Links between information construction and information gain. Entropy and bibliometric distributions.

Thierry Lafouge, Christine Michel

Links between information construction and information gain. Entropy and bibliometric distributions.

Thierry Lafouge
Christine Michel[1]
Laboratoire Recodoc
Université Claude Bernard Lyon 1, Bat 721
43 Bd du 11 novembre 1918
69622 Villeurbanne Cedex, France.
lafouge@enssib.fr
Christine.Michel@montaigne.u-bordeaux.fr

Keywords : bibliometric distribution / entropy / least effort law / information theory

The study of the statistical regularities observed in the field of the information production and use has confirmed the existence of important similarities. Thus, the existence of regularities and measurable ratios allow the prevision and the concept of laws. In the fifties, C. Shannon (Shannon C., Weaver W. 1975 : Théorie mathématique de la communication, Bibliothèque du CELP, 1975) modeled the information circulation theory. The entropy hypothesis of this theory is: the more ranked a system is, the less information it produces. Theoretical studies have tried to formalize the connection between the bibliometric distribution and the entropy. In this paper we try to extend previous results linked with "the least effort principle" and the analytical slope of a bibliometric distribution. In the first and second parts we present some recalls about entropy and bibliometric distributions, and after that, we describe different links between them.

1. Recall about entropy

Let a source of information produce n random events of respective probabilities $p_1, p_2 \ldots, p_n$ where $\sum_{i=1}^{n} p_i = 1$. We call entropy of such a source the following function H (Caumel 1988).

$$H(p_1, p_2 \ldots p_n) = -\sum_{i=1}^{n} p_i Log_2(p_i)$$

As well as in the systems studied in physical science, the higher the entropy source is, the less organized the system is (foreseeable here). Shannon supposes that the more ranked a system is and the less information it produces. Thus, the entropy H is at a maximum if all the events are equiprobable, that is to say if we have $\forall i : \; p_i = \frac{1}{n}$ and in this case $H = Log_2(n)$.

The logarithm function of base two is often used because it is coherent with the electronic information binary digit code (the bit being defined as the maximum entropy of the random binary source).

It is possible to spread and generalize the entropy's definition for a continuous probability's law.

---

[1] Postal adress : CEM-GRESIC, Maison des Sciences de l'Homme d'Aquitaine, Esplanade des Antilles, DU, 33607 Pessac Cedex, France

In this case, we will not talk about distribution but density function (noted *v*) of a random phenomenon. We will define its entropy by the function:

$$H(v) = -\int v(t) Log(v(t)) dt$$

In bibliometry, the events normally studied are: the papers or keywords' output, the books borrowing, the author quoting … The sources taken into account will then be the authors, the bibliographical references and the books. These events are noticeable because they are featured by statistical regularities. Thus, it is interesting to observe how, by the entropy calculus, the quantity of information changes from these various sources according to the random process that lead them.

2. Recall on bibliometric distributions

The classical bibliometric distributions are: the Lotka's law related to the publications' output number, the Zipf's law related to how often words appear in a text and the Bradford's one related to the articles' dispersion in journals. According to the studies phenomena, these laws will be put clearly either in a frequential or in a ranked way.

2.1. Frequency (Lotka):

The frequential approach is the oldest one (Lotka 1926). The probability that an event appears is calculated according to its apparition frequency. An example of a frequential distribution law is the law that analyses the scientific article's output by searchers. Lotka suggests to write the distribution of the number of scientists who have written *i* publications:

$n_i = \frac{n_1}{i^2}$ $i = 1,2.....i_{\max}$ where $i_{\max}$ is the maximum productivity of a scientist. This law is generalized in the formula: $n_i = \frac{K}{i^\alpha}$ where *K* and α are constants depending on the studied field.

2.2. Rank (Zipf):

The statement of a law according to the rank implies that the information source has been previously ranked according to its output. These distributions per rank are used when the source's production ranking is inevitable to point out the apparition of regularity. The most characteristic example of a distribution per rank is Zipf's law. It observes how often words appear in English texts. By ranking these numbers in a decreasing way, he observes that there was an inversely proportional connection between the presentation rank of a word and its apparition frequency. Zipf expresses this regularity with the following equation:

$g(r) = \frac{K}{r^q}$ where g(r) represents the frequency of the rank r.

Numerous works have shown equivalencies between the distributions per rank and the frequential distributions (Egghe 1988). The choice between the one or the other of the presentations depends on the study one wants to carry out. In the study of keyword's distributions, the distribution per rank will be chosen because it is the most significant (Quoniam 1992). In a recent publication (Lhen and al. 1995), in collaboration with the

"CRRM"[2], we have shown the relevance of the generalized entropy's theory (Reyni 1960) to handle distributions.

2.3. The continuous case (Pareto):

Pareto's distribution for the continuous case has the same role as Lotka's distribution in the discrete case . It is written as follows:

$$v(t)=\frac{\alpha}{t_0}\left(\frac{t_0}{t}\right)^{\alpha+1} \quad t \geq t_0 \qquad \alpha > 0$$

to instead of to $n_i = \frac{K}{i^{1+\alpha}} \quad i = 1,2...\infty. \quad \alpha > 0$

Haitun (Haitun 1982) defines a Zipfian distribution with the following hyperbolic density function: $v(t)=\frac{C}{t^{1+\alpha}}$ where t belongs to the interval $[1,\infty)$ and where α and C are positive constants. If $\alpha = 1$, we are in the well-known case of Lokta's law. All the mathematical properties of such distributions have been widely studied. S.D. Haitun opposes this type of distribution to the Gaussian ones.

