

# Information et théorie mathématique: une impasse en science de l'information ?

## Le cas de l'infométrie.

**Thierry Lafouge**

Université Claude Bernard Lyon1 Laboratoire RECODOC  
Bâtiment OMEGA  
43, Boulevard du 11 novembre 1918  
69622 Villeurbanne Cedex  
Tel 04 72 44 58 34

[lafouge@enssib.fr](mailto:lafouge@enssib.fr)

### Résumé

La théorie statistique de l'information de C. Shannon, appelée souvent à tort théorie de l'information ou théorie mathématique de la communication, est souvent réduite et connue en SIC (Sciences de l'Information et de la Communication) au travers du schéma du système général de la communication : source, émetteur, signal...bruit.... La théorie de Shannon est connue en statistique par sa célèbre formule de l'entropie. La formule de Shannon est isomorphe à la formule de l'entropie de Boltzmann en physique. Cette théorie est importante car elle est à la jonction de la théorie du signal et de la statistique. Les mesures de l'entropie sont utilisées comme indicateurs en statistique unidimensionnelle et bidimensionnelle. Nous essaierons au travers de cet article de donner le point de vue de l'infométrie.

**Mot clef:** entropie/ théorie probabiliste de l'information/ statistique

### Abstract

Shannon's theory is commonly boarded in the narrow statement of the general communication scheme : signal, noise, ... The entropy formula in statistics is a characteristic of shannon's theory which is isomorphic to Boltzman formula in physic. It's an important issue in that way this theory is between signal and statistical theory. By using entropy measures in unidimensional and bidimensional statistics, we'll try to point out this issue in a infometric approach.

**Keywords :** entropy/ probabilistic theory of information/ statistic

## Introduction

Le mot information est utilisé dans des contextes très variés, dans des sens totalement différents suivant les disciplines scientifiques : on peut à titre d'exemples citer la thermodynamique avec le concept d'entropie, la physique avec la théorie du signal, la biologie avec la théorie du génome. Se pose alors la question, s'il est possible de construire une théorie de l'information, et si elle est unique. Notre démarche dans cet article vise non pas l'information en tant que telle, mais la quantité d'information. Lorsque l'on parle de quantité d'information et de mesure on pense à la notion de contenu ou de valeur de l'information. La science de l'information de par son objet doit se sentir concernée par ce questionnement. Si on définit l'infométrie comme l'ensemble des techniques de mesures mathématiques et statistiques de l'information, on souhaiterait avoir une définition suffisamment claire du concept de quantité d'information qui puisse nous amener à définir une mesure, c'est à dire un ensemble d'opérations parfaitement définies, nous amenant à des axiomes clairs et dont le résultat est un nombre. La synthèse que nous développons ici n'est pas si ambitieuse. De toute façon à l'heure actuelle, faute de connaissances, ou pire parce que on ne saurait vraiment pas formuler le problème une approche générale du concept de quantité d'information serait vouée à l'échec. Nous nous intéressons ici à la théorie probabiliste de l'information, connue sous le nom de théorie de Shannon, qui est la plus utilisée en science de l'information et de la communication. Ce travail qui à première vue peut paraître « risqué, prétentieux ou obsolète » en science de l'information, nous a semblé nécessaire au vue de prises de position souvent extrêmes de certains chercheurs :

- un rejet de cette théorie, souvent par ignorance et /ou par des présupposés épistémologiques : restriction de la théorie de Shannon au célèbre schéma émetteur, canal, récepteur par exemple,
- une utilisation abusive de cette théorie pour valider des résultats,
- une utilisation naïve de cette dernière.

Nous essaierons de donner au lecteur quelques repères pour lui donner l'envie d'approfondir cette théorie et de se forger sa propre opinion. Nous aborderons principalement dans cet article les relations multiples qu'entretiennent la théorie probabiliste de l'information (travaux d'Hartley, Shannon, Reyni..) et les statistiques en général. Nous n'apporterons pas de résultats théoriques nouveaux mais nous mettrons en parallèle différentes approches utilisant cette théorie et donnerons quelques exemples.

## 1. LA MESURE DE L'INFORMATION : de HARTLEY à SHANNON

### 1.1 Information d'un ensemble : la formule de Hartley en 1928

Etant donné un ensemble  $E$  de  $k$  éléments, où l'on suppose  $k = 2^n$  : si à chaque élément de  $E$  on associe un numéro écrit en base 2 qui permet de le coder, il est trivial de dire que  $n$  digit suffisent pour le repérer. Le nombre  $n$  est dit mesuré la quantité d'information nécessaire pour repérer un élément de  $E$ . On définit alors la quantité d'information de  $E$ , noté  $I(E)$  par la même valeur :

$$I(E) = \log_2 2^n = n .$$

Hartley en 1928 généralise la quantité d'information pour un ensemble  $E$  ayant un nombre quelconque d'éléments par:

$$I(E) = \log_2 (|E|)$$

où  $|E|$  désigne le nombre d'éléments de l'ensemble  $E$ .

Notation

Par la suite on notera  $\log$  au lieu de  $\log_2$  le logarithme en base 2,  $\ln$  le logarithme népérien,  $\text{Log}$  le logarithme lorsque l'on ne précise pas.

### Exemple

Soient les quatre chaînes de caractères, «islamiste, religieux, abcdefghi, xqzrfdk » : elles ont toutes la même quantité d'information, à savoir :  $\log 26^9 = 9 \log 26 = 42,3$  bit. Ici l'ensemble  $E$  est constitué de tous les arrangements possibles avec répétitions de 9 caractères choisis parmi les 26 lettres de l'alphabet soit  $26^9$  éléments; l'unité d'information est le bit, information élémentaire pour repérer les éléments d'un ensemble de cardinal 2. Cet exemple est significatif de ce qu'on appelle quantité d'information d'une chaîne : on prend uniquement en compte sa forme, réduite ici au nombre de caractères. Non seulement la signification est totalement absente mais en plus on ne tient pas compte par exemple d'informations statistiques sur les fréquences des caractères dans la langue, seul le nombre de symboles de l'alphabet est retenu. Si ce nombre de symboles est réduit à deux on retrouve les unités classiques en informatique.

