Production and use of information. Characterisation of informetric distributions using effort function and density function. Exponentiial informetric process

Thierry Lafouge, Camille Prime

Production and use of information. Characterization of informetric distributions using effort function and density function.

Exponential Informetric Process

Abstract
Statistical regularities observed in the production or use of information have been studied for a long time. In this article we define an Exponential Informetric Process to formalize these stochastic process. It is defined by combining an effort function with a density function. Without using the powerful results of Price on the cumulative advantages process this characterization clarifies the principle of least effort.. Some links between statistical theory of information and some informetric distributions are enhanced.

. **Key words** : effort function; exponential process; entropy

1. Introduction

Scientific production is cumulative by nature. If we look from a scientometric point of view and evaluate the number of articles produced by researchers, we know that each new scientific article published is usually built on previous results. In his article, an author quotes the bibliographic references of other work produced earlier (which may be his) in order to validate his work. Furthermore, a known social phenomenon, "success breeds success", will then occur: the n^{th+1} publication will be easier than the preceding one. It will require less effort than the n^{th} publication. This law may prove false for a given period. Let us take the example of a known researcher having published numerous articles and tackling a new research topic, who wants to publish his results in a journal that does not know him: it is possible that his publications will not be accepted readily by this journal and need lots of work from him for his publication to be accepted easily again. Various aspects of this well-known phenomenon are examined in scientometrics. The best known result is that of cumulative advantages formulated by Price in 1976 (Price 1976). He shows that a law of probability - often called the cumulative advantages process - explains these phenomena when we pass to extreme cases. This is known in informetrics through the laws of Bradford, Lotka (production of articles by the aforementioned researchers) and Zipf. These laws are called power laws in the information production process (Egghe 2005).
In this article we introduce an effort function. Mathematically, this effort function is defined simply through an exponential function. We shall speak of the Exponential Informetric Process. This mathematical formalism will allow us to establish simply a linear relationship between the information content, or entropy within the meaning of Shannon's theory information, and the average amount of effort. This average amount of effort produced by the process is obtained by using an distribution of effort. This formulation clarifies certain well-

known characteristics of the power distributions quoted previously, namely their link with the maximum entropy principle (Yablonsky 1980).

2. Information Product Process and effort function

The study of statistical regularities observed in the production or utilization of information confirms the existence of significant similarities. Also, the existence of regularities and measurable ratios allow us to validate the concept of laws of information. These laws are known under the names of the researchers who observed and analyzed these statistical regularities: Bradford (distribution of articles on a given topic in scientific journals), Lotka (production of articles by researchers in a scientific community), and Zipf (regularity of the words in the texts). These distributions, which the bibliometrician very often encounters when statistically analyzing collections, generally fit into simple unidimensional models. We represent these productions in the diagram of figure 1, introduced into informetric systems by Leo Egghe (Egghe 1990) and called "Information Production Process" (IPP). An IPP is a triplet made up of a bibliographical source, a production function, and all the elements (items) produced. Here, the definition of bibliographical source is very broad. It enables us to describe, with the same term, all the authors in a scientific community, and also all the words in a text.

