Olsgaard, John N. (ed.). 1989. Principles and applications of information science for library professionals. Chicago: American Library Assoc.

CHAPTER

2

Bibliometrics and Citation Analysis

Danny P. Wallace

Bibliometrics is the application of quantitative methods to the study of information resources. Although work that is now recognized as being bibliometric in nature was conducted more than seventy years ago, the term itself dates only from 1969, when it was proposed by Alan Pritchard as a replacement for the earlier term "statistical bibliography."[1] Two other areas of study that are very closely related to bibliometrics are scientometrics, which encompasses all quantitative analyses of scientific productivity, and citation analysis, which refers to the practices and patterns of scholarly references.

The major focus of bibliometrics is the search for regular patterns related to the characteristics of information sources. Some of the questions asked by bibliometric studies are:

- In what countries is the literature of a particular subject or discipline published, and what is the balance of the contributions of those nations?
- In what countries do the authors who contribute to a particular subject literature work, and what is the balance of the contributions of those nations?
- What languages are most used in publications on a particular subject?
- How are words used in publications, and what patterns describe their use?
- What types of information sources (books, articles, theses, etc.) are most important to a particular subject?
- What research methodologies (historical, case study, experimental, survey, etc.) are most used in a particular subject's research information sources?
- What types of articles (research report, opinion piece, news story, etc.) make up the periodical literature of a subject?
- What patterns pertain to age of information sources at the time they are used?
- What is the distribution of authors' contributions to a literature, and why do some authors publish more than others?

The author wishes to thank Joan Giglierano for her assistance in the preparation of this chapter.

1. Alan Pritchard, "Statistical Bibliography or Bibliometrics?" *Journal of Documentation* 25:348–349 (December 1969).

- What is the distribution of articles on a particular subject among the journals in which they are published?
- What patterns apply to the circulation of items within libraries or their use within libraries?
- How can knowledge of bibliometric processes contribute to the operations of libraries and other information systems?

The Origins of Bibliometrics

There were a few efforts in the late nineteenth century that might be considered protobibliometric studies, but the first work that can truly be considered bibliometric in nature was that of Cole and Eales in 1917.[2] In this work, the authors provided a detailed analysis of three centuries of publications in comparative anatomy, emphasizing the growth of the literature and the relative contributions of the European countries. A landmark early work was E. Wyndham Hulme's 1923 study of the *International Catalogue of Scientific Literature*; it was in this work that the term "statistical bibliography" was first used.[3] Hulme addressed a number of bibliometric characteristics of the publications listed in the *Catalogue*, including the sizes of the literatures of different sciences, the number of journals in each science, and the number of journals produced by each country represented in the list. Most of the bibliometric studies carried out during the first half of the twentieth century were independent efforts by scholars in diverse fields who were apparently unaware of one another's works. Since the late 1940s, bibliometrics has flourished, becoming a major subdiscipline within information science.

Lotka's Law: Author Productivity

In 1926, Alfred J. Lotka examined author productivity and proposed the mathematical pattern that has since become known as Lotka's Law.[4] The basis for Lotka's Law, like that of most bibliometric principles, is straightforward and intuitive: for any body of literature, there will be a substantial number of authors who have each contributed only one publication, a smaller number of authors who have each contributed a small number of publications, and a very small group of authors who have each contributed a substantial number of publications. Based on a study of the literatures of physics and chemistry, Lotka concluded that the number of authors making a given number of contributions (n contributions) to a specific body of literature is about $1/n^2$ of the number of authors

2. F. J. Cole and Nellie B. Eales, "The History of Comparative Anatomy. Part I. A Statistical Analysis of the Literature," *Science Progress* 11:578–596 (1917).

3. E. Wyndham Hulme, *Statistical Bibliography in Relation to the Growth of Modern Civilization* (London: Grafton and Company, 1923).

