temps pendant lequel la moitié de la littérature active a été citée. Les études d'obsolescence des différentes littératures ont montré des variations importantes de cette caractéristique : 4,6 années en physique, 7,2 années en psychologie, 10,5 années en mathématiques. De façon identique, connaissant le nombre total de citations reçues par une revue, la demi-vie de cette revue mesure le nombre d'années pendant lesquelles elle a reçu 50 % de ces citations. À titre d'exemple, voici les valeurs de ces demi-vies pour quelques revues de science de l'information :

Revues	Demi-vies (années)
J AM SOC INFORM SCI	**6,8**
SOC STUD SCI	**9,6**
SCIENTOMETRICS	**5,1**
INFORM PROCESS MANAG	**6,8**
J INFORM SCI	**6,2**

Tableau 1 : Demi-vie des revues en science de l'information (année 1999) (source JCR)

2.2 - un ensemble d'informations:

- ***La logique classique booléenne et le repérage de l'information:***

La logique classique booléenne du nom du mathématicien George Boole (1815-1864) (encore appelée logique mathématique) identifie, sur des ensembles finis, trois relations de dépendance grâce aux opérateurs booléens ET, OU et NON. Ces trois opérateurs permettent d'effectuer les importantes opérations ensemblistes (figure 3) que sont respectivement l'intersection, l'union et le complémentaire.

ET (produit logique) relie les composantes d'une phrase,

OU (somme logique) relie les termes synonymes ou quasi synonymes,

NON (négation logique) élimine les termes.

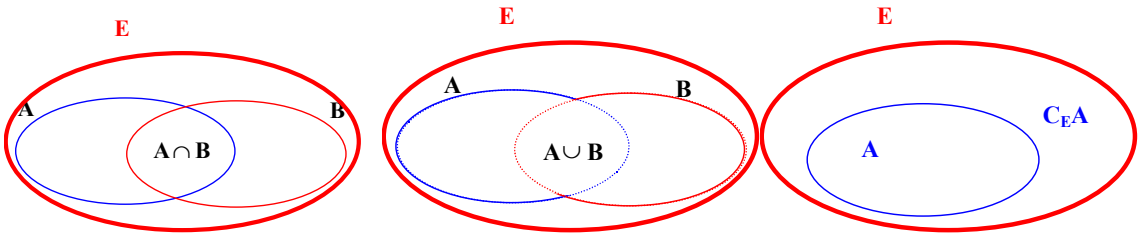

Figure1 – Opérations ensemblistes

ATTENTION, le OU utilisé ici est le « ou » logique et non pas le « ou » exclusif utilisé dans le langage courant.

Application :

Un exemple d'équations de recherche booléenne lors d'une interaction informationnelle personne-ordinateur (P-O) (U représente l'usager et O l'ordinateur)

U - question 1 = "Qu'avez-vous sur l'esclavage aux Etats-Unis?"
interrogation = (slave?) and (United(w)States) or America?)
O - réponse 1 = 2504 références

U - question 2 = " et sur les soulèvements des esclaves dans le Sud avant la guerre de sécession?"
interrogation = (slave?) and (rebellion? or uprising?) and (south?) and HP=1800h)
O - réponse 2 = 21 références

U - question 3 = "plus précisément, sur l'effet de la rébellion de Nat Turner en Virginie?"
interrogation = Nat(w)Turner and Virginia
O = réponse 3 = 13 références[4].

- Les vecteurs et la similitude entre questions et réponses :

Dans l'espace à trois dimensions de la géométrie euclidienne, on appelle vecteur un segment de droite orienté. Si (a_1, $a_{1,...}$ a_1) est un point dans cet espace, alors la ligne qui va de l'origine (0,0,…0) à ce point est le vecteur. Il est représenté par une flèche.

