île Rousse 2005
Journée sur les systèmes d'information élaborée

FORMULATION MATHEMATIQUE PRAGMATIQUE DE LA LOI DU MOINDRE EFFORT - PROCESSUS EXPONENTIEL INFOMETRIQUE

Thierry Lafouge
Lafouge@univ-lyon1.fr

Laboratoire Ursidoc Université Claude Bernard Lyon1
43 Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France

Les régularités statistiques observées lors de la production ou de l'usage de l'information sont connues et étudiées depuis longtemps. Elles sont à nouveau d'actualité aujourd'hui sur Internet comme en témoignent de nombreux travaux. Ces dernières sont présentes historiquement dans trois activités informationnelles qui intéressent nos champs d'application.
En scientométrie, dans l'évaluation de la production des chercheurs, "Loi de Lotka", en bibliométrie pour étudier la dispersion des articles dans les périodiques scientifiques, "Loi de Bradford", enfin en infométrie pour les fréquences des mots dans un texte : "Loi de Zipf ".

Elles se matérialisent lors d'études du trafic sur Internet par des phénomènes d'invariance d'échelle (P. Aby... 2004). On les trouve aussi lorsqu'on comptabilise les fréquences du nombre de pages ou du nombre de degrés entrants ou sortants sur les pages web d'une collection de sites. (C. Prime et... 2005).

La formulation la plus courante est celle de Lotka:

$$v(t) = \alpha . \frac{1}{t^{\alpha+1}} \qquad \alpha > 0 \qquad t \in [1 \quad \infty]$$

Les propriétés de ces distributions hyperboliques ou lois puissance inverse ("Inverse power law") ou distributions Zipfiennes ont été largement étudiées dans différentes disciplines. Elles s'opposent à ce que l'on appelle les distributions gaussiennes.

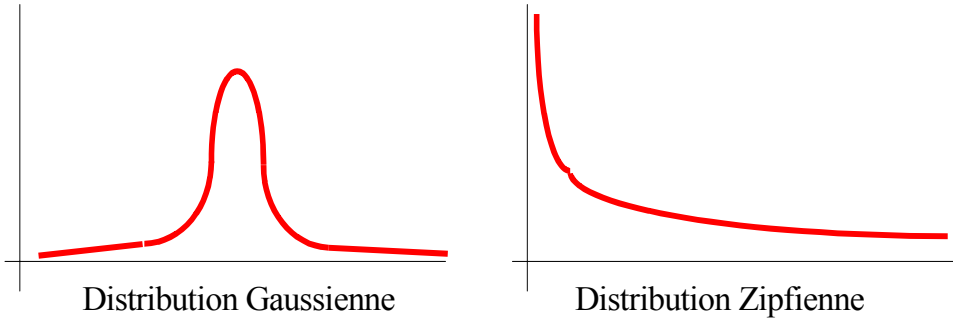


Figure 1. Deux types de distribution statistique

Tout d'abord elles ne sont pas symétriques, d'autre part elles ont une très longue queue et convergent vers 0 très lentement. C'est cette dernière propriété qui nous intéresse ici.

Différentes explications ont été avancées pour expliquer ce phénomène. Les plus connues en infométrie sont celle de Price et de Naranan. En 1972, Price propose un modèle probabiliste, connu sous le nom de loi des avantages cumulés pour expliquer les caractéristiques de ces distributions. Pour cela, il utilise le modèle de l'urne de Polya; cette loi peut s'énoncer ainsi: plus une source produit des

items, plus grande est sa chance d'en produire. De nombreux travaux théoriques ont montré par la suite l'équivalence sous certaines conditions des différents phénomènes décrits précédemment.
Si Price montre que son modèle offre un cadre d'interprétation probabiliste pour les différentes lois, le lien entre les comportements sociaux et la description statistique n'est pas toujours très clair. Le fait par exemple, de ne pas publier un article à un moment précis dans un domaine, ne peut être considéré comme un échec mais plutôt comme un non évènement. Il faut cependant modérer cette critique car les modèles prédictifs expliquent rarement les causes.