4. Alfred J. Lotka, "The Frequency Distribution of Scientific Productivity," *Journal of the Washington Academy of Sciences* 16:323 (1926).

making one contribution. He further suggested that authors making only one contribution will typically account for about 60 percent of the total number of publications. If, for instance, a body of 1,000 publications is considered, about 600 authors will have each contributed one article. The number of authors who have each contributed two publications will be approximately one-fourth of the number contributing one publication each, or about 150. The number of authors who have each made three contributions will be about one-ninth of the number who have contributed one publication each, or about 67. The number of authors who have made ten contributions each will be about one one-hundredth of 600, or 6. In practice, Lotka's Law tends to break down at the upper end of productivity: most bodies of literature include a very small number of authors who are much more productive than the Law would indicate.

Zipf's Law: The Distribution of Words in Text

Another intuitive principle that has become part of the research base of bibliometrics is the tendency for people to use only a small part of their available vocabulary for most communication. The frequency with which words are represented in text was extensively studied by George K. Zipf. Zipf's work was first presented in 1933,[5] and was later expanded upon in his *Human Behavior and the Principle of Least Effort*.[6] Zipf introduced the concept of "word-types" and "word-tokens." Word-types are distinct words used in a body of text; a listing of all word-types in the text would constitute the complete vocabulary of the text. Word-tokens are distinct occurrences of word-types. When word-tokens are counted, it is possible to generate a ranked list of the frequency with which each word-type occurs in the text. According to Zipf's Law, if r is the rank order of the frequency of occurrence of a given word-type, and f is the actual frequency of occurrence, then $r * f = c$, where c is some constant number unique to the body of literature being studied. C has usually been found to be approximately equal to one-tenth of the total number of word-tokens in the body of text being studied. For instance, comparing the three words "library," "circulation," and "acquisitions" in a hypothetical text might produce the following results:

Word	Frequency (f)	Rank (r)	r*f=c
Library	66	1	66
Circulation	22	3	66
Acquisitions	33	2	66

A graph of the relationship between word-type frequency and word-type rank produces the very distinctive curve shown in Figure 2.1. For most bodies of text, the formula provided by Zipf is quite accurate for the middle range of frequencies, but not for very frequent or very infrequent words. At one extreme of the distribution there are typically a few words, such as conjunctions, prepositions, and articles, that occur with a greater frequency than that allowed for by the formula. At the other extreme there are usually a few rare and sometimes esoteric words that occur much less frequently than the formula would suggest. The general pattern, however, has been found to hold for a substantial variety of bodies of text, and its regularity is rather amazing.

5. George K. Zipf, *The Psycho-Biology of Language* (Boston: Houghton Mifflin Company, 1935).

6. George K. Zipf, *Human Behavior and the Principle of Least Effort* (Cambridge, Mass.: Addison-Wesley Press, 1949).

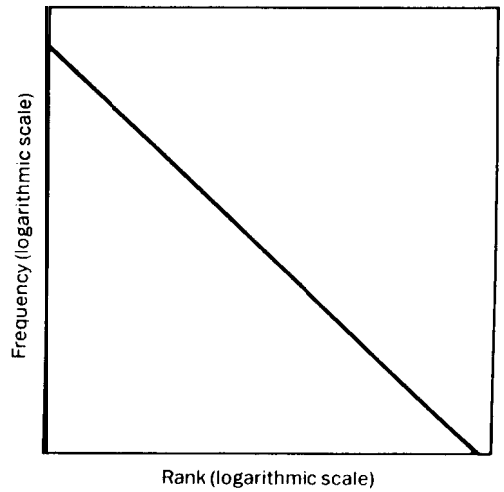

Figure 2.1. The Zipf rank-frequency distribution

Bradford's Law: The Scatter of Literature

The bibliometric principle that has probably received the most attention is scatter. Scatter (also called dispersion or productivity) is based on the frequently observed fact that the use of any collection of items is rarely distributed evenly: some items are heavily used, others receive moderate use, and some are used rarely or not at all. It has been found that the distribution patterns of the use of such items are quite regular and predictable. Measures of scatter include "yearly circulation or other use of a book or journal, number of articles on a given field in the given journal," and "number of references to the given article in subsequent journals."[7]