Application :

Comment peut-on mesurer la proximité de deux ensembles informationnels qui sont définis selon plusieurs critères ? Un des modèles de description possible des ensembles est celui des espaces vectoriels, développé par Salton[5].
Soit un ensemble *D* de documents et *M* l'ensemble des *m* mots $\{M_1, M_2 .... M_i, ...M_m\}$ présents dans les documents. Chaque document sera représenté sous la forme d'un vecteur ayant *m* composantes :

Document A : $\vec{A} = [a_1 \quad a_2 \quad ... \quad a_m]$

Document B : $\vec{B} = [b_1 \quad b_2 \quad ... \quad b_m]$.

[4]KENNEDY L., COLE C., CARTER S. - Connecting on-line strategies and information needs: a user-centered focus labeling approach - RQ, 36, 4, 1997.

[5] G. Salton and M.J. McGill, *Introduction to modern information retrieval,* New York, McGraw-Hill, 1984.

Dans un espace à trois dimensions, les documents seront donc représentés de la façon suivante :

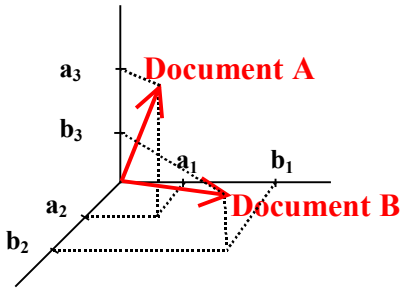

Figure 4 – Représentation vectorielle des documents A et B dans un espace à trois dimensions

Les valeurs a_i et b_j sont les « poids » des mots M_i et M_j présents dans les documents A et B. Ils quantifient la manière dont A et B sont représentés par ces deux mots.

Ce type de modèle a été utilisé pour calculer la proximité d'une question (composée de m mots) et d'un document, et pour calculer la proximité de deux documents.

Pour déterminer cette proximité, on calcule le cosinus de l'angle que forment les deux vecteurs documents entre eux :

Le cosinus ou coefficient de Salton : $$Cos(\vec{A},\vec{B}) = \frac{\vec{A} \bullet \vec{B}}{\left\|\vec{A}\right\|.\left\|\vec{B}\right\|} = \frac{\sum_{k=1}^{m} a_k b_k}{\sqrt{\sum_{k=1}^{m} (a_k)^2}\sqrt{\sum_{k=1}^{m} (b_k)^2}}$$

$\vec{A} \bullet \vec{B}$ est le produit scalaire des vecteurs $\vec{A}\ et\ \vec{B}$ et $\left\|\vec{A}\right\|\ et\ \left\|\vec{B}\right\|$ désignent la norme euclidienne des vecteurs $\vec{A}\ et\ \vec{B}$.

- III - LA STATISTIQUE S'APPLIQUE

La statistique, une branche de la mathématique, s'applique à l'analyse des valeurs numériques ; en particulier, celles pour lesquelles une étude exhaustive est impossible, à cause de leur grand nombre et de leur complexité. La valeur statistique obtenue pour une variable est une estimation de la valeur vraie de cette variable. Une fois collectées, les valeurs numériques devront être analysées de façon à les mettre en ordre, à leur donner un sens :

- l'analyse peut être simplement descriptive, donnant par exemple un état des usages faits de l'information ou du système d'information par les usagers. On fera alors appel à la statistique descriptive.
- l'analyse peut être aussi interprétative, permettant de dire ce que signifient ces valeurs. C'est alors la statistique bidimensionnelle qui décrit et mesure la liaison entre deux variables informationnelles et à la statistique multidimensionnelle qui décrit les relations existant entre trois et plus de trois variables informationnelles.

Le dimensionnement de ces analyses sera différent selon que l'on a en vue un travail consistant, c'est-à-dire de recherche approfondie, ou une évaluation rapide. Dans le premier cas, recherchant dans les valeurs des relations qui permettront d'infirmer ou de confirmer les hypothèses formulées, il sera nécessaire de travailler avec un grand nombre de variables informationnelles. Dans le second cas, on aura seulement besoin d'une analyse à deux ou trois dimensions. La démarche traditionnelle statistique qui consiste à confirmer les hypothèses formulées a considérablement évolué avec la généralisation d'outils d'analyse statistique multidimensionnelle (encore appelés en France analyse de données) qui, en particulier grâce aux outils infographiques, permettent de formuler des hypothèses que l'on vérifiera ensuite en utilisant d'autres méthodes, comme les statistiques exploratrices ou « fouilles de données » (texte mining, data mining, Web mining).