### Remarques

Nous parlons de mesure de quantité d'information sans avoir défini avec précision ce qu'on appelle information. Ici  $n$  est le nombre de cases élémentaires, mémoires, pour coder l'information qui est le cardinal de l'ensemble  $E$ . L'exemple ci-dessus justifie en partie l'appellation théorie du codage utilisée pour désigner la théorie de Hartley et ses prolongements que l'on développera par la suite.

## **1.2 Information d'un événement : la formule de Wiener en 1948**

On se place ici dans le cadre de la théorie des probabilités qui repose sur la théorie des ensembles et de la mesure en mathématique. Parmi toutes les mesures d'information d'un événement  $O$ , dont la probabilité de réalisation est notée  $p = P(O)$ , comment mesurer l'information apportée par cet événement ? On parlera désormais d'entropie d'un événement (on justifiera par la suite cette appellation qui est selon nous abusive et trompeuse). On suppose que cette mesure ne dépend que de  $p$ . On note  $H$  cette fonction sur  $]0, 1]$ . Construire une théorie dans le cadre des probabilités nous amène à poser les propriétés (1) et (2) :

$$(1) H \text{ est non négative } \quad H \geq 0$$

*En effet la probabilité d'un événement est un nombre positif,*

$$(2) \quad H \text{ est additive } \quad H(pq) = H(p) + H(q) \quad p, q \in ]0, 1],$$

*En effet si deux événements A et B sont indépendants on a la relation :  $p(A \cap B) = p(A).p(B)$ .*

$$(3) \text{ Enfin pour des questions de normalisation (propriété non obligatoire) on suppose } H\left(\frac{1}{2}\right) = 1.$$

On montre que la fonction vérifiant ces trois axiomes (J. Aczel J., Z. Daroczy chapitre 1 ) est nécessairement de la forme :

$$H(p) = -\log(p) \text{ ou } H(O) = -\log(P(O))$$

La condition (2) signifie que l'entropie apportée par la conjonction de deux événements indépendants est la somme des entropies apportées par chaque événement. Comme on vient de le voir cette condition est naturelle<sup>1</sup>. La troisième condition qui n'est pas essentielle assigne l'unité d'information (appelé bit) à l'événement de probabilité  $\frac{1}{2}$  équivalent quant à sa mesure à son opposé .

### Exemple

Soit une distribution de descripteurs caractérisant un domaine scientifique. Les mots très fréquents ( $p$  voisin de 1) ont une entropie très faible ( $H$  voisin de 0), c'est ce qu'on appelle les mots triviaux, que connaît parfaitement l'expert du domaine. Les mots ayant une basse fréquence qui sont très nombreux ont une entropie très forte, on parle alors de bruit ou de marginalité. C'est parmi eux qu'on trouve ce qu'on appelle les signaux faibles en veille. Le problème bien connu dans les analyses bibliométriques de références bibliographiques est que ces mots sont très nombreux et ont des fréquences identiques très faibles.

---

<sup>1</sup> Naturelle par rapport à la notion d'indépendance. Rappelons que cette question d'indépendance n'a pas d'équivalent dans la théorie de la mesure en mathématique comme en ont les notions d'espérance, de variable aléatoire.....

### 1.3 Information d'une suite d'évènements : la formule de Shannon en 1948

#### 1.3.1 Les différentes approches

L'entropie de Shannon, notée  $H$  ou  $H_n$ , mesure la quantité d'information moyenne, elle peut être introduite de plusieurs façons, elle généralise les mesures précédentes.

- Le point de vue ensembliste : information apportée par un caractère

Soit  $E_i$   $i \in I$ , une partition de l'ensemble  $E$  par le caractère  $I$  où l'on note

$$p_i = \frac{|E_i|}{|E|}, \text{ on montre facilement (M. Volle) en se servant de la formule de Hartley que la connaissance}$$

de  $I$  permet d'économiser pour repérer les éléments de  $E$  la quantité d'information suivante :

$$H(I) = - \sum_i p_i \cdot \log(p_i)$$

$H(I)$  est aussi appelée entropie de la partition de  $E$  définie par  $I$ .

On remarquera que  $H(I)$  ne dépend pas du caractère  $I$ , ni même du type de ces modalités, mais uniquement de la distribution des fréquences  $p_i$ .

- La démarche de Shannon

Voici rapidement les hypothèses que formule Shannon. (W. Weaver, CE. Shannon) Supposons que nous ayons un ensemble  $n$  d'événements possibles dont les probabilités d'occurrence sont  $p_1, p_2 \dots p_i \dots p_n$ . Comment trouver une mesure de l'incertitude du résultat, c'est à dire du nombre de choix possibles? Les probabilités sont connues a priori et c'est tout ce que nous connaissons sur le futur. Si tous les événements sont équiprobables il est raisonnable de considérer qu'il est souhaitable que l'incertitude soit maximum. Shannon impose à cette mesure  $H$  trois conditions:

- $H$  est une fonction continue des  $p_i$ ,

-si tous les  $p_i$  sont égaux alors  $H$  est une fonction monotone croissante de  $n$ ,

-si un choix se décompose en deux choix successifs le  $H$  original devra être la somme pondérée des valeurs individuelles (Cette propriété est appelée par la suite récursivité ou "branching process"). Shannon montre que la seule fonction  $H$  satisfaisant aux trois hypothèses ci dessus est de la forme :

$$H = k \cdot \sum_{i=1}^n p_i \cdot \text{Log}(p_i) \text{ où } k \text{ est une constante dépendant des unités.}$$

Donnons une forme non normalisée: soit  $p_i$  une suite de  $n$  nombres positifs ou nul quantifiant la probabilité

de  $n$  évènements et vérifiant la relation :  $\sum_{i=1}^n p_i \leq 1$  l'entropie de cette suite est définie par la quantité :

$$H_n(p_1, p_2 \dots p_i, \dots p_n) = - \sum_{i=1}^n p_i \cdot \text{Log}(p_i) / \sum_{i=1}^n p_i$$

Si la suite définit une distribution de probabilité on retrouve la formule bien connue de Shannon qui est à l'origine, rappelons le, une théorie élaborée en vue de modéliser la transmission des signaux électriques.