The most prominent model for scatter is Bradford's Law. In 1934, Samuel C. Bradford examined the literatures of applied geophysics and lubrication and observed a marked regularity in the distribution of articles in relationship to the

7. Philip M. Morse and Ferdinand F. Leimkuhler, "Exact Solution for the Bradford Distribution and Its Use in Modeling Informational Data," *Operations Research* 27:187 (January–February 1979).

journals in which they had been published.[8] Bradford suggested that for each of the two bibliographies studied it was possible to divide the articles into three zones, each of which included an approximately equal number of articles, while the number of journals required to produce those articles increased substantially and regularly from one zone to the next. Bradford formulated a "law of scattering" to characterize this regularity:

> If scientific journals are arranged in order of descending productivity of articles on a given subject, they may be divided into a nucleus of periodicals more particularly devoted to the subject and several groups or zones containing the same number of articles as the nucleus, when the number of periodicals in the nucleus and succeeding zones will be as $1:n:n^2$.[9]

The bibliography of applied geophysics used by Bradford, for instance, conformed to the following pattern:

Zone	Number of Journals	Number of Articles
1	9	429
2	59	499
3	258	404

The starting point for a Bradford analysis is preparation of a ranked list of journals, beginning with the journal that produced the most articles and ending with those journals that each produced one article. This list is then turned into two columns of related figures: the first column provides a cumulative listing of the number of journals, while the second provides a cumulative listing of the number of articles contributed by that number of journals.

When the numbers of journals and of articles are cumulated and plotted on a semilogarithmic graph, the very distinctive curve of Figure 2.2 is produced. The curve consists of three major areas: a long straight section that represents the major portion of the bibliography, a lower curved section that has been called the "Bradford restriction," and an upper curved portion sometimes referred to as the "Groos droop." A number of authors have developed sophisticated formulae for characterizing the shape of the curve and determining the extent to which a particular bibliography is truly Bradfordian, and a large portion of the literature of Bradford's Law has to do with the comparison, refinement, and testing of these formulae.

Bradford's Law has been found to hold for the distribution of articles among journals in most literatures, although the specific shape of the curve varies according to the size and nature of the bibliography. Although most tests of the Bradford distribution have utilized subject bibliographies, Bradford's Law has sometimes been applied to more general collections, including

8. S. C. Bradford, "Sources of Information on Specific Subjects," *Engineering* 137:85–86 (1934).

9. Ibid., 86.

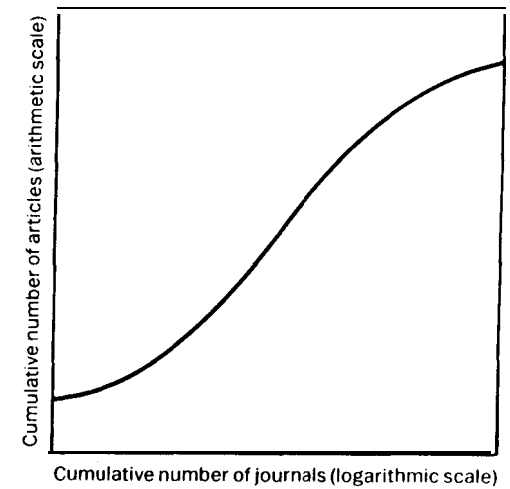

Figure 2.2. The Bradford distribution

those of a special library[10] and a liberal arts college.[11] Bradford's law has also been successfully applied to the study of monograph publishers,[12] library circulation,[13] "the scatter among periodicals of references actually read by a group of scientists,"[14] the distribution of reference questions per requester,[15] the distribution of users of journals in a circulating collection,[16] and the distribution of photocopy requests.[17]

Garfield has suggested that there is actually a Bradford distribution that applies to *all* scientific journals, and has referred to this as a "law of concentration."[18] Holdings of the British Library suggest that the number of

10. E. F. Hockings, "Selection of Scientific Periodicals in an Industrial Research Library," *Journal of the American Society for Information Science* 25:132 (March–April 1974).