En résumé, une statistique efficace fournit des méthodes descriptives, interprétatives et exploratrices permettant d'évaluer la validité des modélisations des phénomènes informationnels qu'elle propose.

ATTENTION, ce peut être un moyen de mentir ! Stade suprême de l'impérialisme mathématique, la statistique prétend formaliser la démarche scientifique en proposant des règles pour évaluer la validité d'un modèle. Il est, bien entendu, que l'on peut développer toutes sortes de modèles statistiques autour des phénomènes sociaux et en particulier des phénomènes informationnels. Mais ce qui est suspect, c'est cette tendance à la complication non nécessaire. C'est aussi la pénombre discrète où on laisse l'évaluation des limites d'un modèle.
Pourtant un des mérites de l'attitude scientifique classique est de connaître ses propres limites. Ici, les insuffisances, quand elles sont reconnues, sont justifiées par le fait qu'il s'agit des débuts d'une nouvelle science[6]. Prédiction et analyses sont faites dans le flou[7].

Quelles sont alors les statistiques efficaces pour décrire, expliquer et maîtriser les phénomènes informationnels ? Et que représente la branche statistique de l'infométrie et quelles sont les principales applications de statistique infométrique ? Ce sera l'objet de notre deuxième partie.

- IV - L'INFOMÉTRIE STATISTIQUE

Quelles sont les premières applications des statistiques à l'étude des phénomènes informationnels? Elles vont constituer la branche statistique de l'infométrie, branche que nous appelons infométrie statistique. Ferons partie de cette branche les applications de ces mêmes statistiques aux bibliothèques (bibliométrie statistique), aux médias (médiamétrie statistique), au WEB (webométrie statistique), à la recherche-développement (scientométrie statistique) et aux musées (muséométrie statistique).

---

[6] XIRDAL Zéphirin, op. cité.

[7] Exception notoire: les fourchettes des pronostics électoraux, un des grands jeux de la télévision technocratique ! Les experts se portent bien mais s'en tirent mal comme on l'a vu en 2002. Du fait même qu'ils sont des experts, il y a des choses que les experts ne peuvent pas prévoir. Ce qui n'empêche pas qu'ils peuvent aussi causer des dégâts.

Les applications statistiques peuvent prendre en compte une variable informationnelle, deux variables informationnelles ou une multiplicité de variables informationnelles.

- **- une variable informationnelle :**

La statistique unidimensionnelle fournit des méthodes et des procédures permettant de résumer des grands ensembles de valeurs numériques d'une variable afin de les rendre intelligibles, de communiquer l'essence de ces valeurs.

- ***Les taux et l'évaluation des produits et des services d'information :***

Le taux de croissance (ou de décroissance) est une catégorie de taux particulièrement intéressante. Il est calculé en déterminant la différence entre la valeur d'une variable au début d'une période donnée et sa valeur à la fin de cette période et en divisant cette quantité par la valeur de la variable au début de la période.

Application :

Le taux de croissance d'un service en ligne qui est passé de 5 000 connexions en 1997 à 15 000 en 2002 est de :

$$Taux\,de\,croissance = \frac{15000-5000}{5000} = 2$$

En pourcentage, le nombre de connexions s'est accru de 200 % en 5 ans, soit 40 % par an. Le nombre de connexions a été multiplié par 3. Mais attention, il n'y a pas 300 % d'augmentation !.

- **- deux variables informationnelles :**

La statistique bidimensionnelle est plus audacieuse et donc plus risquée. Elle permet de découvrir les liens qui existent entre deux de ces variables.