L'approche de Shannon généralise celle de Hartley et de Wiener. Si la suite se réduit à un seul élément on retrouve la formule précédente de Wiener.  $H(p) = -\log(p)$

Enfin si  $p_i = \frac{1}{n}$ , c'est à dire si tous les évènements sont équiprobables on obtient la formule précédente de

Hartley  $I(E) = \log(|E|) = \log(n) = H_n(\frac{1}{n}, \frac{1}{n}, \dots, \frac{1}{n})$ . On montre que  $\log(n)$  est la valeur maximum

de l'entropie : on définit alors la quantité d'information relative  $H_r$ , variant entre 0 et 1 qui est le rapport de l'entropie sur l'entropie maximum.

$$H_r = \frac{H_n}{\log(n)}$$

- Le point de vue informationnel pragmatique

Si l'on veut définir une fonction (notée aussi  $H$ ) qui mesure l'entropie d'une suite d'évènements on peut ( Un système de production bibliographique (Egghe 1990) est un ensemble de sources qui produisent des items) utiliser le formalisme de la bibliométrie distributionnelle (Lafouge 1991).

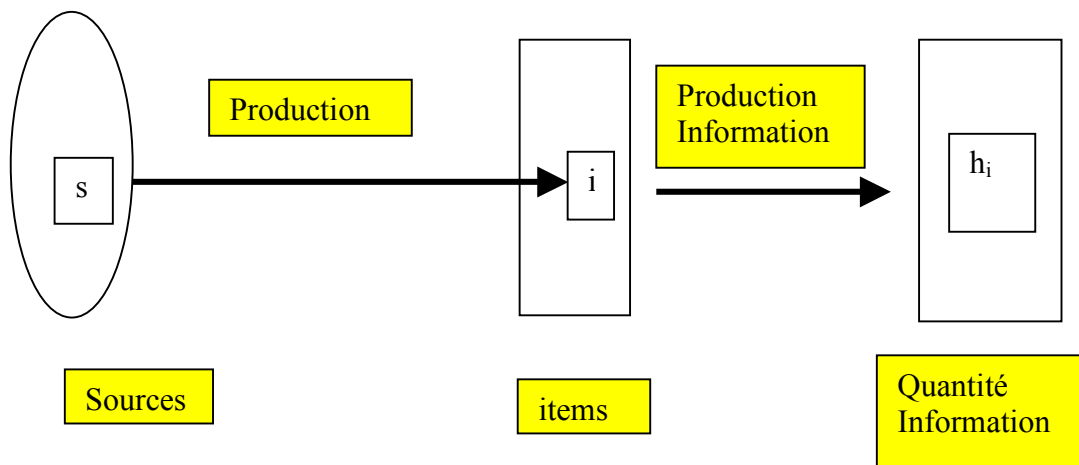
- les chercheurs produisent des articles,
- les formes graphiques<sup>2</sup> d'un texte produisent des occurrences,
- les mots clefs de références bibliographiques produisent des occurrences,
- .....

On écrit  $H$  sous la forme  $H = \sum_i h_i p_i$  où  $p_i$  est la probabilité qu'un événement élémentaire  $i$  se produise ( un chercheur produit  $i$  articles.....) et  $h_i$  l'information apportée par cet événement  $i$ . On définit alors  $h_i$  par  $h_i = -\log(p_i)$  (On s'inscrit dans les hypothèses de Wiener)

$H$  est l'espérance de la variable aléatoire  $i \rightarrow \log(p_i)$ . Le schéma ci-dessous résume le problème.

---

<sup>2</sup> Ensemble de caractères délimités par un séparateur.



$$H = \sum_i h_i \log(p_i)$$

$H$  mesure une quantité moyenne d'information. Nous obtenons le même résultat que Shannon. Ceci n'est pas un hasard car nous nous situons dans le cadre de la théorie des probabilités. Alors que dans la théorie de Shannon on parle de transmission de quantité d'information, ici on parle de production de quantité d'information.

### Exemple

Les phénomènes précédents de production bibliométrique peuvent être décrits suivant différentes formes. Nous utilisons la représentation générale des fonctions zipfiennes étudiées par Haitum (S. D Haïtum) où une telle distribution est définie par la fonction de densité hyperbolique suivante :

$$v(t) = \frac{C}{t^{\alpha+1}} \text{ où } C \text{ est une constante, } \alpha \text{ un nombre positif, et où } t \in [1 \ \infty]$$

Il est possible de généraliser la définition de l'entropie lorsque on a une distribution continue :

$$H(v) = - \int v(t) \cdot \text{Log}(v(t)) dt$$

Les entropies des distributions continues héritent de la plupart des propriétés du cas discret défini précédemment. Si on calcule l'entropie d'une zipfienne on a :

$$H(\alpha) = - \text{Log}(\alpha) + \frac{1}{\alpha} + 1 . \text{ Il est aisé de montrer que l'entropie est une fonction décroissante de } \alpha . \text{ On}$$

retrouve l'interprétation classique de la loi de Lotka qui stipule que plus  $\alpha$  est élevé, plus grand est le fossé entre le petit nombre de chercheurs qui produisent beaucoup et le grand nombre de chercheurs qui produisent très peu, et donc plus grand est la quantité d'information. Yablonsky (A. L Yablonsky) montre qu'il existe un lien entre ce type de distribution et le principe du maximum d'entropie (notée MEP), lui même lié au principe de la loi du moindre effort (PLE), résultat que nous avons prolongés (T. Lafouge, C. Michel) pour le cas de la distribution binomiale négative. Nous pensons que cette voie de recherche encore peu explorée peut être féconde.

- Le rêve idéaliste : entropie et information

Soit un langage constitué de  $n$  symboles chacun ayant une fréquence  $N_i$  le nombre total  $W$  de messages possibles de  $N$  symboles respectant les fréquences précédentes (c'est à dire l'égalité :  $N = \sum_{i=1}^n N_i$ ) est :

$$W = \frac{N!}{N_1! \cdot N_2! \cdot N_3! \cdot \dots \cdot N_n!} \text{ (formule de Brioullin)}$$

Si l'on pose  $p_i = \frac{N_i}{N}$   $i = 1..n$  et que l'on utilise la formule de Stirling pour calculer factoriel on

obtient :  $\frac{\log(W)}{N} = k \cdot H_n(p_1, \dots, p_i, \dots, p_n)$  : sous cette forme la mesure de Shannon est bien la quantité moyenne d'information apportée par un symbole. Son analogie avec la fonction entropie (Voir ci-dessous) de la thermodynamique est patente (D. Parrochia).