2.4. The geometrical case

Very often the geometrical distribution is used to quantify some regularities observed in bibliometry, especially in the field of documentary uses in libraries (Lafouge et al. 1997).
A geometrical distribution is written as follows:

$$G(q)(i) = (1-q)q^{i-1} \quad i = 1,2....\infty \quad 0 < q < 1$$

If the continuous equivalent of this distribution is written the following exponential is obtained:

$$v(t) = pq^t = p\,e^{t\log q} = pe^{-t\log(\frac{1}{q})}, \quad 1-p=q\ , t>0.$$

2.5. The negative binomial law

Another distribution called the negative binomial distribution is often used (Lafouge 1999) to model the use distributions.
It is written as follows:

$$Bn(q,r)(1) = (1-q)^r$$

$$Bn(q,\ r)(i) = r(r+1)....(r+i-1).\frac{q^{i-1}.(1-q)^r}{i!} \qquad i = 2....\infty\ ;\ 0 < q < 1$$

If $r = 1$ we have the equation of a geometric distribution. As far as we know there is not any law strictly equivalent to the negative binomial one in the continuous case.

[2] CRRM : Centre de Recherche Retrospectif de Marseille.

3. Entropy and distribution

3.1. Problem

Let (F,I) be a bibliometric distribution where :

F represents the set of all the patterns referring to an identified bibliographic source as for example the authors, publications.… The pattern may also be a word, defined by sequences of characters surrounded with separators, or several words.

I represents the set of all the items: each item is a positive number indicating the pattern occurrence (the number of appearances).

All these values form the source output. We have previously seen that several representations of a distribution are possible. We suppose that this source is ruled by a stable random process that can be observed and it is characterized by a function conveying the effort to produce all the different items. Thus the studied question is as follows: **for a given quantity of effort, what is the connection between the random distribution of the source and the effort function when the quantity of information (considered on the Shannon sense) produced by the source is maximized**? The technique used is called MEP, Maximum Entropy Principle, and has already been used in other studies.

Kantor (Kantor 1998) presents an application where MEP is used to improve information retrieval. Let us consider the K index terms $T_1, T_2, ...T_K$ of a query, Kantor considers the 2^K possible Boolean equations constructed with combinations of $T_1, T_2, ...T_K$. He considers each equation as queries and calls atom A(i) (i=1, …, 2^K) the corresponding answering subset of documents. A(i) is used to represent both the logical combination of terms and the set of documents indexed by that logical combination.

We make the assumption that every document is either relevant or not. In an atom, the relevant documents take the value 1, the others the value 0. An atom A(i) has the probability $p(i,1)$ to be relevant, and $p(i,0)$ to be non relevant. By knowing V_k (k=1, …, K) "the probability of relevance for documents indexed by terms T_k" and V_R, "the probability of relevance for all documents", Kantor hopes to calculate the distribution of $p(i,v)$ which maximizes the entropy, and then presents to the users the first atoms corresponding to the highest $p(i,v)$. The mathematical formalization of this problem is :

> Find a positive (condition 1) distribution $p(i,v), i = 1,...,2^K, \quad v = 0,1$ respecting (condition 2) $\sum_{i=1}^{2^K} \sum_{v=0}^{1} p(i,v) = 1,$
>
> and which maximizes the entropy $H(p) = -\sum_{i=1}^{2^K} \sum_{v=0}^{1} p(i,v) \log(p(i,v))$
>
> subjected to (condition 3') $\sum_{i=1}^{2^K} p(i,1) = \left(\sum_{i=1}^{2^K} f_i \right) V_k$ where $f_i = p(i,1) + p(i,0)$ for each k.
>
> and (condition 3'') $\sum_{i=1}^{2^K} p(i,1) = V_R$

The evaluation study results show that the MEP method is useful in IR for small collections but not for big ones. Final discussions argue that this method may be improved by taking into account finest criterion than "presence or absence of terms in a document" to estimate the document's relevance.

In the case of continuous distributions, Yablonsky (Yablonsky 1980) used the MEP in order to find the distribution of the "least effort principle" in the case of scientific article production. The necessary effort to produce an article is modeled by the function *E*, defined by $E(t)=k.Log(t)$, where *k* is a positive constant.
The effort made to produce the first article is $E(1)$, it is called the "minimum state of the scientist". The effort made to produce the second one $E(2)$ requires less effort from the scientist, and so on.

> The aim is to find a positive (condition 1) density function *v(t)* on the interval [a, ∞[$(a \geq 0)$, respecting (condition 2) $\int_a^{\infty} v(t)dt = 1$ (v density function)
>
> and which maximizes the entropy : $H(v) = -\int_a^{\infty} v(t)Log(v(t))dt$
>
> subjected to (condition 3) $\int_a^{\infty} E(t)v(t)dt = E$ (Constraint of an effort)

This model is known to be the "least effort law": the density function which results from the question of the entropy maximization, under an effort constraint, is the Zipfian function.
Indeed, the entropy maximization for the effort function $E(t)=kLog(t)$ is obtained for a density function whose analytical shape is: $v(t)=\frac{C}{t^{1+\alpha}}$ where $\alpha=\frac{k}{E}$ and $t \geq 1$

The calculation of the entropy according to α gives: $H(\alpha)=-Log(\alpha)+\frac{1}{\alpha}+1$

α being positive by definition, it is easy to show that the entropy is a decreasing function of α. The classical interpretation of Lotka's law is found, that is to say the higher α is, the bigger the gap between the number of scientists who produce a little is. (Knowing that there are few scientists who produce a lot compared to the number of scientists who produce a little).