- La co-occurrence et les cartographies informationnelles

Considérons un ensemble d'articles scientifiques où chacun est caractérisé par différents mots. Nous ne connaissons a priori ni ces mots, ni leur nombre. Les premiers traitements simples que l'on peut faire sont d'établir la liste des mots utilisés et de calculer leurs fréquences (nombre d'occurrences), puis de s'intéresser à la co-occurrence de deux mots, c'est-à-dire au nombre de fois qu'ils apparaissent ensemble dans un texte. Si les mots sont ainsi associés, les intérêts des auteurs des articles le sont aussi.

Le rôle des mots en tant qu'opérateurs de l'auto-structuration des domaines scientifiques et techniques a été en effet mis en évidence. Les mots indiquent quels sont les sujets intéressants dans un domaine de recherche donné à un moment donné. Lorsque deux mots apparaissent simultanément dans un ensemble d'articles, les sujets qu'ils représentent sont associés. Les schémas d'association des mots permettent donc de mettre en évidence les tendances de la recherche, ainsi que les principaux centres d'intérêt des chercheurs (figure 5).

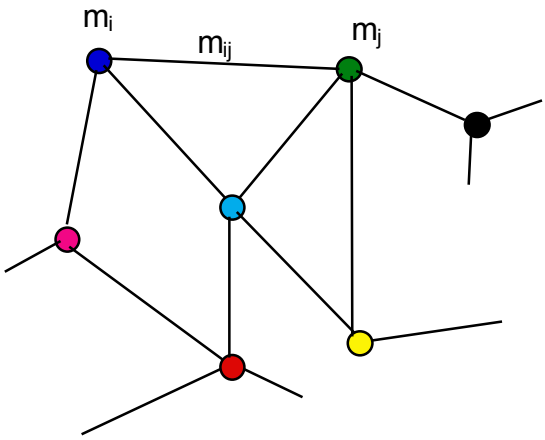

Figure 5 – Réseau d'associations de mots

Pour construire le réseau des associations de mots, la première étape consiste à calculer le nombre d'occurrences m_i de chaque mot i dans l'ensemble d'articles et le nombre de co-occurrences m_{ij} de chaque paire de mots m_i et m_j. Cependant, la co-occurrence ne permet pas à elle seule de mesurer la force des associations entre les mots, car elle avantage les mots apparaissant un grand nombre de fois par rapport aux autres. On calcule donc un coefficient normalisé (c'est-à-dire dont les valeurs sont comprises entre 0 et 1), croissant avec le nombre de co-occurrences, appelé le coefficient d'association noté E_{ij}. Un coefficient d'association égal à 1 signifie que les mots i et j sont systématiquement trouvés ensemble ; un coefficient à 0 signifie au contraire qu'ils ne sont jamais ensemble dans un document. Il existe plusieurs méthodes de calcul d'un coefficient d'association. Le coefficient utilisé dans la méthode des mots associés est :

$$E_{ij} = \frac{m_{ij}^2}{m_i . m_j}$$

Ce coefficient varie entre 0 et 1. Il vaut 0 si les mots i et j n'apparaissent jamais simultanément et 1 dans le cas inverse.

On trouvera sur la figure 6 un graphe des mots portant sur les revêtements céramiques ; les textes analysés proviennent d'une banque de brevets et sont constitués des titres et des résumés de 16 000 brevets extraits de cette banque.