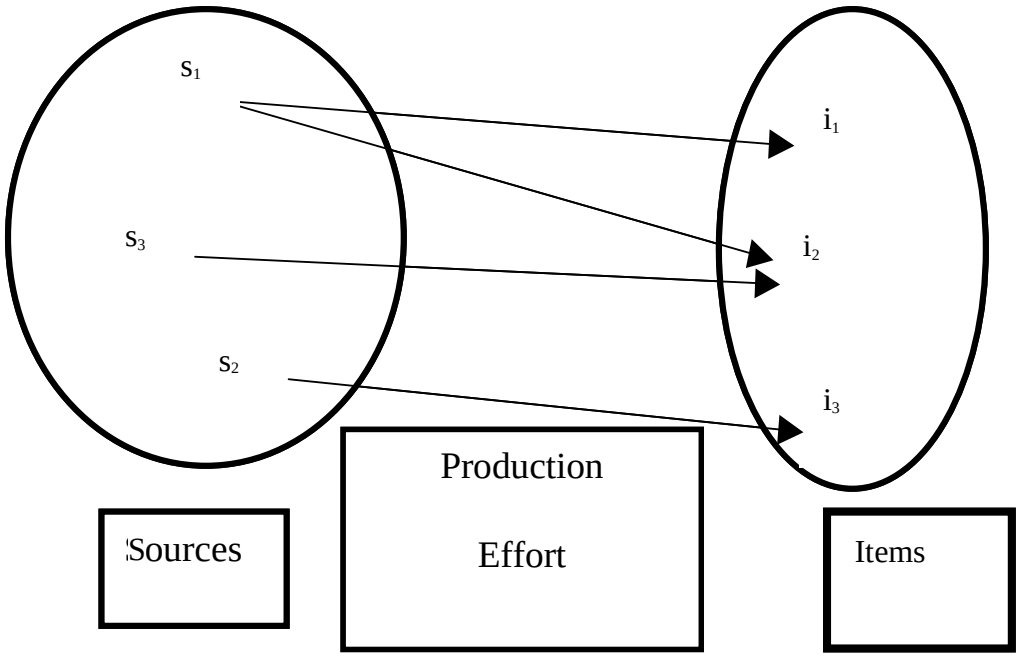

Figure 1
Schematic representation of an Information Production Process with effort function

Any IPP is defined using a production function. As an example, for the best known IPP in informetrics quoted above, we have the following production functions:

- Authors (sources) produce articles (items) - Law of Lotka (Lotka 1926),
- Journals (sources) publishe (produce) articles related to a well determined subject (items) - Law of Bradford (Bradford 1934),
- Words (sources) produce occurrences of words (items) - Law of Zipf (Zipf 1949).

The observation and statistical treatments of these processes lead us to calculate the distribution of the observed frequencies. We use here the size frequency form. The distribution of frequencies is noted ν where $\nu(i)$ represents the number of sources that have

produced i items ($i = 1,2....i_{\max}$ (maximum number of items produced)). In general, we observe a decreasing distribution, which is characteristic of these processes. For example, in Lotka's formulation this means the number of authors having produced i articles is greater than the number of authors having produced $i+1$ articles. Also, these greatly decreasing distributions usually fit power distributions:

$$\nu(i) = \frac{k}{i^{\alpha}} \; i = 1.2..., i_{\max} \; .\alpha > 0, k > 0.$$

where k is a coefficient of standardization and α an indicator of concentration characterising the dispersion of the distribution. The exponent α is only an indicator of concentration within the family of Lotka or power function.

In the work of Egghe (Egghe 2005), we can find a complete panorama of the properties and applications of power laws in the information product process.

We henceforth assume that each item produced requires a certain amount of effort. In this article, we introduce the effort function f where $f(i)$ denotes the average amount of effort from a source needed to produce i items $i = 1,2....i_{\max}$. This amount of effort is characteristic of the process and is not necessarily directly observable. We give possible interpretations of this function. In the first process, it depends on the publication system set up by a scientific community. In the second, it is the editorial system that determines the effort function. For word production, we can quantify the effort produced by the length of the word: the longer the word, the greater the effort. The average amount of effort, denoted $\overline{F}$, produced by such a process is : $\overline{F} = \sum_{i=1}^{i=i_{\max}} f(i).\nu(i)$

If f is the identity function $f(i) = i$, the average amount of effort produced by the process is simply equal to the number of items produced. We will suppose that this function is increasing. The type of growth will characterize the process. A growth of a concave function that is slowing, such as the logarithmic function, $f(i) = Log\,(i)$, will characterize the power distributions that we have just seen.

3. Average information content or entropy

3.1 Definition

In 1948 C. Shannon worked out a statistical theory on the transmission of electrical signals. This statistical theory of information (Shannon, C. 1993). stipulates that the more the states of a system are equiprobable, the more the process produces information. This work extends the theory of Hartley and Wiener, which stipulating that the more an event is unpredictable, the more information it contributes. The average information content of a process is given by the measurement of the entropy (denoted $\overline{H}$) of Shannon. If p_i , $i = 1,...., n$ denotes n probabilities such as $\sum_{i=1}^{n} p_i = 1$ we have

$$\overline{H} = -\sum_{i=1}^{n} p_i.Log\,(p_i) \; .$$

Note: we will use here $Log = Log_e$. All the results are valid for a logarithmic function in any base. With a concern for standardization, the information theory uses the function in base 2.