Our aim is to spread these results to other bibliometric distributions: the geometric distribution and the binomial negative distribution.
We will consider the continuous case, so we will keep Yablonsky's notations:

- Let *v(t)* be the bibliometric distribution density function, *t* being defined on [a, ∞[, such as : $v(t) \geq 0$ on [a, ∞[$(a \geq 0)$ (condition 1).
- Let $E(t)$ be the effort production function (*E* is a positive constant corresponding to the average effort),
- And *H* be the entropy.

We have found couples (*E(t), v(t)*) where :

(condition 2) $\int_a^{\infty} v(t)dt = 1$

(condition 3) $\int_a^{\infty} E(t)v(t) = E$ (Constraint of an effort)

(condition 4) $H(v) = -\int_a^{\infty} v(t)Log(v(t))dt$ is maximized

3.2 The geometrical case

Let us remember the density function of a geometrical distribution :

$$v(t) = pq^t = p\,e^{t\log q} = pe^{-t\log(\frac{1}{q})} \quad (1-q=p)$$

We have chosen to take the linear effort function $E(t)=k(t-1)$.

The case $t=1$ as previously seen corresponds to the minimal state of the scientist who has produced one publication.

If we put ourselves as previously in the case of the scientific output, in the case of a linear function $E(t)$, we want to show that when the MEP is applied with constraint of an effort, the resulting function is a geometric distribution with a density function such as :

$$\boxed{w(t)=\alpha e^{-\alpha(t-1)} \quad where \quad \alpha=\frac{k}{E} \quad \text{et} \quad t\geq 1}$$

Remark : in the case of $t=1$, we have $\frac{1}{\alpha}$ the average of v. The condition 3 thus aims at setting the expectation.

Demonstration

These results are shown using variational calculation techniques.

Let us demonstrate that the function:

$$w(t)=\alpha e^{-\alpha(t-1)} \qquad \text{where} \quad \alpha=\frac{k}{E} \quad \text{et} \quad t\geq 1$$

checks the conditions

$$v \geq 0 \text{ on } [1,\infty[\qquad (1)$$

$$\int_1^{\infty} v(t)dt = 1 \qquad (2)$$

$$\int_1^{\infty} k(t-1)v(t)dt = E \qquad (3)$$

and maximizes the function: $H(v)=-\int_1^{\infty} v(t)Log(v(t))dt$

We can easily show that the function w checks the conditions (1) (2) and (3). Let us show that w maximizes the entropy. We will prove that H reaches its maximum for the function w.

Let F be the following function:

$F(t,v)=v\,Log(v)+\lambda v+\alpha(t-1)v$ where λ is a constant whose value is: $\lambda=-1-\text{Log}(\alpha)$

We have: $\frac{\partial}{\partial v}F(t,v)=Log(v)+1+\lambda+\alpha(t-1)$

We can easily show that this derivative cancels for w.

So for t fixed, we have: $\frac{\partial}{\partial v}F(t,w)=0$ and $\frac{\partial^2}{\partial^2 v}F(t,v)=\frac{1}{v}\geq 0$

F being convex and $\frac{\partial F}{\partial v}$ canceling for w with any value of t determined we can write:

$\forall v, F(t,v)\geq F(t,w)$

So $\forall v,\ v\,Log(v)+\lambda v+\alpha(t-1)v \geq w\,Log(w)+\lambda w+\alpha(t-1)w$

And so $\forall v,\ \int_1^{\infty}\left(v\,Log(v)+\lambda v+\alpha(t-1)v\right)dt \geq \int_1^{\infty}\left(w\,Log(w)+\lambda w+\alpha(t-1)w\right)dt$

Let ν be any function checking the normalization (conditions 2) and (condition 3)

$\int_1^{\infty} v\,Log(v)dt \geq \int_1^{\infty} w\,Log(w)dt$ hence the result is proved.

$\boxed{H(\alpha)=1-Log(\alpha)}$.

We have the same result as previously for the variation of *H* in function of α. Moreover we can notice that the entropy calculated with a linear effort law is always inferior to an entropy calculated with a logarithmic effort law and that the difference varies in an inversely proportional way. The dispersion, and then the entropy, is stronger in the Zipfian case. This result justifies the choice of the entropy of order 1 to feature the diversity of a Zipfian distribution (Lhen et al.1995).

Note

We can show Yablonsky's previous result using the same technique with the function denoted *F*:

$F(t,v)=vLog(v)+\lambda v+(1+\alpha)vLog(t)$ *where* $\lambda=-Log(\alpha)-1$

3.2. The negative binomial case

We have said that we do not know the density function of the negative binomial law in the continuous case. So we will use convolution techniques to build a new distribution that we will call here " pseudo negative binomial".

The convolution technique is defined by :

If X_1 and X_2 are two independent and continuous random variables having respectively as density functions F_1 and F_2 defined on the interval $]-\infty,+\infty[$, then the random variable X_1+X_2 will have the density function F_1*F_2, called convolution product of F_1 and F_2 and defined by :

$$(F_1 * F_2)(x)=\int_{-\infty}^{x} F_1(y).F_2(x-y)dy$$

This definition is generalized for a convolution product of order j.