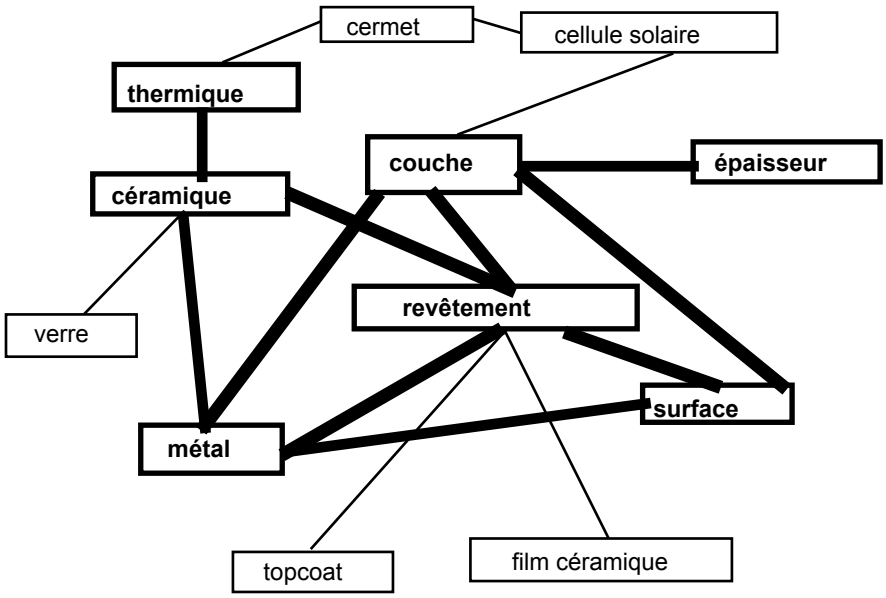

Figure 6 – Graphe « revêtement céramique »

Dans ce graphe, l'épaisseur des traits entre les mots est proportionnelle à l'intensité de liaison des associations. Chaque amas ou cluster est alors caractérisé par les mots du thème et ses associations internes et externes.
Cette classification permet ensuite de construire une représentation cartographique originale appelée diagramme stratégique. La figure 7 représente le diagramme stratégique « revêtement céramique ».

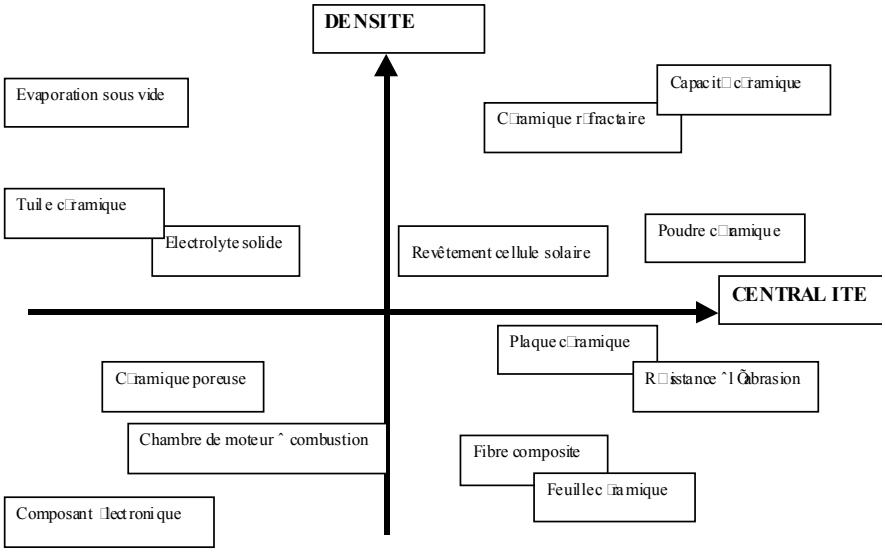

Figure 7 – Diagramme stratégique « revêtement céramique »

Les thèmes de recherche sont ainsi représentés dans un diagramme bidimensionnel défini par deux indicateurs : la densité et la centralité.

- La densité d'un amas est la moyenne de ses associations internes : c'est un indicateur de sa cohérence.

- La centralité d'un amas est la moyenne de ses associations externes : elle indique si l'amas est plus ou moins isolé.

Les sujets qui apparaissent dans le quadrant en haut à droite sont les sujets qui sont dans le front de la recherche (très structurés, très développés). Par contre, ceux qui se situent dans le quadrant en bas à gauche sont des sujets périphériques et faiblement structurés.