In (Lafouge 2003), we showed all the wealth and omnipresence that the Shannon theory has with the information sciences. Properties inherent in the information sciences are often quoted. A new result published from time to time shows an unexpected aspect, as for example in Information Retrieval, the publication of Sandor Dominich (Dominich and al. 2004). In this article the author use the well known property, " The farther apart p_i from each other the smaller the amount of information", to define an UDO (*Uncertainty decreasing operation*) probability space.

3.2 Maximum entropy principle and principle of the least effort

The maximum entropy principle (denoted here MEP) consists of maximizing the average information content imposing on the system a constant average amount of effort (denoted *F*). This latter has been used by Kantor (Kantor 1998) in *Information Retrieval* for modelling information search situations. The principle of the least effort (denoted here PLE), attributed to Zipf (Zipf 1949) in linguistics, consists of minimizing the average amount of effort imposing on the system an average information content. Intuitively, we can say that the MEP consists of choosing the maximum profit solution from among a set of situations requiring the same production effort. Whereas the PLE chooses the solution that minimizes the effort from among a set of solutions giving the same profit. L. Egghe and T. Lafouge have shown (Egghe 2006) that these two principles are equivalent for discrete, finite and decreasing distributions, which we often encounter in informetrics.

4. Exponential Informetric Process

4.1 Continuous distributions

When we mathematically formalize informetric processes, two representations are possible: the discrete mode or the continuous mode. In the preceding, we used a discrete representation to define a stochastic process We then chose to work in continuous mode in order to generalize the results. We define two functions: a density function and an effort function.

In all the following, we shall denote by ν a density function modelling any stochastic process. We suppose that this is defined on the interval $[1..\infty]$ and that the necessary condition for standardization, $\int_1^\infty \nu(t)dt = 1$ [1] is verified.

We introduce effort function f also defined on the interval $[1..\infty]$, positive increasing and not bounded, which verifies the following condition: $\overline{F} = \int_1^\infty \nu(t).f(t)dt \leq \infty$ [2].

This second condition signifies that the average amount of effort to produce all the items is finite. The functions ν and f define what we call in this article an informetric process. These two distributions are not independent. It is natural to think that we can express the production according to the effort. This is what we are going do now by defining an Exponential Informetric Process.

4.2 Definition of an Exponential Informetric Process

Let f be a positive function defined on $[1..\infty]$ and *a*, a positive number greater than 1. We define an Exponential Informetric Process $\nu(f,a)$ by: $\nu(f,a)(t) = k.a^{-f(t)}$ $k > 0$ where f is an effort function.

Condition [2] is then written $\int_1^{\infty} f(t).a^{-f(t)}\,dt \leq \infty$ [3]. The effort f is increasing, not bounded, so we can easily show that $v(f,a)(t) = k.a^{-f(t)}$ verifies the condition [1] of standardization $\int_1^{\infty} a^{-f(t)}\,dt \leq \infty$ and that it is now possible to calculate the constant of standardization k. The geometric and power distributions that we currently encounter in bibliometry are represented by this simple model as we see later.

4.3 Exponential Informetric Process and entropy

We shall now show that an exponential process thus defined verifies the two preceding principles, the MEP and the PLE, and that we have a simple relationship between amount of effort and information content. In continuous mode, if a process is defined by its density function v, its entropy $\overline{H}$ is calculated by the formula: $\overline{H} = -\int v(t).Log\,(v(t))\,dt$. Contrary to the discrete mode, entropy $\overline{H}$ is not necessarily positive.
Firstly, let us recall the mathematical formula of these two principles for a stochastic process
.

Maximum entropy principle (MEP)

The MEP consists of maximizing the entropy, meaning the function H $H(v) = -\int v(t).Log\,(v(t))\,dt$ knowing that $\int v(t)dt = 1$ and $\int f(t).v(t)dt = \overline{F}$ [4] and where f is a given positive function (effort function) and $\overline{F}$ a fixed constant (average amount of effort).

Principle of the least effort (PLE)

The PLE consists of minimizing the effort, meaning the function F, $F(v) = \int v(t).f(t)dt$ where *f is* a given effort knowing that $\int v(t)dt = 1$, and $\int v(t).Log\,(v(t).dt = \overline{H}$ where $\overline{H}$ is a given constant (average information content).