Indeed, if X_j is a finite series of independent identically distributed random variables of density F we show that the random variable $\sum_{i=1}^{j} X_i$ has a density function F_j defined by the following convolution:

$F_1 = F$

$F_2 = F * F \qquad F_2(x) = \int_{-\infty}^{x} F(y).F(x-y)dy$

$$F_j = F_{j-1} * F \qquad j = 2,3....$$

The convolution techniques have been used to give a new interpretation of Lotka's law (Egghe 1994).
We know (Calot 1984) that in the discrete case the sum of *j* independent variables of a geometric law *G(q)* is a negative binomial law $Bn(jq)$.
The exponential distribution is the geometric distribution density function: $v(t) = \alpha e^{-\alpha t} \quad t \geq 0$. If we build a density function v_j from the convolution of *j* exponential distributions, we may consider it as the continuous version of a negative binomial law, we will call it "pseudo negative binomial " law.

A simple calculation shows that the convolution product of order j of the *v* function is :

$$v_j(t) = \alpha^j \frac{t^{j-1}}{(j-1)!} e^{-\alpha t} \qquad t \geq 0$$

Egghe has shown (Egghe 1994) the stability properties for the geometric distribution using this convolution product. If the original distribution is an exponential one, we can interpret this distribution v_j in different ways:

-In a context of articles output, $v_j(i)$ is the proportion of authors who have written i articles, each article having exactly *j* authors.

- In a context of bibliographic references keywords distributions, $v_j(i)$ is the proportion of words used *i* times, each reference having exactly *j* keywords

Then the question that we want to solve is: if we set the effort quantity (noted E_j), what is the nature (linear, logarithmic…) of the effort distribution (noted *EF*) linked to a random process of a negative binomial pseudo type, when the information quantity is maximum? This question is mathematically written:

Let's consider the following distribution : $v_j(t) = \alpha^j \frac{t^{j-1}}{(j-1)!} e^{-\alpha t} \quad t \geq 0, j > 0$

What is the nature of the effort distribution *EF* that checks the following conditions:

$$\int_0^\infty EF(t) v_j(t)\, dt = E_j \qquad \text{(Constraint of an effort)}$$

and maximizes the entropy *H*: $-H(v) = -\int_0^\infty v(t) Log(v(t))\, dt$ for any function ν checking the conditions:

$v(t) \geq 0$ on $[0, \infty[$

$$\int_0^\infty v(t)\, dt = 1$$

We will see if the effort function *EF*, $\boxed{EF(t) = \alpha t - (j-1)\log(t)}$ is valid to solve the problem.

Demonstration:

We can easily check by recurrence that $\int_0^{\infty} v_j(t)dt = 1$.

It's more difficult to know the value of the effort $\int_0^{\infty} EF(t)v_j(t)dt$ **.**

Let us put $K(\alpha, j) = \frac{\alpha^j}{(j-1)!}$

$$\int_0^{\infty} EF(t)v_j(t)dt = K(\alpha, j)\left(\int_0^{\infty} \alpha t^j e^{-\alpha t}dt - \int_0^{\infty}(j-1)\log(t)t^{j-1}e^{-\alpha t}dt\right)$$

$$\int_0^{\infty} \alpha t^j e^{-\alpha t}dt = \left[-t^j e^{-\alpha t}\right]_0^{\infty} + \int_0^{\infty} jt^{j-1}e^{-\alpha t}dt$$

however we have $\left[t^j e^{-\alpha t}\right]_0^{\infty} = 0$ and we have seen that $\int_0^{\infty} K(\alpha, j)t^{j-1}e^{-\alpha t}dt = 1$

So $\int_0^{\infty} \alpha t^j e^{-\alpha t}dt = \frac{j}{K(\alpha, j)}$

So we have shown that: $\int_0^{\infty} EF(t)v_j(t)dt = j - (j-1).K(\alpha, j).\int_0^{\infty}\log(t)t^{j-1}e^{-\alpha t}dt$

Let's now calculated the value of $\int_0^{\infty}\log(t)t^n e^{-\alpha t}dt$

$$\int_0^{\infty}\log(t)t^n e^{-\alpha t}dt = \int_0^{1}\log(t)t^n e^{-\alpha t}dt + \int_1^{\infty}\log(t)t^n e^{-\alpha t}dt$$

$$\int_0^{\infty}\log(t)t^n e^{-\alpha t}dt = \lim_{x\to 0}\left(\left[\log(t)U_n(t)\right]_x^1 - \int_x^1 \frac{U_n(t)}{t}dt\right) + \lim_{y\to\infty}\left(\left[\log(t)U_n(t)\right]_1^y - \int_1^y \frac{U_n(t)}{t}dt\right)$$

With $U_n(t) = \int_0^t s^n e^{-\alpha s}ds$

Let's calculate $U_n(t) = \int_0^t s^n e^{-\alpha s}ds$

We will show recurrently that $U_n(t) = -\sum_{i=0}^{n}\frac{n!t^{n-i}e^{-\alpha t}}{(n-i)!\alpha^{i+1}} + \frac{n!}{\alpha^{n+1}}$

$U_0(t) = \int_0^t e^{-\alpha s}ds = -\frac{(e^{-\alpha t}-1)}{\alpha}$ and $-\sum_{i=0}^{0}\frac{0!t^0 e^{-\alpha t}}{(0-i)!\alpha^{i+1}} + \frac{0!}{\alpha} = -\frac{(e^{-\alpha t}-1)}{\alpha} = U_0(t)$

Now, let suppose that $\forall n > 0$, $U_{n-1}(t) = -\sum_{i=0}^{n-1}\frac{(n-1)!t^{n-1-i}e^{-\alpha t}}{(n-1-i)!\alpha^{i+1}} + \frac{(n-1)!}{\alpha^n}$.

