Journal of Informetrics 11 (2017) 1003–1015

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Regular article

A remarkable example in three-dimensional informetrics. The geometric law: Distribution of use or distribution of structure?

Abdellatif Agouzal[a,b], Thierry Lafouge[a,c,*]

[a] Université de Lyon, France
[b] Université Lyon 1, Institut Camille Jordan, CNRS U.M.R. 5208, Villeurbanne, F-69622, France
[c] Université Lyon 1, ELICO, Villeurbanne, F-69622, France

ARTICLE INFO

Article history:
Received 25 January 2017
Received in revised form 25 June 2017
Accepted 11 September 2017

Keywords:
Geometric law
Three-dimensional informetrics
Probability generating function
Distribution of use
Distribution of structure

ABSTRACT

This article offers a mathematical model that is in line with three-dimensional informetrics. We show that the geometric distribution – often found in informetric processes of use to model the distributions of book loans in libraries, of citations, or of journals downloads in scientific databases – plays a surprising role. To show this, we infer the existence of another distribution, which we have called "distribution of structure". Distributions of use and distributions of structure are the two key processes in our three-dimensional IPP (Information Production Process). Furthermore, we indirectly introduce the time dependence of these distributions through the concept of information obsolescence. To demonstrate the theorems, we use the properties of probability-generating functions (PGF).

1. Introduction and background

To describe an informetric process, we use a general framework that consists in studying a population of sources which randomly produces items over time. Many examples have been studied in the past decades. We mention two historic examples: the production of articles by researchers (Lotka, 1926), and the contributions of research articles in a particular field in scientific journals (Bradford, 1934). A framework, referred to as IPP (Information Production Process) was defined (Egghe, 1990) to study such informetric processes in which the population is itself a set of sources[1] (researchers, articles, web pages...) producing items (articles, citations, links...). These historic examples do not explicitly consider time dependence in their mathematical formulation of the processes. They are stationary distributions in the form of counts (number of articles, number of citations...) observed over a predefined time period. Probabilistic laws are fitted to the empirical distributions. Their properties have been known for a long time in the field of scientometrics (Haitun, 1982). Almost all of the distributions are aggregative (above-average variance), decreasing, and are generally long-tailed. Certain researchers have implemented a stochastic model to take the time dependence explicitly into account and explain such phenomena more clearly. We can, for example, cite the works of Burrell on bibliometric processes (Burrell, 1988). It is also important to mention the works of Price in which the time unit (Price, 1976) is taken into account. The cumulative advantage model (a variant being Success-

* Corresponding author at: Université Lyon 1, ELICO, Villeurbanne, F-69622, France.
E-mail address: lafouge.thierry@gmail.com (T. Lafouge).

[1] It would be more precise to talk about generalized bibliographical sources.

http://dx.doi.org/10.1016/j.joi.2017.09.001
1751-1577/

Breeds-Success (SBS)) describes the evolution of an IPP in time. In Egghe and Rousseau (1995), a model generalising Price's model (general SBS) is defined.

To the best of our knowledge, there are no historic laws – such as Lotka's or Bradford's – that deal with the use of documents. However, the statistical regularities in the distributions of book loans in libraries were observed early on. Such studies were undertaken to improve library management. Predictive tools, which originated from operational research (Morse, 1968) were set up. In his book, Morse chose the Poisson process when implementing Markovian matrices to regulate the circulation of books in libraries. The negative binomial (Bagust, 1983) law was chosen to model stationary distributions of library book loans (Leemans, Maes, Rousseau, & Ruts, 1992). Stochastic models, which depend on several parameters and which take explicit consideration of time dependence, were used (Burrell, 1990; Burrell & Fenton, 1993).

In the digital era, borrowing the physical copy of a document is not as crucial as it used to be. Uses are now more of interest. However, statistical regularities linked to uses are still relevant, and mathematical models are still effective. Recently, studies (Ajiferuke & Famoye, 2016) used very different data sets that represented four broad informetric subfields where different counting models were tested.

These authors created statistics based on variables that are used in altmetrics (Priem & Groth, 2012) such as: statistics for the number of views, statistics for the number of readers.

Distributions of use related to citations are still largely studied and constitute invaluable indicators – notably citations in the evaluations of research. A multitude of types of distributions (Exponential, Weibul, Log-normal, Yule distribution. . .) have been tested to adjust these distributions. A state of the art of these various studies can be found in (Bertoli-Barsotti & Tommaso, 2015). This study leads the authors to suggest a new formula to calculate the h-index, based on the geometric law, to model the distributions of citations. Burrell (2014) noticed that the distribution of citations of three renowned researchers in informetrics followed a geometric law. The body of work is admittedly small, but the result is nevertheless surprising.

The "No Use" sources, which are non-producing sources, are not *a priori* excluded from the framework of an IPP, but they are not studied as a stand-alone element. Taking into account these "non-producers" often modifies the type of model used for the adjustment. For example, in the study of loans (Burrell, 1980), this factor was taken into account. Egghe and Rousseau (2012) described a Lotka distribution which includes non-producers (the so-called shifted Lotka function). We must note that, in practice, the definition of an IPP generally implies that the entire set of sources is defined with the help of the produced items. Taking into account the affiliated time factor and "No Use" sources is, of course, crucial in the analysis of citations and in many other cases. In longitudinal studies, over time, the set of sources varies (some entering, some leaving). To the best of our knowledge, there are few theoretical and/or applied studies that take into account the time factor.

For a long time, some studies noticed that two IPPs could be linked in a natural way. These IPPs are called Three-dimensional Information Production Processes (Egghe, 2005, chap. 3). Rousseau (1992) considered the case in which researchers who published articles then also received citations. Burrell (1992) defined a stochastic model to study this type of problem: a population of researchers who publish articles over time and who subsequently receive citations. Burrell suggested a model in which he counted the number of published papers of a publishing author during [0,t], with a time-dependent geometric distribution.

Looking at data from the journal provider, we studied the demand for scientific articles (Lafouge, 1998) by researchers.

This experimentation led us to define a three-dimensional IPP with the following linear framework: journals produce scientific articles by including them in volumes, and the scientific articles are then requested by researchers.

We then conducted several theoretical studies (Lafouge, 2001; Lafouge & Guinet, 1999; Lafouge & Lainé Cruzel, 1997). The results from these articles are re-explained in the present paper using a more general method. To the best of your knowledge, few theoretical or practical studies on three-dimensional informetrics exist.

The aim of this article is to revisit the three results mentioned above. We had shown, under certain conditions, that if the distribution of article production in journals was of a certain type (Poisson, geometric. . .), then the distribution of use was of the same type. Here, we are building a generic model that is in line with the three dimensional IPP. It uses the properties of the probability generating function. We introduce time dependence in a non-explicit way by having the proportion of sources that no longer produce items (No-use) – those that are no longer used – tending towards 1. Such a reality illustrates what is known as information obsolescence.

Our article is organized in three parts:

- Defining the problem: we describe the informational process in which our study operates (see Section 2).
- Defining the general theory (see Section 3).

This section is divided in two subsections. In the first part (see Subsection 3.1), we study a stationary problem in which the geometric law is highlighted (see Theorem 3.6). In a second section (see Subsection 3.2), we indirectly introduce the time dependence (see Theorem 3.7).

- Discussion and conclusion (see Section 4).