We then have the following results, which characterize an Exponential Informetric Process.

Theorem: Exponential Informetric Process, MEP and PLE

With an Exponential Informetric Process, $v(f,a)(t) = k.a^{-f(t)}$ $k > 0, a > 1$ f an effort function (increasing and verifies the condition [3]) we have the following properties:

(a) $v(f,a)$ is decreasing.

(b) The two principles, maximum entropy and least effort are verified simultaneously.

(c) If $\overline{H}$ and $\overline{F}$ describe the average information content and effort produced by the process, we have the following proportional relationship: $\overline{H} = -Log\,(k) + Log\,(a).\overline{F}$ [6].

Proof

Note: we no longer specify the interval of variation of v and f which is $[1..\infty]$.

Demonstration of (a)

We can easily show that $\boldsymbol{v(f,a)}$ is a decreasing density function because we have $a > 1$ and f increasing.

Demonstration of (b) and (c)

- For the MEP

Next [3] $v(f,a)$ verifies the condition [4], let us put $F = \int f(t).k.a^{-f(t)}dt$

Let us show that H reaches its maximum for the function $v(f,a)$. Let G be the following function: $G(t,v) = v\,Log\,(v) + \lambda v + fv.Log\,(a)$ where λ is a constant whose value is: $\lambda = -1 - Log\,(k)$.

We have: $\frac{\partial}{\partial v}G(t,v) = Log\,(v) + 1 + \lambda + f.Log\,(a)$

So we can easily show that the derivative is cancelled for $v(f,a)$ (to simplify, we then denote $v(f,a)$ by v_f)

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

G being convex and $\frac{\partial G}{\partial v}$ cancelling whatever the value of t fixed, for any v function. Checking [4] we can then write: $G(t,v) \geq G(t,v_f)$.

Hence $\forall v$, $v\,Log\,(v) + \lambda v + fv.Log\,(a) \geq v_f\,Log\,(v_f) + \lambda v_f + f.v_f.Log\,(a)$

or $\forall v$, $\int (v\,Log\,(v) + \lambda v + f.vLog\,(a))dt \geq \int (v_f\,Log\,(v_f) + \lambda v_f + f.v_f Log\,(a))dt$

Finally, we have the result: $\int (v\,Log\,(v))dt \geq \int (v_f\,Log\,(v_f))dt$

- For the PLE

To verify the condition [5] of the PLE, let us calculate the value of the entropy $\overline{H}$:

We have $\overline{H} = -\int v_f.Log\,(v_f).dt$

$= -\int k.a^{-f}.Log(k.a^{-f})dt = -Log\,(k).k.\int a^{-f}dt + Log\,(a).\int f.ka^{-f}dt = -Log\,(k) + Log\,(a).\overline{F}$

This calculation demonstrates the condition (c) of linearity.

Let us demonstrate that EF reaches its minimum for the function v_f. Let G, the following function $G(t,v) = v\,Log\,(v) + \lambda v + fvLog\,(a)$ where λ is a constant with the value $\lambda = -1 - Log\,(k)$. As for the preceding case, we can easily conclude. (In the case of the PLE, the condition $a > 1$ is necessary to conclude. We will find the case $a < 1$ examined in the article already quoted (Egghe 2004) for the finite discrete case). We note that the condition of decrease is not necessary to demonstrate the result (b) and (c).

4.4 Examples

Note here we will use $a = e$. All the results are valid for any number $a > 1$.

The geometric and power distributions that we currently encounter in informetrics can be represented by this simple model.

- Geometrical model: the effort function is the linear function $f(t) = \alpha(t-1)\ \alpha > 0, t \geq 1$

The exponential informetric corresponding process is then written: $v(f,a) = \alpha.e^{-\alpha(t-1)}$ The entropy is equal to $H(\alpha) = 1 - Log\,(\alpha)$.