It's easy to show that $U_n(t) = -\frac{t^n e^{-\alpha t}}{\alpha} + \frac{n}{\alpha}U_{n-1}(t)$

So $U_n(t) = -\frac{t^n e^{-\alpha t}}{\alpha} - \sum_{k=1}^{n}\frac{n!t^{n-k}e^{-\alpha t}}{(n-k)!\alpha^{k+1}} + \frac{n!}{\alpha^{n+1}}$

And So $U_n(t) = -\sum_{k=0}^{n} \frac{n! t^{n-k} e^{-\alpha t}}{(n-k)! \alpha^{k+1}} + \frac{n!}{\alpha^{n+1}}$

In Annex 1 we have the result :

$$\lim_{x\to 0}\left(-\log(x)U_n(x) - \int_x^1 \frac{U_n}{t}dt \right) = \sum_{k=0}^{n-1} \frac{n!}{(n-k)!\alpha^{k+1}} \lim_{x\to 0}\left(\int_x^1 t^{n-k-1} e^{-\alpha t} dt \right) + \frac{n!}{\alpha^{n+1}} \lim_{x\to 0}\left(\int_x^1 \frac{\left(e^{-\alpha t} - 1\right)}{t} dt \right)$$

In Annex 2 we have the result :

$$\lim_{y\to \infty}\left(-\log(y)U_n(y) - \int_1^y \frac{U_n(t)}{t}dt \right) = \sum_{k=0}^{n-1} \frac{n!}{(n-k)!\alpha^{k+1}} \lim_{y\to \infty}\left(\int_1^y t^{n-k-1} e^{-\alpha t} dt \right) + \frac{n!}{\alpha^{n+1}} \lim_{y\to \infty}\left(\int_1^y \frac{e^{-\alpha t}}{t} dt \right)$$

So

$$\int_0^\infty \log(t) t^{j-1} e^{-\alpha t} dt = \sum_{k=0}^{n-1} \frac{n!}{(n-k)!\alpha^{k+1}} \int_0^\infty t^{n-k-1} e^{-\alpha t} dt + \frac{n!}{\alpha^{n+1}} \lim_{x\to 0}\left(\int_x^1 \frac{\left(e^{-\alpha t} - 1\right)}{t} dt \right) + \frac{n!}{\alpha^{n+1}} \lim_{y\to \infty}\left(\int_1^y \frac{e^{-\alpha t}}{t} dt \right)$$

We have $\int_0^\infty t^{n-k-1} e^{-\alpha t} dt = \frac{(n-k-1)!}{\alpha^{n-k}}$

So $\int_0^\infty \log(t) t^n e^{-\alpha t} dt = \frac{n!}{\alpha^{n+1}} \left[\sum_{k=0}^{n-1} \frac{1}{(n-k)} + \int_0^1 \frac{\left(e^{-\alpha t} - 1\right)}{t} dt + \int_1^\infty \frac{e^{-\alpha t}}{t} dt \right]$

$$\int_0^\infty \log(t) t^n e^{-\alpha t} dt = \frac{n!}{\alpha^{n+1}} \left[\sum_{k=1}^{n} \frac{1}{k} + \int_0^1 \frac{\left(e^{-\alpha t} - 1\right)}{t} dt + \int_1^\infty \frac{e^{-\alpha t}}{t} dt \right]$$

By using complete Gamma function it is possible to say that :

$\int_0^1 \frac{\left(e^{-\alpha t} - 1\right)}{t} dt + \int_1^\infty \frac{e^{-\alpha t}}{t} dt = -\gamma - \log(\alpha)$ (γ is Euler's constant :0.5772….)

So $\int_0^\infty \log(t) t^n e^{-\alpha t} dt = \frac{n!}{\alpha^{n+1}} \left[\sum_{k=1}^{n} \frac{1}{k} - \gamma - \log(\alpha) \right]$

With initial notations : $\int_0^\infty \log(t) t^{j-1} e^{-\alpha t} dt = \frac{j-1!}{\alpha^j} \left[\sum_{k=1}^{j-1} \frac{1}{k} - \gamma - \log(\alpha) \right]$

So $\int_0^\infty EF(t) v_j(t) dt = j - (j-1) \left[\sum_{k=1}^{j-1} \frac{1}{k} - \gamma - \log(\alpha) \right]$

So $E_j = j - (j-1) \left[\sum_{k=1}^{j-1} \frac{1}{k} - \gamma - \log(\alpha) \right]$

Now let us show that EF maximizes the entropy.

We will show that the entropy *H* reaches its minimum for the function $v_j(t) = K(\alpha, j) t^{j-1} e^{-\alpha t}$.

Let *F* be the following function:

$F(t, v) = v\,Log(v) + \lambda v + EF(t) v$ where λ is the constant $\lambda = -Log\left(K(a, j)\right) - 1$

We have: $\frac{\partial}{\partial v}F(t,v)=Log(v)+1+\lambda+EF(t)$

Let us put : $\nu_j(t)=K(\alpha,j)t^{j-1}e^{-\alpha t}$

$\frac{\partial}{\partial v}F(t,\nu_j)=Log\big(K(\alpha,j)\big)+(j-1)log(t)-\alpha t+1+\lambda+\alpha t-(j-1)log(t)$

$\frac{\partial}{\partial v}F(t,\nu_j)=Log\big(K(\alpha,j)\big)+1+\lambda$

If we replace λ by its value then this derivative cancels for ν_j .

