Bibliometrics for Dummies

Tina Jayroe

University of Denver

Dottore Shimelis G. Assefa, PhD
Informazioni Scienza
23 settembre 2008

Bibliometrics (or bibliometry) is included with other mathematical and statistical analysis methods under the umbrella of informatics/information science. Bibliometrics has been used for almost a century as a descriptive and evaluative science.[1] The term “bibliometrics” was coined by Alan Pritchard in 1969. Prior to this, the field had been labeled “statistical bibliography” by British librarian E. Wyndham Hulme in 1922 (Diodato, 1994, p. 154).

Bibliometrics can overlap with informetrics, the mathematical and statistical study of patterns in documentation and information; and scientometrics, the mathematical and statistical analysis of research patterns in life and physical sciences. It is associated with cybermetrics, the study of the quantitative analysis of scholarly and scientific communications in the Internet. And even cliometrics, the study of historical data by use of statistical techniques, uses some of its aspects because “the study of history involves the study of information” (Diodato, 1994, p. 42).

Philosophically, bibliometrics contributes to a better understanding of the information universe because it is the study of how humans relate to information (Bates, 1999, p. 1048). Ronald Rousseau states that “bibliometric (or informetric) research has developed a rich body of theoretical knowledge, which in turn aims at finding wider applications in all practical aspects of information work” (1990, p. 197).

In library science, bibliometrics usually refers to the analysis of patterns of data and information as it pertains to literature—especially scholarly authorship, publication, and communication—by employing certain methods and “laws.” It has a significant impact in the area of information retrieval, and now scientists are using bibliometrics to study link structures in the Internet, also known as webometrics. Ultimately, all applications of bibliometrics help to make meta-information explicit in the paradigm of information science. This paper focuses on the three major bibliometric laws, citation analysis, and a brief look at webometrics.

The major empirical laws of bibliometrics are Bradford’s law, Lotka’s law, and Zipf’s law. They are “the ones that have been used most often in the literature to model the distribution of information phenomena” (Chen & Leimkuhler, 2006, p. 308). While all three laws result in similar probability distributions, the differences lie in the data being analyzed. These are what Dr. Virgil Diodato calls *items* and *sources:*

> In a Bradford analysis, the items are often journal articles, and the sources are then the journals that produce the articles. . . . In a Lotka analysis . . . the sources are authors, and the items are

> the documents that the authors produce. In Zipf's law, the sources are the text, or more precisely the rank of the words in the text, and the items are the numbers of occurrences that each ranked word produces (1994, p. 93).

In library and information science these are applied and analyzed in order to predict patterns of recorded information and communications. Libraries can then develop theories on why certain structures and patterns occur, subsequently creating practices that enable the best use of information behavior based on those findings.

Bradford's Law

In 1934 mathematician and librarian Samuel C. Bradford came up with the bibliometric formula **1: *n: n²***, which basically states that most articles are produced by few sources and the rest are made up of many separate sources. This law can also be referred to as *Bradford's Distribution*, *The Law of Scatter*, and *The Law of Scattering* or simply *core and scattering* (Diodato, 1994, p. 24). The reason for this labeling is that when the data are demonstrated visually, there is a cluster (or nucleus) near the journal that produces the most articles on a given topic. When displayed in relation to the topic, the distribution for the remaining data tends to be scattered among the many journals that produce significantly fewer articles. Bradford gives these data their own mathematical equations and designates them to fall within certain "zones."

To enter these data into a graph would result in what is called a *Bradford Curve*, which looks like the letter "J" or "S" when displayed on a chart. Bradford's law is thought to be the basis of the *80/20* rule in librarianship—80% of the items used are 20% of the library's collection (Raber, 2003, p. 73).

Lotka's Law

Formulated by Alfred J. Lotka in 1926, it is also known as the *inverse square law* or *inverse exponential law,* and is used to determine the most prolific authors within a discipline. It usually proves that there are only a few authors who contribute to a field as a publication increases in quantity.

Lotka was a mathematician. While working at the Statistical Bureau of the Metropolitan Life Insurance Company, he started counting the names of authors and compared them to the numbers of publications in *Chemical Abstracts* between 1907 and 1916 (Hertzel, 2003, p.303). From this he devised

the formula $\mathbf{x^n y = c}$ "where: y is the portion of authors making x contributions each; n and c are parameters that depend on the field being analyzed" (Diodato, 1994, p. 105, 106).

Lotka's law has been highly applicable to various areas of publication but especially in determining "patterns of productivity among chemists" (Chen & Leimkuhler, 1986, p. 307). It has been debated whether it is applicable to librarianship (Hertzel, 2003, p. 304, 305), but is definitely useful in the area of collection development and information retrieval (Raber, 2003, p. 72).

Zipf's Law

Zipf's law, sometimes called *Zipf's Law for Words of High Frequenc*y, or the *Principle of Least Effort* is based on natural-language corpus and the frequency in which a word is used in a text. It is named after philologist George Kingsley Zipf and actually consists of two laws. Specifically, it is the first law that analyzes the frequency/distribution of a word in a text and then uses that analysis to determine its rank among other frequently used words within the text. Zipf's second law is a formula for words with low frequencies (Diodato, 1994, p. 167, 168).

The formula for Zipf's law is ***rf = c*** "where *r* is the rank of a word-type, *f* is the frequency of occurrence of the word-type, and *c* is a constant, dependent on the corpus" (Wyllys, 1981, p. 1). Using Zipf's law, one can determine words of less value such as "a" "the" or "of" and calculate the frequency of rarer words, thereby determining a text's true aboutness.[2]

According to Andrew D. Booth (who revised Zipf's law in 1967 to include a count of all the words in a text), J. B. Estroup was the first to state the formula in 1916, however it was Zipf who made it popular (Diodato, 1994, p. 15).