- Power model: the effort function is a logarithmic function, $f(t) = (\alpha+1)Log\,(t)\ \alpha > 0, t \geq 1$ is then the exponential informetric corresponding is then written: $v(f,a) = \alpha.e^{-(\alpha+1).Log\,(t)}$ The

form used is in general: $\nu(f,a)(t)=\frac{\alpha}{t^{(\alpha+1)}}$. In this case the entropy (Yablonsky 1981) is equal to $H(\alpha)=1-Log\ (\alpha)+\frac{1}{\alpha}$

We note that in both cases, the entropy is a decreasing function of α . The interpretation of the law of Lotka has been verified. The greater α , the greater the gap between the small number of researchers who publish a lot and the large number of researchers who publish little.

- Mixed case

In this case the effort function is composed of two functions: the first one is linear (effort constant) and the other logarithmic (least effort law). The effort function is : $f(t)=\alpha.t-(j-1).Log\ (t)\ \alpha>0,\ j=1,2,.....,\ t\geq 0$. The exponential informetric process corresponding process is : $v(t)=\alpha^{j}\frac{t^{j-1}}{(j-1)!}.e^{-\alpha t}$. The reader will find more details in (Lafouge 2001).

- Another example

The effort function is : $f(t)=k.t^{\alpha}\ \ \alpha>0\ \ \ k>0.\ t\geq 1$

The exponential corresponding process is then written : $\nu(f,\ a)=e^{-kt^{\alpha}}$. We must show that $\overline{F}<\infty$ .It is easy to show the condition [3] is verified for we have:

: $\int_{1}^{\infty} t^{\alpha}e^{-t^{\alpha}}\ dt\ =\frac{1}{\alpha}.\Gamma(\frac{\alpha+1}{\alpha},\ 1)$

where Γ is the Gamma function : $\Gamma(a,b)=\int_{b}^{\infty}e^{-t}.t^{\ a-1}.dt\ \ a>0\ \ b\geq 0$

In the geometrical case, the effort function indicates the fact that the production of each item requires on average the same amount of effort. Whereas in the power case, the effort function, which is concave, means that the production of each item requires less and less effort. This last property enables us to say that a power distribution is an exponential process with a logarithmic effort function. This characterization clarifies, without using the powerful result of Price on the cumulative advantages process, the principle of least effort, which is expressed by the properties of the logarithmic function. This property, with that of invariance of scale, (Egghe 2005) gives *Lotkaian informetrics* all its force.

References

Bradford, S. (1934). Sources of information on specific subjects. In *Engineering*, 137, pp 85-88.

Dominich, S., Goth, J., Kiezer, T. and Szlavik, Z. (2004). An Entropy-Based Interpretation of Retrieval Status Value-Based Retrieval, and its Application to the Computation of Term and Query Discrimination Value. In *Journal of the American society for Information Science and Technology,* 55(7), pp 613-627.

Egghe, L. (1990). On the duality of Informetric systems with applications to the empirical law. In *Journal of Information Science*, 16, pp 17-27.

Egghe, L. (2005). *Power* Laws in the information production process: Lotkaian informetrics, Elsevier, Oxford.(to be published in 2005)

Egghe, L. and Lafouge, T. (2004). On the Relation Between the Maximum Entropy Principle and the Principle of Least Effort. In *Mathematical and Computer Modelling*, (to be published in 2006).

Kantor, P.B. and Jung, J.L. (1998).Testing the maximum entropy principle for information retrieval. In *Journal of the American Society for Information Science* 49(6) 1998, pp 523-527.

Lafouge, T., and Michel C. (2001). Links between information construction and information gain. Entropy and
distributions. In *Journal of Information Science*, 27(1), pp 39-49.

Lafouge, T. (2003). Information et théorie mathématique: une impasse en Science de l'Information?. In *Decision Making* 6 Mars 2003.
http://lepont.univ-tln.fr/isdm/PDF/isdm6/isdm6a34_lafouge.pdf

Lotka, A.J. (1926). The frequency distribution of scientific productivity. In *Journal of the Whashington Academy of Sciences*. 16, pp 317-323.

Price, D.S. (1976). A general Theory of Bibliometric and other Cumulative Advantage Processes. In *Journal of the American Society for Information Science*, pp 292-306.

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

Yablonsky, A.L. (1981). On fundamental regularities of the distribution of scientific productivity. *Scientometrics* 2(1), pp 3-34.

Zipf, G.K. (1949) Human Behaviour and the Principle of Least Effort. Addison-Wesley, Cambridge, Massachusetts, USA, 1949. Reprinted: Hafner, New York, USA, 1965.