For t fixed we have: $\frac{\partial^2}{\partial^2 v}F(t,v)=\frac{1}{v}\geq 0$

F being convex and $\frac{\partial F}{\partial v}$ canceling in $\nu_j(t)$ for any value of t fixed we can write:

$\forall v \quad F(t,v)\geq F(t,\nu_j(t))$

that is to say:

$\forall v \quad v\,Log(v)+\lambda v+\alpha(t-1)v\geq \nu_j(t)Log(\nu_j(t))+\lambda\nu_j(t)+\alpha(t-1)\nu_j(t)$

$\forall v \quad \int_0^\infty \big(v\,Log(v)+\lambda v+\alpha(t-1)v\big)dt\geq \int_0^\infty\Big(\nu_j(t)Log(\nu_j(t))+\lambda\nu_j(t)+\alpha(t-1)\nu_j(t)\Big)dt$

Let ν be any function checking the normalization (conditions 2) and (condition3):

$\forall v \quad \int_0^\infty \big(v\,Log(v)\big)dt\geq \int_0^\infty\Big(\nu_j(t)Log(\nu_j(t))\Big)dt$

So the entropy is maximum for $\nu_j(t)$.

Calculation of the entropy $-H(v)=\int_0^\infty v(t)Log(v(t))dt$

$-H(\nu_j)=\int_0^\infty K(\alpha,j)t^{j-1}e^{-\alpha t}log\big(K(\alpha,j)t^{j-1}e^{-\alpha t}\big)dt$

$-H(\nu_j)=\int_0^\infty K(\alpha,j)t^{j-1}e^{-\alpha t}\big(log\big(K(\alpha,j)\big)+(j-1)log(t)-\alpha t\big)dt$

$-H(\nu_j)=\int_0^\infty K(\alpha,j)Log\big(K(\alpha,j)\big)t^{j-1}e^{-\alpha t}dt+\int_0^\infty K(\alpha,j)(j-1)t^{j-1}e^{-\alpha t}\log(t)dt-\int_0^\infty \alpha K(\alpha,j)t^{j}e^{-\alpha t}dt$

We have: $\int_0^\infty K(\alpha,j)t^{j-1}e^{-\alpha t}dt=1$, $\int_0^\infty \alpha t^{j}e^{-\alpha t}dt=\frac{j}{K(\alpha,j)}$ and

$\int_0^\infty \log(t)t^{j-1}e^{-\alpha t}dt=\frac{(j-1)!}{\alpha^j}\left[\sum_{k=1}^{j-1}\frac{1}{k}-\gamma-\log(\alpha)\right]$

So

$-H(\nu_j)=Log\big(K(\alpha,j)\big)+(j-1)\left(\sum_{k=1}^{j-1}\frac{1}{k}-\gamma-\log(\alpha)\right)-j$

We can remark that for j=1 we have the same results as for the geometrical low, i.e. $H(v_1) = 1 - Log(\alpha)$

We can remark that $-H(v_j) = Log(K(\alpha, j)) - Ej$

We want now to analyze the characteristics of the two functions Ej and $H(v_j)$.

In all cases of α, Ej and $H(v_j)$ are increasing functions of j, the author's number. It means that publishing with many authors required more effort but induce, in all cases, a gain of information. We analyze now the difference $H(v_j) - Ej$.

$H(v_j) - Ej = -Log(K(\alpha, j)) = Log(\frac{(j-1)!}{\alpha^j})$ j =1,2…….

$H(v_j) - Ej$ is depending on j and α. The 3 following curves give the results for α=1 (figure 1), α =2 (figure 2) and α =3 (figure 3).

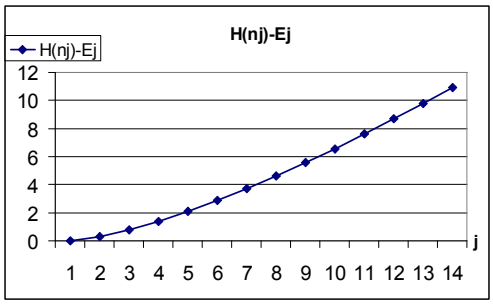

Figure 1 : case of α =1

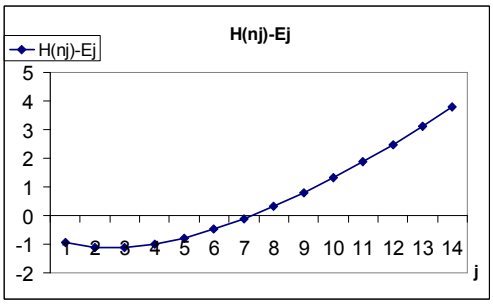

Figure 2 : case of α =3

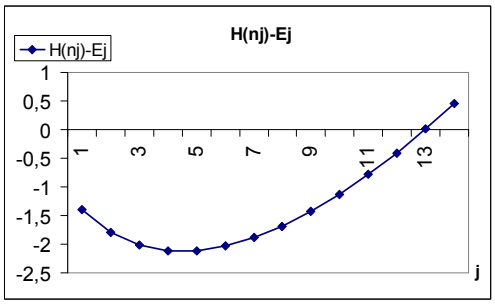

Figure 3 : case of α = 5

As it's represented in figure 1, in the case of α=1 we have always $H(v_j) > Ej$, which means that the gain of information is always higher than the effort need to produce it. On the other case, if α>1 (figure 2 and 3), the curves of $H(v_j) - E_j$ is first negative, cut the x-axe on a particular point m_α and stay positive after than. The value m_α define the minimum number of authors required for having a gain of information relative to a paper, biggest than the effort due to produce it.