Les notions d'entropie et d'information découlent de la thermodynamique. C'est Sadi Carnot qui en formulant le premier principe va initier ces travaux. Enfin Boltzman en étudiant la mécanique va obtenir

une formule identique en calculant l'énergie d'un gaz comme la moyenne des énergies correspondantes . Si on désigne par  $S$  la fonction entropie, cette dernière peut s'écrire :

$S = K.Ln(\Omega(E))$  où  $\Omega(E)$  désigne le nombre d'états possibles d'un système ayant une énergie donnée  $E$  et où  $K$  est une constante . Tous ces résultats permettent de montrer qu'il existe une isomorphie entre l'entropie de Boltzmann et l'entropie de Shannon.

Il est alors tentant de postuler une équivalence entre énergie et information. L'exploitation qu'on peut faire d'une telle analogie montre maintenant ses limites. L'analogie entre entropie et information est séduisante mais inopérante pour notre discipline. Nous avons cependant gardé ici comme dans beaucoup d'ouvrages et d'articles le terme entropie.

Remarque

Lorsque que l'on travaille avec les fréquences il est souvent intéressant de mettre l'entropie de Shannon sous la forme :

$$H = \log(F) - \frac{1}{F} \cdot \sum_{i=1}^n f_i \cdot \log(f_i) \quad F = \sum_{i=1}^n f_i$$

Exemple

On parle en linguistique quantitative de l'entropie des lettres de l'alphabet d'une langue. Ainsi en français la fréquence moyenne de la lettre E est 0,175 ..... On définit alors l'entropie moyenne d'une lettre par  $H(p_1, p_2, \dots, p_{26})$  ; en langue française le calcul nous donne 3,98 bit. On peut alors étudier l'utilisation des caractères dans plusieurs langues et faire des comparaisons. Il n'en est pas de même pour les formes graphiques ou les mots d'une langue car on ne travaille pas dans un univers fermé.

1.3.2 Les propriétés de la mesure de Shannon

Nous rappelons sans les démontrer les principales propriétés algébriques de cette mesure :

(a) Symétrie  $H_n(p_1, \dots, p_i, \dots, p_n) = H_n(p_{k(1)}, \dots, p_{k(i)}, \dots, p_{k(n)})$  où  $k$  est une permutation arbitraire sur l'ensemble  $\{1 \dots n\}$

(b) Normalité  $H_2(\frac{1}{2}, \frac{1}{2}) = 1$

(c) Décision  $H_2(1, 0) = 0$

(d) Linéarité  
Soient deux distributions de probabilité :  $p_i \quad i = 1..n \quad q_j \quad j = 1..m$   
 $H_{nm}(p_1 \cdot q_1, \dots, p_i \cdot q_j, \dots, p_n \cdot q_1, p_n \cdot q_2, \dots, p_n \cdot q_m) = H_n(p_1 \cdot p_i \cdot p_n) + H_m(q_1 \cdot q_j \cdot q_m)$

(e) Linéarité forte  
Soient une distribution de probabilité  $p_i \quad i = 1..n$  et  $m$  distributions de probabilités :  $q_{jk} \quad k = 1..n \quad j = 1..m$   
 $H_{nm}(p_1 \cdot q_{11}, p_1 \cdot q_{12}, \dots, p_1 \cdot q_{1m}, \dots, p_i \cdot q_{i1}, p_i \cdot q_{i2}, \dots, p_i \cdot q_{im}, \dots, p_n \cdot q_{n1}, p_n \cdot q_{n2}, \dots, p_n \cdot q_{nm}) =$   
 $H_n(p_1, \dots, p_i, \dots, p_n) + \sum_{j=1}^m p_j \cdot H_n(q_{j1}, q_{j2}, \dots, q_{jn})$

(f) Récursivité  $H_n(p_1, \dots, p_i, \dots, p_n) = H_{n-1}(p_1 + p_2, p_3, \dots, p_n) + \frac{p_1}{p_1 + p_2} H_2(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$

Les propriétés (a) (b) (c) sont communes à toutes les mesures statistiques de l'information ( Voir les mesures de Reyni au paragraphe 1.4, et les mesures d'ordre supérieure dans la conclusion).

- La symétrie signifie que c'est une mesure globale d'un ensemble d'événements,
- la normalité traduit que l'incertitude est maximum lorsque tous les événements sont équiprobables. Elle croît en fonction du nombre d'événements.
- la propriété de décision signifie qu'il n'y a pas d'incertitude si un événement est sûr.
- la linéarité est l'équivalent de la propriété de Wiener pour des événements indépendants,
- la linéarité forte joue un rôle important : elle va nous permettre de donner un sens à la notion d'entropie conditionnelle ; elle est à la base de la construction d'une série d'indicateurs (cf. 3.1.1).

#### 1.4 Une autre approche : le gain d'information

A première vue la notion de gain d'information dans la vie courante semble plus intuitive que la notion de quantité d'information. Dans la théorie de l'information un gain d'information sera matérialisé par une quantité d'information.

Pour introduire cette notion on peut procéder comme pour le point de vue ensembliste précédent (Voir paragraphe 1.3.1) en utilisant les mêmes notations. On définit  $F$  un sous ensemble de  $E$  et on construit la

partition de  $F$  :  $F_i = F \cap E_i$  avec  $q_i = \frac{|F_i|}{|F|}$ . La question est alors : pour repérer un élément de  $E$ , qu'apporte le

fait de savoir que cet élément appartienne à  $F$ ? Cette quantité  $\Delta(J)$  est égale est égale à l'information pour repérer un élément suivant  $I$  si on ignore que l'élément est dans  $F$  moins l'information si l'on sait que l'élément est dans  $F$  ( Voir M. Volle p58 ).

$$\Delta(J) = \sum_i q_i \log\left(\frac{q_i}{p_i}\right) \text{ (Cette information est dite information de Kullback.)}$$

Le lecteur peut penser que nous jouons avec les mots (ou avec les notations en mathématiques) Il vérifiera cependant que cette quantité est différente de  $H_n(p) - H_n(q)$ , et donc que le gain d'information n'est pas une différence de quantité d'information, c'est bien un autre concept.