Une deuxième explication s'appuie sur les hypothèses de croissance exponentielle des sources (nombre de journaux) et du nombre d'items dans chaque source (nombre d'articles). Naravan utilise l'exemple de Bradford pour démontrer son résultat, mais les arguments restent universels. Un des intérêts de cette approche est le lien avec la théorie fractale (Egghe 2005).
Nous proposons dans cet article, non pas une nouvelle explication mais une lecture différente de la formule mathématique précédente en introduisant la notion de fonction d'effort. Cette dernière se trouve justifiée par un résultat mathématique obtenu lorsque qu'on passe à la limite.

1. Système d'information bibliographique généralisé

L'étude des régularités statistiques observées dans la production ou l'utilisation d'information est souvent représentée avec ce que nous appelons un « système d'information bibliographique généralisé » (Voir figure 1)

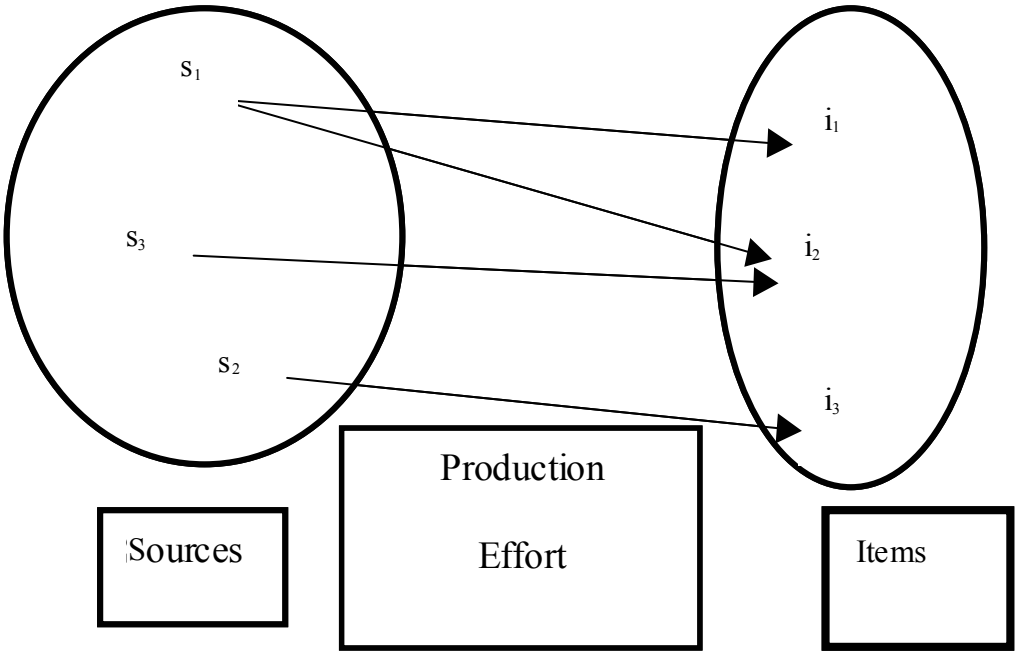


Figure 2 : Représentation schématique d'un système d'information bibliographique généralisé

La définition d'un « système d'information bibliographique généralisé » suppose l'existence d'un ensemble de sources (chercheurs, revues scientifiques, mots…), d'une fonction de production et de l'ensemble de tous les items produits (articles, occurrence) . Cette définition est très large. On suppose que la production, l'usage, de chaque item nécessite une certaine quantité d'effort. Cette dernière est définie grâce à une fonction nommée par la suite fonction d'effort.

2. Processus exponentiel infométrique

Dans cette communication nous allons présenter mathématiquement ces lois bibliométriques en utilisant la formulation (fréquence, effectif) de Lotka, selon le mode continu. Nous définissons ce que nous appelons un processus bibliométrique exponentiel, à l'aide d'une fonction dite fonction d'effort.
Soit f une fonction positive de $[1..\infty]$ dans $\Re^+$. (ou du moins croissante à partir d'une valeur positive supérieure à 1) croissante, non bornée, et a un nombre supérieur à 1, nous appelons processus bibliométrique exponentiel la distribution statistique $v(f,a)$ suivante :

$v(a,f)(t)=k\,a^{-f(t)}$ (k constante de normalisation) où la fonction f appelée fonction d'effort vérifie la condition : $\int_1^{\infty} f(t).a^{-f(t)}dt < \infty$

Cette dernière condition impose que la "Quantité d'effort" nécessaire pour produire ce processus ou ce qui équivalent, l'entropie, est finie.