4. Conclusion

We have observed that, in the Zipfian distribution law case, the corresponding effort function was logarithmic, in the geometric law case, it was linear. In each case the entropy is

decreasing with α. In the pseudo negative binomial case, the effort function is composed by two functions: the fist one is linear (effort constant) and the other logarithmic (the least effort law). For *j* fixed the corresponding entropy is decreasing with α.
Let us now vary j. Let us remember a possible interpretation of the pseudo negative binomial law: in a context of articles output, $v_j(i)$ is the proportion of authors who have written *i* articles, each article having exactly *j* authors. There is a link between the effort and the entropy variation. We have shown that for each value of α, there is a number m_α representing the minimum number of authors required to write a paper having more information than the effort due to produce it. In the case of $\alpha=1$, we have $m_1=1$, i.e. whatever the number of author, each new article produce more information that the effort developed to write it. That needs to be verify.

References

EGGHE L.(1988): *On the classification of the classical bibliometric laws.* Jounal of Documentation, Vol 44(1), 1988, p. 53-62.

EGGHE L. (1994): *Special features of the author publication relationship and a new explanation of Lotka's law based on convolution theory.* Journal of the American Society for Information Science, Vol 45(6),1994, p. 422-427.

CALOT G. (1984): Cours de calcul de probabilité. Chapitre 12. Dunod décision 1984, 476 pages.

CAUMEL Y. (1988): Probabilités, théorie et applications. Chapitre 8. Eyrolles, 1988, 290 pages.

HAITUN SD. Stationary Scientometrics Distributions. Scientometrics N°4, 1982, Part I p.5-25, Part II p.89-104, Part III p.181-194.

KANTOR Paul B., JUNG Jin Lee: *Testing the Maximum Entropy Principle for information Retrieval.* Journal of the American Society for Information Science, Vol 49 (6), 1998, p 557-556.

LAFOUGE T**.,** LAINE CRUZEL S. (1997): *A new explanation of the geometrical law in the case of library circulation data* . Information Processing and Management, Vol 33 (4), 1997, p. 523-527.

LAFOUGE T.,GUINET E. (1999): *A new explanation of the negative binomial law and the Poisson law with regard to library circulation data.* Journal of Information Science Vol 25 (1), 1999, p 89-93.

LHEN J., LAFOUGE T., ELSKENS Y., QUONIAM L., DOU H. (1995): *La « statistique des lois de Zipf.* Revue Française de Bibliométrie N°14, 1995, p. 135-146.

LOTKA A. (1926) : *The frequency distribution of scientific productivity*. Journal of the Washington Academy of Sciences, 16, 1926, p.317-323.

QUONIAM L. (1992): *Bibliométrie sur les références bibliographiques: méthodologie*, p. 244, 261.La Veille technologique; l'information scientifique, technique et industrielle. Dunod, 1992, 436 pages.

REYNI A. (1961): *On measures of entropy and information.* In NEYMAN J. ed Berkeley symposium on mathematical statistic and probability .

SHANNON C.,WEAVER W.(1975): Théorie mathématique de la communication. Bibliothèque du CEPL, 1975, 188 pages.

YABLONSKY A.L (1980): *On fundamental regularities of the distribution of scientific productivity.* Scientometrics, Vol 2,(1), 1980, p. 3-34.

Annex 1

$$\lim_{x\to 0}\left(-\log(x)U_n(x)\right)=\lim_{x\to 0}\left(\sum_{k=0}^{n}\frac{\log(x)n!x^{n-k}e^{-\alpha x}}{(n-k)!\alpha^{k+1}}-\frac{n!\log(x)}{\alpha^{n+1}}\right)$$

$$\lim_{x\to 0}\left(-\log(x)U_n(x)\right)=\lim_{x\to 0}\left(-\sum_{k=0}^{n-1}\frac{\log(x)n!x^{n-k}e^{-\alpha x}}{(n-k)!\alpha^{k+1}}-\frac{\log(x)n!e^{-\alpha x}}{\alpha^{n+1}}+\frac{n!\log(x)}{\alpha^{n+1}}\right)$$

$$\lim_{x\to 0}\left(-\log(x)U_n(x)\right)=\lim_{x\to 0}\left(-\sum_{k=0}^{n-1}\frac{\log(x)n!x^{n-k}e^{-\alpha x}}{(n-k)!\alpha^{k+1}}-\frac{\log(x)n!\left(e^{-\alpha x}-1\right)}{\alpha^{n+1}}\right)$$

$$\lim_{x\to 0}\left(-\log(x)U_n(x)\right)=-\frac{n!}{(n-k)!\alpha^{k+1}}\sum_{k=0}^{n-1}\lim_{x\to 0}\left(\log(x)x^{n-k}e^{-\alpha x}\right)-\frac{n!}{\alpha^{n+1}}\lim_{x\to 0}\left(\log(x)\left(e^{-\alpha x}-1\right)\right)$$

We know that $\forall k<n,\quad \lim_{x\to 0}\left(\log(x)x^{n-k}e^{-\alpha x}\right)=0$ And $\lim_{x\to 0}\left(\log(x)\left(e^{-\alpha x}-1\right)\right)=0$