De la même façon soient deux caractères  $X$  et  $Y$  définis chacun par une distribution de probabilité,  $p_i$   $i = 1, n$  et  $q_j$   $j = 1, m$ , on définit  $\Delta(X, Y)$  le gain d'information lorsque l'on passe des distributions marginales de  $X$  et de  $Y$  à la distribution des fréquences des événements  $(i, j)$ . On montre alors que le gain d'information est égal à :

$$\Delta(X, Y) = \sum_{ij} p(i, j) \log \frac{p(i, j)}{p_i \cdot p_j}$$

On remarque que cette expression est symétrique en  $i$  et  $j$  et qu'elle est nulle si les deux distributions sont indépendantes.

Lorsque l'on élabore l'axiomatique de la théorie probabiliste de l'information deux approches sont possibles.

La première consiste à suivre les étapes suivantes :

Définition de la quantité d'information -> axiomatisation -> définition du gain d'information.

La deuxième :

Définition du gain d'information -> axiomatisation -> définition de la mesure d'information

Dans (A. Renyi) ce dernier développe la deuxième approche qui lui permet d'axiomatiser la notion de gain d'information et de construire dans un deuxième temps un ensemble de nouvelles mesures dont on trouvera la formule ci-dessous :

Soit une distribution de probabilité  $p_i$   $i = 1, n$  et un nombre  $\alpha \neq 1$  il pose :

$$H_\alpha = \frac{1}{1-\alpha} \text{Log} \left( \sum_{i=1}^n (p_i)^\alpha \right)$$

#### Attention

On vient de voir dans tout ce qui précède que l'entropie de Shannon est une mesure de l'incertitude. Au sens de Shannon, c'est à dire la théorie mathématique de la communication l'entropie mesure l'imprévisibilité moyenne que l'observateur a sur le message émis. Plus cette valeur est élevée plus l'incertitude augmente (en physique on dit que le désordre augmente), ce que nous traduisons en disant que l'information diminue. Quantité

*d'information et entropie varient en sens contraire. Si un événement est sûr, son entropie est nulle, par contre la quantité d'information est maximum. Les physiciens parlent de négentropie pour désigner la quantité d'information..*

## 2. STATISTIQUES ET MESURE de L'INFORMATION

L'objectif de ce paragraphe est de montrer les liaisons qui existent entre les statistiques classiques utilisées en infométrie et les mesures précédentes. Certains ouvrages de statistique ont un chapitre entier consacré à la théorie statistique de l'information (Voir par exemple M. Volle, A. Reyni dans la bibliographie ci dessous)

### 2.1 Statistique unidimensionnelle et mesures informationnelles

#### 2.1.1 Indicateurs de concentration et de diversité

Pour caractériser les distributions de fréquences les statistiques unidimensionnelles utilisent trois types de mesure que sont les indicateurs de tendance centrale, de dispersion et de concentration. Les distributions rencontrées en science de l'information sont connues sous le nom de Zipfienne et ont été largement étudiées dans la littérature (S. D Haitum). Elles sont en général de forme hyperbolique et décroissantes (cf 1.3.1) et possèdent une longue queue avec un écart type supérieur à la moyenne. En général on les oppose aux distributions gaussiennes que l'on rencontre fréquemment lorsque l'on étudie des populations physiques. Aussi les indicateurs classiques (moyenne, variance, coefficient de variation ) issues de la théorie des moments ne sont pas toujours adaptés pour résumer ces distributions. Les chercheurs en biologie, économie, infométrie ont développé de nombreux autres indicateurs.

Nous allons nous intéresser aux indicateurs de concentration ou de diversité. Donnons rapidement une présentation des indices de ce type d'indice. Soit un ensemble de  $n$  sources (ouvrages, auteurs, mots, ...) qui produisent des items (prêts, articles, occurrences...). Nous désignons par  $f_i$  le nombre d'items produits par la  $i^{\text{ème}}$  source ; un indice de concentration (respectivement de diversité) de cette distribution doit vérifier les trois propriétés :

$D(f_1, f_2, \dots, f_i, \dots, f_n) = D(f_{k(1)}, f_{k(2)}, \dots, f_{k(n)})$  : invariance par permutation, cela signifie que c'est bien un indice global.

$D(f_1, f_2, \dots, f_i, \dots, x_n) = D(af_1, af_2, \dots, af_i, \dots, af_n)$  : invariance si on modifie l'échelle de mesure.

Si ces deux propriétés sont classiques pour de nombreux indicateurs il n'en n'est pas de même pour celles qui vont suivre que L. Egghe a longuement étudié (L. Egghe, R. Rousseau 1990) et appelle principe de transfert si on a :

Pour tout  $f_i \leq f_j$  et  $0 < f \leq f_i$  :

un indice de concentration vérifie:

$$D(f_1, f_2, \dots, f_i, \dots, f_n) < D(f_1, f_2, \dots, f_i - f, \dots, f_j + f, \dots, f_n)$$

Cela signifie qu'un transfert d'une source pauvre vers une source riche augmente l'indice de concentration.

Un indice de diversité vérifie :

$$D(f_1, f_2, \dots, f_i, \dots, f_n) > D(f_1, f_2, \dots, f_i - f, \dots, f_j + f, \dots, f_n)$$

Cela signifie qu'un transfert d'une source pauvre vers une source riche diminue l'indice de diversité.

La mesure de Teil ( provenant de l'entropie qui est une mesure de diversité) est une « bonne mesure de concentration » au sens qu'elle vérifie toutes les propriétés souhaitées (L. Egghe, R. Rousseau) ; elle est souvent

écrite sous la forme :  $Th = \frac{1}{n} \sum_{i=1}^n \frac{f_i}{\mu} \cdot \log\left(\frac{f_i}{\mu}\right)$  où  $\mu = \frac{1}{n} \sum_{i=1}^n f_i$  on montre aisément que l'on a l'égalité :

$$H = Th - \text{Log}n$$

Pielou a généralisé ce type d'indice. Ce dernier plus connu sous le nom d'entropie généralisée a été développé par Reyni, dans un tout autre contexte (Voir paragraphe 1.4). On définit l'entropie généralisée d'ordre  $\alpha$  par :

$$H_\alpha = \frac{1}{1-\alpha} \text{Log} \left( \sum_{i=1}^n (p_i)^\alpha \right) \quad \alpha \neq 1 \text{ on montre que } H = \text{Lim} (\alpha \rightarrow 1) H_\alpha .$$

Hill propose une approche (M. Hill) qui va définir un ensemble de mesures de diversité :  $D_\alpha = \exp(H_\alpha)$  . Lorsque  $\alpha = 2$  on retrouve l'indice bien

$$\text{connu de Simson : } D_2 = \sum_{i=1}^n p_i^2 .$$

Il est intéressant de noter que deux approches différentes, une en statistique unidimensionnelle et l'autre liée à la notion de gain d'information aboutissent à des résultats identiques.