Propriétés d'un processus bibliométrique exponentiel

On montre (T. Lafouge et… 2005) qu'un tel processus vérifie les propriétés suivantes.

$v(f,a)$ est décroissante,

$v(f,a)$ vérifie simultanément le principe du maximum d'entropie (PB. Kantor et.. 1998) et le principe du moindre effort. Ce résultat est implicitement démontré dans les textes de la théorie de Schannon (C. Shannon 1993).
Cette dernière propriété vient du fait que notre distribution est définie avec une fonction f qui joue le rôle de « fonction d'effort ». En effet la quantité d'effort produite par un processus v *est* $EF=\int v.(t).f(t).dt$ Pour démontrer cette dernière propriété il est nécessaire que *a* soit supérieur à 1, la croissance de la fonction f n'est par contre pas nécessaire.

L'entropie (*EH*) et la quantité d'effort (*EF*) sont liées par la relation linéaire suivante

$$EH(v(f,a))=-\log(k) + \log(a).EF(v(f,a))$$

3. Caractérisation d'un processus exponentiel infométrique

On peut définir un processus exponentiel infométrique par sa seule fonction d'effort (on choisit *a* égal à *e*). La croissance de la loi d'effort caractérise alors notre processus exponentiel.

Soit α un nombre positif on définit 3 types de distribution à l'aide de leur fonction d'effort. Ici on a normalisé la fonction d'effort par : $f(1)=0$

Effort parabolique : distribution de type gaussien : $v(t)=k.e^{-\alpha(t-1)^2}$

Effort linéaire : distribution de type exponentiel : $v(t)=k.e^{-\alpha(t-1)}$

Effort logarithmique : distribution de type Zipfienne : $v(t)=e^{-(\alpha+1)Log(t)}$

Ces trois fonctions bien connues sont représentées ci dessous.