So $\lim_{x\to 0}\left(-\log(x)U_n(x)\right)=0$

Moreover $\lim_{x\to 0}\left(\int_x^1\frac{U_n}{t}dt\right)=\lim_{x\to 0}\int_x^1\left(-\sum_{k=0}^{n}\frac{n!t^{n-k-1}e^{-\alpha t}}{(n-k)!\alpha^{k+1}}+\frac{n!}{\alpha^{n+1}t}\right)dt$

$$\lim_{x\to 0}\left(\int_x^1\frac{U_n}{t}dt\right)=\lim_{x\to 0}\int_x^1\left(-\sum_{k=0}^{n-1}\frac{n!t^{n-k-1}e^{-\alpha t}}{(n-k)!\alpha^{k+1}}-\frac{n!}{\alpha^{n+1}}\frac{e^{-\alpha t}}{t}+\frac{n!}{\alpha^{n+1}t}\right)dt$$

$$\lim_{x\to 0}\left(\int_x^1\frac{U_n}{t}dt\right)=-\sum_{k=0}^{n-1}\frac{n!}{(n-k)!\alpha^{k+1}}\lim_{x\to 0}\left(\int_x^1 t^{n-k-1}e^{-\alpha t}dt\right)-\lim_{x\to 0}\left(\frac{n!}{\alpha^{n+1}}\int_x^1\frac{\left(e^{-\alpha t}-1\right)}{t}dt\right)$$

So $\lim_{x\to 0}\left(-\log(x)U_n(x)-\int_x^1\frac{U_n}{t}dt\right)=\sum_{k=0}^{n-1}\frac{n!}{(n-k)!\alpha^{k+1}}\lim_{x\to 0}\left(\int_x^1 t^{n-k-1}e^{-\alpha t}dt\right)+\lim_{x\to 0}\left(\frac{n!}{\alpha^{n+1}}\int_x^1\frac{\left(e^{-\alpha t}-1\right)}{t}dt\right)$

Annex 2

$$\lim_{y\to\infty}\left(-\log(y)U_n(y)-\int_1^y\frac{U_n(t)}{t}dt\right)=\lim_{y\to\infty}\left(-\log(y)U_n(y)\right)+\lim_{y\to\infty}\left(-\int_1^y\frac{U_n(t)}{t}dt\right)$$

$$\lim_{y\to\infty}\left(-\log(y)U_n(y)\right)=\lim_{y\to\infty}\left(-\log(y)\sum_{k=0}^{n}\frac{n!y^{n-k}e^{-\alpha y}}{(n-k)!\alpha^{k+1}}\right)+\frac{n!}{\alpha^{n+1}}\lim_{y\to\infty}\left(\log(y)\right)$$

$$\lim_{y\to\infty}\left(-\log(y)U_n(y)\right)=\sum_{k=0}^{n}\left(\frac{n!}{(n-k)!\alpha^{k+1}}\lim_{y\to\infty}\left(-\log(y)y^{n-k}e^{-\alpha y}\right)\right)+\frac{n!}{\alpha^{n+1}}\lim_{y\to\infty}\left(\log(y)\right)$$

And $\forall k \le n, \quad \lim_{y\to\infty}\left(-\log(y)y^{n-k}e^{-\alpha y}\right)=0$

So $\lim_{y\to\infty}\left(-\log(y)U_n(y)\right)=\frac{n!}{\alpha^{n+1}}\lim_{y\to\infty}\left(\log(y)\right)$

$$\lim_{y\to\infty}\left(-\int_1^y\frac{U_n(t)}{t}dt\right)=\lim_{y\to\infty}\left(\int_1^y\left(\sum_{k=0}^{n}\frac{n!t^{n-k-1}e^{-\alpha t}}{(n-k)!\alpha^{k+1}}-\frac{n!}{\alpha^{n+1}t}\right)dt\right)$$

$$\lim_{y\to\infty}\left(-\int_1^y\frac{U_n(t)}{t}dt\right)=\lim_{y\to\infty}\left(\int_1^y\left(\sum_{k=0}^{n-1}\frac{n!t^{n-k-1}e^{-\alpha t}}{(n-k)!\alpha^{k+1}}+\frac{n!}{\alpha^{k+1}}\frac{e^{-\alpha t}}{t}-\frac{n!}{\alpha^{n+1}t}\right)dt\right)$$

$$\lim_{y\to\infty}\left(-\int_1^y\frac{U_n(t)}{t}dt\right)=\sum_{k=0}^{n-1}\frac{n!}{(n-k)!\alpha^{k+1}}\lim_{y\to\infty}\left(\int_1^y t^{n-k-1}e^{-\alpha t}dt\right)+\frac{n!}{\alpha^{k+1}}\lim_{y\to\infty}\left(\int_1^y\frac{e^{-\alpha t}}{t}dt\right)-\frac{n!}{\alpha^{n+1}}\lim_{y\to\infty}\left(\log(y)\right)$$

$$\lim_{y\to\infty}\left(-\log(y)U_n(y)-\int_1^y\frac{U_n(t)}{t}dt\right)=\sum_{k=0}^{n-1}\frac{n!}{(n-k)!\alpha^{k+1}}\lim_{y\to\infty}\left(\int_1^y t^{n-k-1}e^{-\alpha t}dt\right)+\frac{n!}{\alpha^{n+1}}\lim_{y\to\infty}\left(\int_1^y\frac{e^{-\alpha t}}{t}dt\right)$$