### 2.1.2 Variance et Entropie

Une des propriétés selon nous la plus intéressante de l'entropie pour notre discipline est le caractère agrégatif de cette mesure semblable à celui de la variance. Comme la variance l'entropie peut être découpée en fractions après un regroupement en classes. Une classe est un regroupement de sources, on peut citer :

- en scientométrie : un laboratoire regroupement des chercheurs,
- en bibliométrie : un journal regroupant des articles,
- en infométrie : un thème regroupant des mots spécifiques

On notera :

$n$	nombre de sources,
$m$	nombre de classes,
$f_i$	nombre d'items produit par la $i^{\text{ème}}$ source,
$F_k$	nombre d'items produits par la $k^{\text{ème}}$ classe,
$F$	nombre total d'items, $F = \sum_{i=1}^n f_i$
$H_k$	entropie de la $k^{\text{ème}}$ classe.

On a alors le résultat suivant :  $H = H_{\text{inter}} + H_{\text{intra}}$

Où comme pour la variance (voir variance interclasse, intraclasse) :

$$H_{\text{inter}} = \sum_{j=1}^m \frac{F_j}{F} \cdot H_j \qquad H_{\text{intra}} = - \sum_{j=1}^m \frac{F_j}{F} \log \left( \frac{F_j}{F} \right)$$

Cette propriété est intéressante car la quantité d'information ainsi décomposée donnera une vision de la quantité d'information (ou plutôt diversité) apportée par chaque classe et par leur différenciation au sein du groupe. Cette propriété est intéressante quand il est possible de regrouper un ensemble de codes, de mots (Voir Thésaurus, Plan de classement...), en unités hiérarchiquement supérieures. C'est dans cet esprit que nous avons préconisé l'emploi de nouveaux indicateurs pour résumer les distributions Zipfiennes (J. Lhen.), l'entropie de Shannon étant la diversité d'ordre 1.

## 3.1 Statistique bidimensionnelle et mesures informationnelles

La statistique unidimensionnelle résume avec des indicateurs une distribution de valeurs. La statistique bidimensionnelle a pour objectif de découvrir des liens éventuels qui existent entre deux variables.

### 3.1.1 Entropie conditionnelle et information mutuelle

Nous allons tirer les conséquences de la propriété de linéarité forte de la mesure de Shannon. Supposons deux évènements  $X$  et  $Y$  chacun étant défini par une suite de  $n$  (respectivement de  $m$ ) éventualités définis par une suite  $p_i$  (respectivement  $p_j$ ).

On note :

$p(i, j)$  probabilité d'avoir les évènements  $i$  et  $j$  simultanément,  
 $p(j/i)$  probabilité d'avoir  $j$  sachant  $i$  réalisé (probabilité conditionnelle),

$$p(j/i) = \frac{p(i, j)}{\sum_{i=1}^n p(i, j)}$$

On a  $H(X, Y)$  entropie conjointe de  $X$  et  $Y$  :

$$H(X, Y) = - \sum_{i, j} p(i, j) \cdot \log(p(i, j))$$

Si les évènements sont indépendants ( $p(i, j) = p_i \cdot p_j$ ) on obtient la relation :  
 $H(X, Y) = H(X) + H(Y)$  (Voir propriété (d) de linéarité)

On définit l'entropie conditionnelle de  $Y$  sachant  $X$  comme la moyenne de l'entropie  $Y$  pondérée pour chaque valeur de  $X$  pondérée.

$$H(Y/X) = \sum_{i, j} p(i, j) \cdot \log(p(j/i))$$

On montre alors facilement que les propriétés suivantes (Voir propriété de linéarité forte) :

$$H(X, Y) = H(Y/X) + H(X)$$

D'où on tire :  $H(X) + H(Y) \geq H(X, Y) = H(X) + H(Y/X)$

On en déduit le résultat : si  $X$  et  $Y$  dépendent l'un de l'autre alors la connaissance de  $X$  entraîne une diminution de l'entropie c'est à dire de l'imprévisibilité qu'on a sur  $Y$  :  $H(Y/X) \leq H(Y)$ . De même on montre :  $H(X/Y) \leq H(X)$ .

Ce résultat nous permet de construire les indicateurs souvent utilisés en théorie mathématique de la transmission du signal et par certains chercheurs en scientométrie que l'on verra par la suite lors d'un exemple :

Information sur  $X$  contenu dans  $Y$  :  $T(X, Y) = H(X) - H(X/Y)$

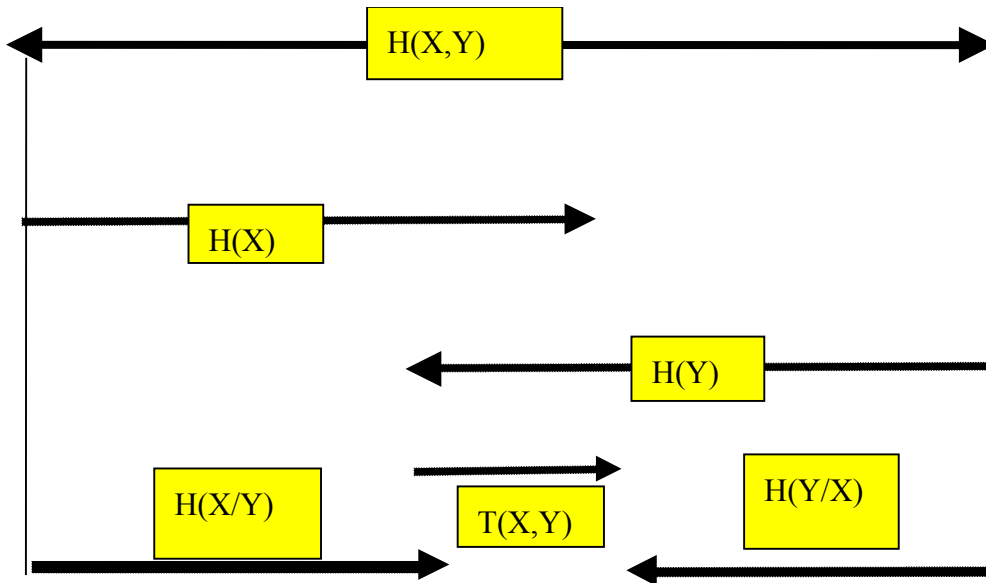
Information sur  $Y$  contenu dans  $X$  :  $T(Y, X) = H(Y) - H(Y/X)$

Ces indicateurs sont souvent appelés liens en statistique, ils ne sont pas symétriques.