Citation Analysis

Citation analysis encompasses many quantitative measurements that are crucial to: enhancing document retrieval; determining aboutness; and revealing relationships between texts, subjects, and authors. Some of the methods used are: citation age (or obsolescence); co-citation**;** impact factor; bibliographic coupling; self citation; clustering; allocitation; and many more.

Scientist Eugene Garfield built off of Bradford's work in the 1960s by creating a citation index which showed how works are often cited or referenced by a few; how they tend to cluster around a

publishing source; and most importantly, that the amount of citations does not necessarily correlate with the significance of a work to its discipline. He termed this research a citation's *impact factor*—a measurement for determining the quality of a work regardless of how many times it has been cited (Hertzel, 2003, p. 319, 320).

While there are many specific reasons for employing a citation tool, one purpose of citation analysis is to gain a better understanding of bibliographic distribution such as: which authors are being cited and by whom; from what disciplines and organizations; and in what countries. These data can reveal pertinent information about productivity within a given field. Citation indices such as Institute for Scientific Information's *Journal Citation Reports,* the *Arts & Humanities Citation Index,* the *Science Citation Index,* and the *Social Sciences Citation Index* journals all track and publish such information.

Since citing can be complicated, citation analysis requires serious evaluation and cannot always be taken at face value. For example, an author may cite another author, but the context could be negative; the second or third author may not be cited at all, even though they made significant contributions to an article; or the citation may not be relevant to the work (Raber, 2003, p. 78).

Another tricky area in citation analysis is allocitation, which is the opposite of self-citation [*I still need to elaborate on this*]. And then there is the *Matthew effect* (or biased citations), which identifies the concept "success-breeds-success or the rich-get-richer phenomenon. In scholarship this occurs when already well-known individuals receive disproportionately high recognition for their new work compared to the relatively low recognition received by lesser known colleagues who do comparable work" (Diodato, 1994, p. 100).

Webometrics: Link Analysis[3]

Commercial search engines and web crawlers have enabled an environment where bibliometrics can be applied to study information patterns on the World Wide Web. Since the Web contains recorded information, areas such as hyperlinking can be measured for scholarly communication, much like journal citation analysis.

With online computations there exists an ability and a benefit in studying social networks and other collaboration structures.[4] Björneborn and Ingwerson draw an analogy (and strongly assert it is *only* an analogy) to using inlinks, outlinks, and selflinks, etc. in the Internet much the way citation, co-citation,

and self-citation are used among journals for gathering analytical data (Thelwall, M., Vaughan, L., & Björneborn, L., 2005, p. 84).

Conclusion

Whether it is link analysis, citation analysis, applied mathematics, or theoretical interpretation, bibliometrics contributes to understanding the phenomenon of information by revealing patterns and structures of data in relation to recorded information and communications. Once these data become prevalent and analyzed, they can provide better contexts and tools to enhance organizational evaluations, information retrieval, system design, social behaviors, and human knowledge.

References:

Bates, M. (1999). The invisible substrate of information science [Electronic version]. *Journal of the American Society for Information Science, 50*(12) 1043–1050.

Björneborn, L. & Ingwerson, P. (2004). Toward a basic framework for webometrics [Electronic version]. *Journal of the American Society for Information Science, 55*(14) 1216–1227.

Chen, Y., & Leimkuhler, F. (1986). A relationship between Lotka's law, Bradford's law, and Zipf's law [Electronic version]. *Journal of the American Society for Information Science, 37*(5) 307–314.

Diodato, V. (1994). *Dictionary of bibliometrics*. Binghampton, NY: The Haworth Press, Inc.

Hertzel, D. H. (2003). Bibliometrics History. In Drake, M. (Ed.), *Encyclopedia of library and information science* (pp. 288–328). New York, NY: Marcel Dekker, Inc.

Raber, D. (2003). *The problem of information: An introduction to information science*. Lanham, MD: Scarecrow Press, Inc.

Rousseau, R. (1990) Relations between continuous versions of bibliometric laws [Electronic version]. *Journal of the American Society for Information Science, 41*(3) 197–203.

Thelwall, M. & Vaughan, L. (2004). Webometrics: An introduction to the special issue [Electronic version]. *Journal of the American Society for Information Science, 55*(14) 1213–1215.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics [Electronic version]. *Annual Review of Information Science and Technology, (39)*1 81–135

Wyllys, R. (1981). Empirical and theoretical bases of Zipf's law [Electronic version]. *Library Trends, 30*(1) 53–64.

Notes

[1] Reousseau's bibliometric timeline. Retrieved September 19, 2008 from http://users.telenet.be /ronald.rousseau/html/timeline_of_bibliometrics.html

[2] "Zipf's law is primarily about the pattern of word occurrences/distribution in a corpus. That authors tend to use a select core of words time and again. When a given corpus (body of text) is parsed and ranked by the number of occurrences, and you multiply the rank by the number of occurrences, the result is an inverse relationship with a constant. I think, having the knowledge of the highest frequency of words in a corpus is a clue to what that text is about. I would argue, you can use that knowledge to gain an insight of what that particular piece is all about. However, the initial thesis of the Zipf's law is more of economics in that the principle of least effort (after Zipf's original book, - Human Behavior and the Principle of Least-Effort) comes into play" (Dr. Assefa, personal communication, September 22, 2008)

[3] Because "Webometrics encompasses all quantitative studies of Web related phenomena" (Thelwall & Vaughan, 2004, p. 1213), I was only able to touch on link analysis in the interest of time, page limit, and the fact that I am not a computer scientist ☺.

[4] The exception of temporary communications such as online chat rooms goes beyond webometrics and falls within the field of cybermetrics (Thelwall, M., Vaughan, L., & Björneborn, L., 2005, p. 84).