- **– de multiples variables informationnelles :**

La statistique multidimensionnelle permet de rechercher les relations existant entre plusieurs variables. Elle fournit des outils dont on attend d'abord qu'ils aient une efficacité pratique, la justification théorique n'étant recherchée qu'en second. On distingue trois grands types de méthodes : la classification, l'analyse factorielle des correspondances et l'analyse relationnelle. Par exemple :

- *L'analyse factorielle des correspondances et la typologie des productions d'informations scientifiques techniques selon les disciplines :*

L'analyse factorielle consiste à traiter des grands tableaux de nombres, difficiles à lire, en les remplaçant par des tableaux plus simples qui soient une bonne approximation des premiers. Les diverses méthodes d'analyse factorielle (en particulier l'analyse factorielle des correspondances, que nous étudions ci-dessous) emploient le même procédé : étant donné un nuage de points (les individus), munis de masse (les effectifs), dans un espace dont le grand nombre de dimensions interdit la visualisation du nuage, espace muni d'une métrique (qui mesure la distance entre les individus), il s'agit de trouver les axes d'inertie du nuage et d'obtenir des visualisations sur des plans formés par les couples d'axes. On pourrait résumer ceci en disant que ces méthodes d'analyse permettent de « géométriser des tableaux de nombres ».

Application :

La production de littérature scientifique et technique des chercheurs d'une université est ventilée dans les quatre grandes disciplines de la façon suivante :

	Article	Livre	Brevet	*Total*
Sciences de la matière	13	2	5	20
Sciences de la vie	20	2	8	30
Sciences sociales	10	5	5	20
Sciences de l'ingénieur	7	1	22	30
Total	50	10	40	100

Tableau 2 – Production de littérature scientifique et technique

La lecture de ce tableau nous apprend que 50 % de la production scientifique des chercheurs est faite d'articles *(ligne 5, colonne 1)*. Si on applique ce pourcentage aux sciences de la matière, on constate que, sur une production totale de 20, il ne devrait y avoir que 10 articles :

$$\frac{20.50}{100} = 10$$

Or, il y en a 13. Les scientifiques des sciences de la matière produisent en priorité des articles. Donc, il n'y a pas indépendance entre la discipline et le type de production choisi : le respect de la proportion moyenne correspond à ce que l'on appelle la situation d'indépendance.

Grâce *à l'analyse factorielle des correspondances*, on obtient une représentation de la typologie de la production de littérature scientifique et technique selon les disciplines :