Comme pour l'entropie relative on utilise de préférence des indicateurs quantifiant les informations relatives :

$$\frac{T(X, Y)}{H(X)} ; \frac{T(Y, X)}{H(Y)}$$

On résume traditionnellement toutes ces propriétés dans le schéma ci-dessous reproduit dans des contextes différents. On va développer un exemple en s'inspirant des travaux de Loet Leydesdorf (L. Leydesdorff) qui fait partie des chercheurs utilisant cette théorie dans ses nombreux travaux.



Exemple

Soit un ensemble  $A$  de  $m$  articles concernant un thème scientifique. Chaque article est caractérisé par un ensemble  $M$  de  $n$  de mots (mots clés, mots du résumé, du titre....) caractérisant son contenu. La fréquence de chaque mot (appelée également nombre d'occurrences) est connue. Soit  $f_{ij}$  la fréquence du mot  $i$  dans l'article  $j$  on alors un tableau croisé (Articles \* Mots) de  $n*m$  valeurs qu'on va traiter dans un premier temps classiquement :

$$N = \sum_{i,j} f_{ij} \text{ nombre total d'occurrences dans les } m \text{ articles,}$$

$$p(i, j) = \frac{f_{ij}}{N} \text{ probabilité d'occurrence du mot } i \text{ dans l'article } j,$$

$$a_j = \frac{\sum_{i=1}^n f_{ij}}{N} \text{ probabilité d'occurrence d'un mot dans l'article } j$$

$$m_i = \frac{\sum_{j=1}^m f_{ij}}{N} \text{ probabilité d'occurrence du mot } i \text{ dans les articles.}$$

On utilise la formalisation précédente. On calcule en premier lieu les entropies des trois distributions :

$$H(A) = - \sum_{j=1}^m a_j \cdot \log(a_j) \text{ distribution des articles au travers des mots,}$$

$$H(M) = - \sum_{i=1}^n m_i \cdot \log(m_i) \text{ distributions des mots au travers des articles,}$$

$$H(A, M) = - \sum_{i,j} p(i, j) \cdot \log(p_{ij}) \text{ distribution conjointe.}$$

On en déduit les entropies conditionnelles :

$$H(A/M) = H(A, M) - H(M) \qquad H(M/A) = H(A, M) - H(A)$$

Puis ce qu'on a appelé les informations mutuelles

$$T(A,M)=H(A)-H(A/M)$$

$$T(M,A)=H(M)-H(M/A)$$

On construit alors les indicateurs (compris entre 0 et 1) qui vont résumer l'information de ces mots distribués dans les articles :

$Hr(A,M)=\frac{H(A,M)}{\log(n.m)}$  : pourcentage d'entropie de la distribution des valeurs du tableau par rapport à l'entropie maximale .

$Tr(A,M)=\frac{T(A,M)}{H(A)}$  : réduction de l'entropie concernant la distribution des mots dans les articles.

$Tr(M,A)=\frac{T(M,A)}{H(M)}$  : réduction de l'entropie concernant la distribution des articles à travers les mots.

C'est en partie en utilisant ces trois indicateurs que L. Leydesdorff analyse les résultats d'une étude menée par l'Ecole des Mines qui est à l'origine de la méthode des mots associés. Le lecteur trouvera en bibliographie la référence de cet article ( L. Leydesdorff(1992)) ainsi que la réponse de J. P Courtial un des nombreux chercheurs à l'origine de cette méthode. (J. P Courtial). Leydesdorff calcule les paramètres ci-dessous sur des données avant classification et après classification (après formation de ce qu'on appelle généralement des agrégats ); dans ce cas il fait les mêmes calculs sur ces données en regroupant les mots de chaque agrégats. Il obtient un tableau Articles\* Agrégat qu'il est obligé de compléter par les mots de fréquence faible qui ont été éliminés lors de classification. Cette étude est intéressante car elle manipule les mêmes données et permet de faire des comparaisons. Cette démarche est malheureusement bien rare dans notre discipline. Sans vouloir donner raison à l'un ou à l'autre nous voulons faire une ou deux remarques sur l'utilisation de ces indicateurs comme outil d'évaluation de la méthode des mots associés.

Les mesures faite avec l'entropie sont des moyennes. Dans la pratique l'écart type calculé sur la distribution informationnelle (cf 1.3.1 approche informationnelle pragmatique ) est souvent supérieur à l'entropie et donc relativise les conclusions car il existe une dispersion tres forte autour de la moyenne. Nous avons conscience qu'il faudrait développer cette affirmation d'un point de vue théorique. Cette étude a été entreprise.

La méthode des mots associés utilise un indice de proximité appelé coefficient d'équivalence qui est la mesure du cosinus de Salton au carré. L'évaluation de la méthode des mots associés est faite avec l'entropie d'ordre 1 qui est une mesure lié à la distance du khi2. Pourquoi tester uniquement l'ordre 1 ?

### 3.1.2 Distance du $\chi^2$ et mesure du gain d'information

Explicitons les relations entre la distance du  $\chi^2$  et de la mesure du gain d'information vue au paragraphe 1.4.