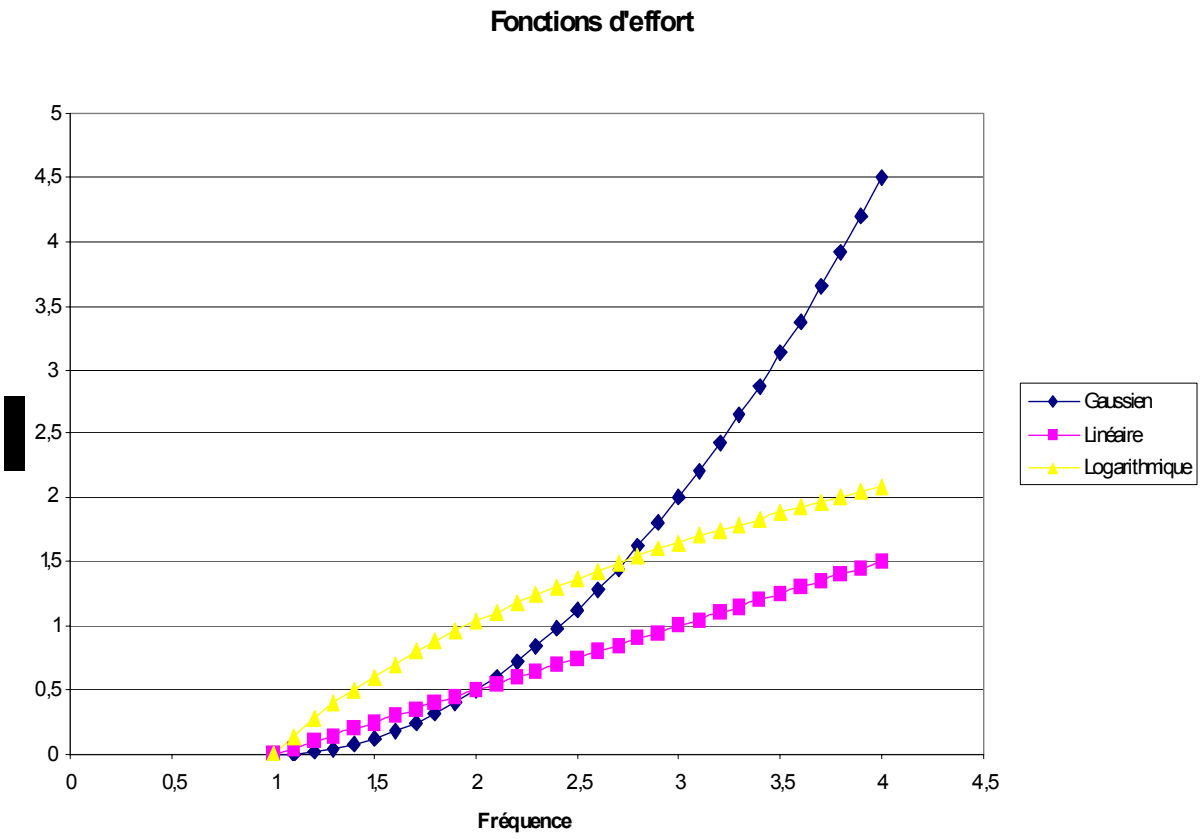


Figure 3 : Fonctions d'effort avec $\alpha = 0.5$

Le cas Gaussien correspond à une fonction d'effort qui augmente de plus en plus vite à l'opposé du cas Zipfien qui montre une fonction d'effort qui augmente de moins en moins vite d'où le nom de loi du moindre effort. Le cas exponentiel correspond au cas neutre, où la production d'effort est constante. On peut visualiser cette fonction d'effort en observant la vitesse de décroissance des trois distributions correspondantes (Voir figure 4).
La Gaussienne s'aplatit très vite au contraire de la distribution Zipfienne qui a une très longue queue, la distribution exponentielle se situant au centre.

Il est nécessaire de caractériser à l'aide des fonctions d'effort un processus exponentiel infométrique . Tout d'abord rappelons un résultat bien connu en infométrie et très facile à démontrer.
Si β est strictement supérieur à 1 la fonction d'effort, $f(t)=\beta.Ln(t)$ définit un processus exponentiel infométrique ayant pour quantité d'effort moyen $\frac{\beta}{(\beta-1)^2}$
Dans le cas contraire on n'a pas de processus. Cette fonction apparaît comme limite. On a le résultat suivant.

Résultat

Soit g une fonction d'effort quelconque, si on étudie son comportement par rapport à la fonction d'effort correspondant à la loi du moindre effort, on a le résultat suivant :