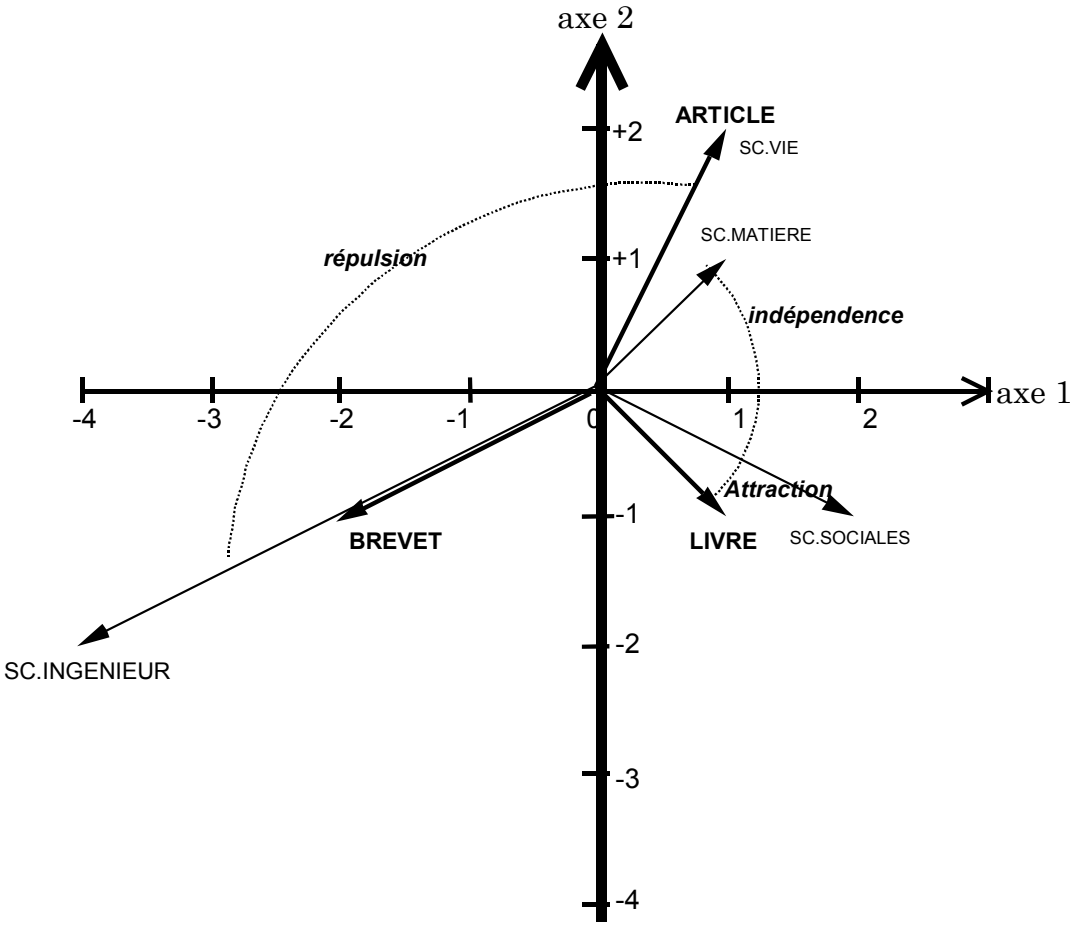

Figure 8 – La littérature scientifique et technique selon les disciplines

(Plan des axes 1 et 2)

- **CONCLUSION**

Alors que les techniques qui produisent de l'information, les techniques qui la mémorisent, les techniques qui la véhiculent frôlent tous les jours les limites de l'infiniment petit et les limites de l'infiniment grand,

mathématique et statistique nous permettent d'explorer plus facilement ces univers inconnus de l'information. Elles nous aident aussi à mieux saisir cette information devenue de ce fait infiniment croissante, infiniment rapide et infiniment complexe pour mieux maîtriser sa production, sa communication et son usage. Les exemples que nous avons présentés ne sont qu'un premier pas dans le sens d'un engagement plus profond de l'outil mathématique en Science de l'information.

Les développements actuels des activités scientifiques, techniques et industrielles dans les différents secteurs de l'information et de la culture laissent présager un usage plus intensif de cet outil mais aussi, il faut le souhaiter, la découverte de nouvelles méthodes, de nouvelles lois et de nouvelles techniques mathématiques et statistiques encore mieux adaptées à l'objet information.

A côté des diverses cultures qu'elle incorpore jusqu'à maintenant, la Science de l'information ajoute une culture que peu attendaient peut-être, la culture mathématique.

REFERENCES

BOUDON Raymond – L'analyse mathématique des faits sociaux – Plon, Paris,1970.

BORGMAN C.L. (ed.) – Scholarly communication and bibliometrics – Sage Publications, London, 1990.

BOYCE B.R., MEADOW C.T., KRAFT D.H. – Measurement in information science – Academic Press, San Diego, 1994

CALLON M., COURTIAL J.P., PENAN H. – La scientométrie – Que sais-je ?, PUF, Paris, 1993.

COLEMAN James S. – Introduction to mathematical sociology – The Free Press, New York, 1964

EGGHE L., ROUSSEAU R. - Introduction to informetrics : quantitative methods in library, documentation and information science, Elsevier, Amsterdam, 1990.

ELKANA Y. (ed.) – Towards a metric of science – John Wiley & sons, New York, 1978.

LAFOUGE T, LE COADIC Y.F., MICHEL C. – Eléments de Statistique et de Mathématique de l'information : infométrie, bibliométrie, médiamétrie, scientométrie, muséométrie, webométrie. – Les Presses de l'ENSSIB, Lyon, 2001.