Journal of Informetrics 8 (2014) 136–146

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regularity in the time-dependent distribution of the percentage of never-cited papers: An empirical pilot study based on the six journals

Zewen Hu[a,b], Yishan Wu[a,*]

[a] Institute of Scientific & Technical Information of China, Beijing, China
[b] School of Information Management, Nanjing University, Nanjing, China

ARTICLE INFO

Article history:
Received 5 May 2013
Received in revised form 4 November 2013
Accepted 4 November 2013
Available online 1 December 2013

Keywords:
Percentage of uncited papers
Citation distribution
Regularity in time-dependent distribution
Influencing factors
Negative exponential model
Journal

ABSTRACT

The non-citation rate refers to the proportion of papers that do not attract any citation over a period of time following their publication. After reviewing all the related papers in *Web of Science*, *Google Scholar* and *Scopus* database, we find the current literature on citation distribution gives more focus on the distribution of the percentages and citations of papers receiving at least one citation, while there are fewer studies on the time-dependent patterns of the percentage of never-cited papers, on what distribution model can fit their time-dependent patterns, as well as on the factors influencing the non-citation rate. Here, we perform an empirical pilot analysis to the time-dependent distribution of the percentages of never-cited papers in a series of different, consecutive citation time windows following their publication in our selected six sample journals, and study the influence of paper length on the chance of papers' getting cited. Through the above analysis, the following general conclusions are drawn: (1) a three-parameter negative exponential model can well fit time-dependent distribution curve of the percentages of never-cited papers; (2) in the initial citation time window, the percentage of never-cited papers in each journal is very high. However, as the citation time window becomes wider and wider, the percentage of never-cited papers begins to drop rapidly at first, and then drop more slowly, and the total degree of decline for most of journals is very large; (3) when applying the wider citation time windows, the percentage of never-cited papers for each journal begins to approach a stable value, and after that value, there will be very few changes in these stable percentages, unless we meet a large amount of "Sleeping Beauties" type papers; (4) the length of an paper has a great influence on whether it will be cited or not.

1. Introduction

In the scientific community, there exists frequent uncitedness to the mediocre, the low quality, the unintelligible, the irrelevant, the valuable but undiscovered or forgotten, the par excellence, and the well known documents (Garfield, 1973). Price (1965) estimated that about 35% of all published papers in 1961 were not cited at all in any given year and 10% were never cited over a period of ten years. Koenig (1983) reported a 50% uncitedness rate in the pharmaceutical literature in a

* Corresponding author. Permanent address: Institute of Scientific & Technical Information of China, 15 Fuxing Road, Beijing 100038, China. Tel.: +86 010 13501127855.
E-mail addresses: wuyishan1957@163.com, huzewen915@163.com (Y. Wu).

1751-1577/$ – see front matter
http://dx.doi.org/10.1016/j.joi.2013.11.002

citation time window of four years following publication. It is an accepted fact of academic life that some papers, in whatever discipline and wherever published, will never be cited (Burrell, 2002; Egghe, 2000; Rousseau, 1994).

Though papers that are less frequently cited in a shorter citation time window may be of great value that has not been found and utilized due to various reasons, if they fail to attract any attention from the peers over a long period of time following their publication, then they might have weakness in relevance, importance, popularity, novelty, quality or impact. Of course, there are also some exceptions, for example, some "Sleeping Beauties" papers that remain uncited in a long period of time and subsequently receive several citations, as if sleeping beauties are awakened (Burrell, 2005; van Raan, 2004). To certain degree, the percentage of papers never attracting any citation over a long period of time following their publication may be taken as a negative indicator to evaluate the quality of journals and the academic impact of researchers, organizations or countries. MacRoberts and MacRoberts (2010) also thought that an evaluative citation analysis should take the uncited work into account. A decreasing function relationship between impact factor of journals and their uncitedness rate has been verified by van Leeuwen and Moed (2005), Egghe (2008, 2010) and Hsu and Huang (2012).