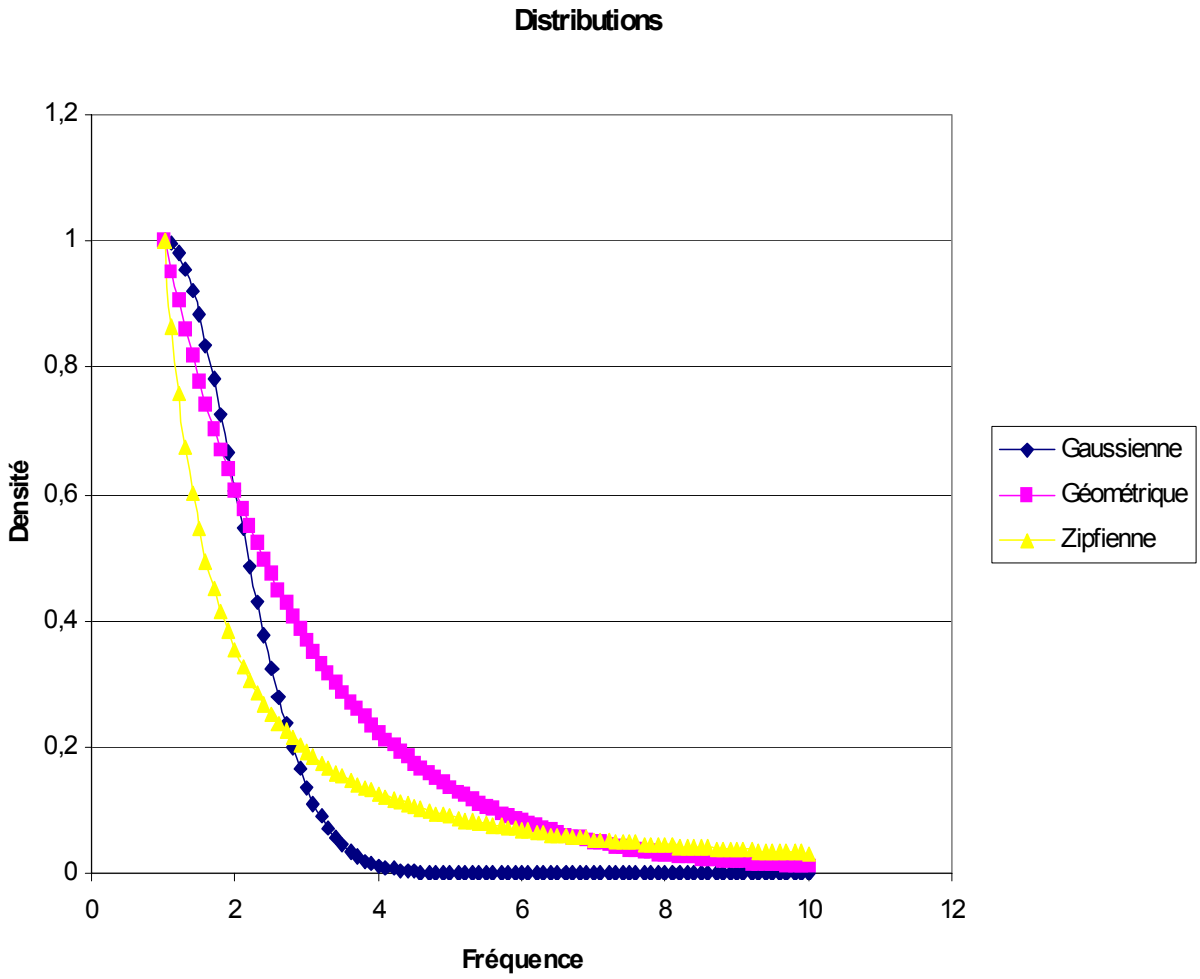
Si la limite existe et si $Limite(t \to \infty)\frac{g(t)}{Ln(t)} = \alpha$ alors on a :

Si $\alpha > 1$, g définit un processus exponentiel infométrique (c'est aussi vrai si α *est* ∞ .) .

Si $\alpha < 1$, g n'est pas un processus infométrique.

Dans le cas où α est égale à 1, les deux cas sont possibles.

La démonstration de ce théorème qui est aisé n'est pas donné ici . Remarquons cependant que la condition, « fonction croissante non bornée », de la fonction d'effort est nécessaire pour obtenir ce résultat ;.



En résumé nous ne venons de découvrir un nouveau modèle mais simplement à donner une lecture mathématique simple de la fonction logarithmique qui soit pertinente en regard de ces lois, à savoir la loi du moindre effort.

Références bibliographiques

P. Aby, P. Flandrin, N. Hohn, D. Veitch (2004) . Invariance d'échelle dans l'internet, paru dans Mesures de l'Internet sous la direction de Eric Guichard, Les Canadiens en Europe, p. 96-111.

L. Egghe. (2005) Power laws in the information production process: Lotkaian Informetrics. Elsevier

PB. Kantor, et J.L. Jung. (1998) Testing the maximum entropy principle for information retrieval. Journal of the American Society for Information Science 49 (6) 1998 p 523-527.

T. Lafouge, C. Prime Claverie (2005). Production and use of information. Charcterization of informetric distributions using effort function and density fonction .. Information Processing and Management (à paraitre juillet 2005).

C. Prime Claverie, M. Beigbeder, T. Lafouge (2005). Limits and feasibility of cositation method on the web an experiment on the web French speaking. A paraître, conference internationale de scientométrie , ISSI juillet 2005.. (2005)

C. Shannon. (1993) Collected papers edited by N.J.A. Sloane, Aaron D. Wyner. New York : IEEE Press c1993.

.