Rappel : la métrique du  $\chi^2$

Si l'on considère les trois distributions  $p, q, r$  ( $p_i, q_i, r_i, i = 1, n$ ) la distance entre  $p$  et  $q$ , calculée avec la métrique du  $\chi^2$  centrée sur  $r$  par :

$$D^2_r(p,q)=\sum_{i=1}^n \frac{(p_i - q_i)^2}{r_i}$$

L'utilisation de cette métrique est justifiée par le test du  $\chi^2$  (Voir la loi de probabilité du  $\chi^2$ ) et ses propriétés qui en font une distance adaptée à de nombreux problèmes en Statistique lorsque l'on croise deux variables nominales ou que l'on veuille ajuster une distribution par exemple. Elle est également utilisée en analyse factorielle où la distance entre les points est celle du  $\chi^2$  centrée sur le centre du

nuage de points.. Nous allons voir que cette distance est liée avec la notion d'entropie. Nous utilisons par la suite les notations du paragraphe 3.1.1 . On définit le *Lien* entre deux caractères par :

- $p(i, j)$  distribution du caractère conjoint  $(X, Y)$ ,
- $p_i \cdot p_j$  distribution des caractères conjoints  $(X, Y)$  supposés indépendants,

$$Lien(X, Y) = \sum_{i, j} \frac{(p(i, j) - p_i \cdot p_j)^2}{p_i \cdot p_j}$$

D'après ce qui précède, le *Lien* mesure donc la distance au sens du  $\chi^2$  entre les deux distributions ci dessus. Ce dernier mesure la dépendance entre les deux caractères, plus la valeur sera élevée, plus les caractères seront dépendants. Si  $X$  et  $Y$  sont indépendants le *Lien* entre  $X$  et  $Y$  est nul. Cette mesure symétrique est équivalente à une constante près au gain d'information mutuelle  $\Delta(X, Y)$  défini au paragraphe 1.4 (Voir M. Volle p 65)

$$\text{on a } \Delta(X, Y) \approx k \cdot Lien(X, Y)$$

Cette approximation est valable uniquement pour des faibles valeurs du Lien c'est à dire si les caractères ne sont pas « trop indépendants ».

On a des résultats identiques entre la mesure du cosinus distributionnel de Salton et la mesure du *Lien* de Reyni d'ordre  $\frac{1}{2}$ .

### 3. CONCLUSION

Toutes ces mesures de quantité d'information sont basées sur la notion d'évènement et sont des mesures de type probabiliste qui sont construites suivant le schéma mental suivant :

à un évènement on fait correspondre sa probabilité  $p$ , puis un nombre qui est fonction de  $p$  caractérisant une mesure de sa quantité d'information. Nous avons choisi cette approche pour introduire la mesure de Shannon de façon pragmatique. Cette dernière est mathématisée dans (J. Aczel J., Z. Daroczy chapitre 3) où on définit ce qu'on appelle une fonction d'information (fonction de Shannon), puis une fonction d'information d'ordre  $\alpha$ . Cette dernière permet de définir les mesures d'information d'ordre supérieur. Pour  $\alpha \neq 1$  on pose :

$$H_n^\alpha(p_1, p_2, \dots, p_n) = \frac{1}{2^{1-\alpha} - 1} \cdot (\sum_{k=1}^n p_k^\alpha - 1)$$

Ces mesures qui sont des mesures de type entropie généralisent celle de Shannon. Comme pour la mesure de Reyni on montre le résultat suivant :  $H_n = \lim_{\alpha \rightarrow 1} H_n^\alpha$ . Cet ensemble de mesures au vu des propriétés mathématiques (J. Aczel J., Z. Daroczy chapitre 6) (Voir linéarité, linéarité forte, récursivité) est vraiment la généralisation de la mesure de Shannon... Nous ne connaissons pas d'utilisation de ces mesures en infométrie, encore faudrait t-il donner des raisons d'utiliser ces mesures entropiques généralisées et savoir donner une interprétation au coefficient  $\alpha$ . Les deux approches, celle de Reyni et celle de Shannon ne sont pas indépendantes, il existe une relation mathématique entre ces deux formes.

Nous espérons avoir ouvert des directions de Recherche qui peuvent encore selon nous révéler des résultats nous permettant de mieux comprendre et utiliser les mathématiques mesurant les quantités d'information.

### Remerciements

Je remercie Sylvie Lainé Cruzel , maitre de conférence de Recodoc, qui a bien voulu critiqué ce travail.

### Références bibliographiques

J. Aczel Z. Daroczy

On Measures of Information and their Characterisations. Mathematics in Science and Engineering Vol 115.

J. P. Courtial (1992)

Comments on Leydesdorf's a validation study of Leximappe. Scientometrics 25 (1992)  
p 313-316.

L. Egghe, R. Rousseau

Sensitivity Aspects of inequality measures

(preprint)

L. Egghe (1990)

*The duality of informetric systems with applications to the empirical law.*

Journal of Information Science, Vol 16, 1990, p 17-27

L. Egghe

Development of hierarchy theory for digraphs using concentration theory based on a new type of Lorentz curve

(preprint)

L. Egghe R. Rousseau (1990),

Introduction to Informetrics.

Quantitative Methods in Library, Documentation and Information Science.

Elsevier, 1990, 450 pages, Amsterdam.

S. D. Haitum (1982)

*Stationary Scientometric Distributions.*

Scientometrics n°4, 1982, Part I p.5-25, Part II p.89-104, Part III

p.181-194.

M. O. Hill (1973)

*Diversity and Evenness: a unifying notation and its consequences.*

Ecology, 1973, Vol 54, N°2, p. 427-433.

T. Lafouge, C. Michel (2001).

*Links between information construction and information gain. Entropy and bibliometric distributions.* Journal of

Information Science, 27 (1) 2001, p 39-49.

T. Lafouge, L. Quoniam (1991).

*Les distributions bibliométriques.*

Revue française de bibliométrie N° 9, 1991, p. 128-138.

L. Leydesdorff (2001)

The challenge of Scientometrics . The developpment, Measurement, and Self-Organisation of Scientific Communications

Published by Universal Publishers

<http://www.upublish.com/books/leydesdorff-sci.htm>

L. Leydesdorff(1992)

*A validation study of Leximappe Scientometrics 25 (1992) p295-312.*

J. Lhen, T. Lafouge, Y. Zilsken, L. Quoniam, H. Dou (1995)

*La " Statistique" des lois de Zipf.*

Revue Française de Bibliométrie N°14, 1995, p. 135-146.

D. Parrochia (1994)

Cosmologie de l'Information

Chapitre 1, 2, 3.

Hermès 282 pages

A. Renyi (1966)

Calcul des probabilités, Edition Jacques Gabay 1992,

618 pages

M. Volle (1985)

Analyse des données Collection "Economie et Statistiques avancées ", Economica.323 pages.

W.Weaver, C.E Shanon (1975)

Théorie mathématique de la Communication.

Les classiques de sciences humaines La bibliothèque du CEPL 188 pages.

A. L. Yablonsky (1980)  
On fundamental regularities of the distribution of scientific productivity.  
Scientometrics 2(1) 1980 p3-34.