Studies on uncitedness have become an important research topic in Scientometrics field. Some relevant classical reviews (Garfield, 1972, 1991, 1998; Hamilton, 1990, 1991) bring us up to date and some empirical studies (Abt, 1981; Bott & Hargens, 1991; Egghe, Guns, & Rousseau, 2011; Ghosh & Neufeld, 1974; Ghosh, 1975; Hopson, 2008; Schwartz, 1997; Sen & Patel, 2012) have investigated and compared the percentages of non-citation papers in fixed citation time window for certain domains or journals.

However, there are very few scholars who take the percentage of never-cited (or unused) papers in a fixed collection of publications as a whole and observe its development patterns in a series of citation time windows with different length t $[0,\infty]$ following their publication, as well as study what distribution model can fit their time-dependent patterns.

Fortunately, some scholars have earlier explored the future development (or use) patterns of books or papers over time due to the obsolescence and developed some exponential decay functions to model the exponential decay process of the number of cited (or used) books or papers as time elapses. In particular there is a paper written by Gosnell (1944), who studied the publication dates of titles included in the three selected book lists and found some use patterns of books over time following their publication. For instance, there is a rapid rise in number of used titles for the first few years preceding the date of issue of each respective list. Generally speaking, the maximum number of used titles per year occurs within three years preceding the publication of each list. After this maximum is reached, the number of used titles per year drops rapidly at first, and then drops more slowly, approaching the base line of zero asymptotically. The initial rise may be ascribed to the lag in selection of books. This lag is due to the delay in appearance of reviews and acceptance by scholars, while the subsequent drop may be ascribed to the obsolescence of books. As no satisfactory models were found to fit this special distribution curve of rising at first, then dropping, Gosnell dropped data for the rising portion of the curve for the first few years and only examined the curve of decline after the early maximum was reached. Then he proposed a function to model this kind of decay curve over time:

$$y = y_0 b^t \tag{1}$$

where y is the number of used titles, y_0 the number of used titles at the maximum or initial point with lag eliminated, and t is time elapsed. The omega $\omega = 1 - b$ becomes the annual rate of decrease in the curve, or the rate of obsolescence.

Following that, many scholars also studied in depth the future use pattern (e.g. citation pattern) of literature over time in face of aging following their publication (Alvarez, Escalona, & Pulgarín, 2000; Brookes, 1970a,b; Cole, 1963; Egghe & Ravichandra, 1992; Egghe, 1993, 2001; Griffith, 1979; Gupta, 1990; Line, 1970; Line & Sandison, 1974; Oliver, 1971; Stinson & Lancaster, 1987; Yu & Li, 2010). Among them, Brookes gave out the concrete proof of Gosnell' model. Griffith (1979) applied such model to determine obsolescence for all the scientific literature cited in the SCI up to that date. Gupta's analysis of citations to the 1983 volume of Physical Review (1990) showed a decay curve of citation over time, which can be fitted by an exponential model. So, under the condition of assuming that there is an exponential decay of citation over time due to the obsolescence when the effect of citation lag is eliminate, a classic negative exponential decay model can be obtained as follows:

$$f(t) = K \cdot e^{-at} \tag{2}$$

In Eq. (2), t is the age of literature, K is a constant, and a is the aging coefficient. Interestingly, the two models (Eqs. (1) and (2)) are equivalent in theory, Yu and Li (2010) deduced Eq. (1) from Eq. (2). For further detail, we advise readers to see the original papers.

From the above description about the development patterns of the number of cited literature over time following their publication, we can see that such curves show common features: "dropping rapidly at first, and then dropping more slowly, approaching the base line of zero asymptotically, when the effect of citation lag is excluded." Although there is a very lager difference in expressing two terms: the number of cited literature and the percentage of never-cited literature, there is similar development patterns. For example, the time-dependent pattern of the percentage of never-cited papers being studied here also presents a similar distribution curve. In initial citation time window approaching to zero, nearly all the publications are in a non-cited status due to the lag of citation. Along with the rapid increase of citation, the percentage of non-cited publications drops rapidly at first, and then just like the case of fast obsolescence of literature, those never-cited publications in previous years have smaller chance to get cited in subsequent years, so the percentage of never-cited publications drops more slowly over time, approaching the base line of zero asymptotically after going through an infinitely