11. Martin Gordon, "Periodicals Use at a Small College Library," *Serials Librarian* 6:63–73 (Summer 1982).

12. James C. Baughman, "Toward a Structural Approach to Collection Development," *College and Research Libraries* 38:241–248 (May 1977); Dennis B. Worthen, "The Application of Bradford's Law to Monographs," *Journal of Documentation* 31:19–25 (March 1975).

13. Stephen Bulick, "Book Use as a Bradford-Zipf Phenomenon," *College and Research Libraries* 39:218 (May 1978); Allen Kent and others, *Use of Library Materials: The University of Pittsburgh Study* (New York: M. Dekker, 1977), 38.

14. B. C. Vickery, "Bradford's Law of Scattering," *Journal of Documentation* 4:198 (December 1948).

15. P. F. Cole, "The Analysis of Reference Question Records as a Guide to the Information Requirements of Scientists," *Journal of Documentation* 14:197–207 (December 1958).

16. Paul B. Mayes, "The Use of the Bradford-Zipf Distribution to Estimate Efficiency Values for a Journal Circulation System," *Journal of Documentation* 31:287–289 (December 1975).

17. Donald J. Morton, "Analysis of Interlibrary Requests by Hospital Libraries for Photocopied Journal Articles," *Bulletin of the Medical Library Association* 65:425–432 (October 1977).

18. Eugene Garfield, "Citation Analysis as a Tool in Journal Evaluation," *Science* 178:476 (October 27, 1972).

scientific journals published worldwide is something like 50,000. According to Garfield's Law of Concentration, this number can be divided into three zones as follows:

Zone	Number of Journals
1	1,000
2	6,500
3	42,500

The first zone is presumably made up of multidisciplinary journals and journals that constitute the core of research in specific subject areas. The second zone includes journals that are of a more specialized nature and therefore do not have the broad significance of the first zone journals, and the third zone consists of journals that are of a very highly specialized nature, are of purely local interest, or for some other reason do not contribute as strongly to the overall body of scientific literature.

Obsolescence: The Decline in Use of Materials as They Age

Most experienced librarians are well aware that items in their collection tend to attract less use the older they get. In many libraries this factor is used as a rationale for discarding older items, and the process of determining the point at which an item has become so old that it should be discarded is an important collection management issue. This aging process has also been observed in the context of the references included in scholarly publications: most references tend to be to relatively recent publications, and the likelihood of a publication being cited appears to decline over time. This aging process is generally referred to as "obsolescence," although the aging of information sources and the obsolescence of a technology or methodology clearly are not directly analogous. When a piece of machinery is said to be obsolescent, there is usually the implication that it has been replaced by a better piece of machinery and is therefore no longer of use. Obsolescence in bibliometrics, however, suggests only that older materials are not used, not that they are no longer useful.

Obsolescence has usually been studied in the context of the circulation of items in a library collection, or of the citation of one body of literature by another. The results of obsolescence studies are quite consistent: when items are ranked according to their age at the time they are used (circulated, requested, cited, etc.), recent items account for a very large proportion of the items used, while very old items receive very little use. This distribution tends to conform to the pattern shown in Figure 2.3. Obsolescence is sometimes described in terms of the "half-life" of a literature. The basic definition of the half-life of a literature is "the time during which one-half of all the currently active literature was published."[19] Although most studies of obsolescence have addressed aging

19. R. E. Burton and R. W. Kebler, "The 'Half-Life' of Some Scientific and Technical Literatures," *American Documentation* 11:18–19 (January 1960).

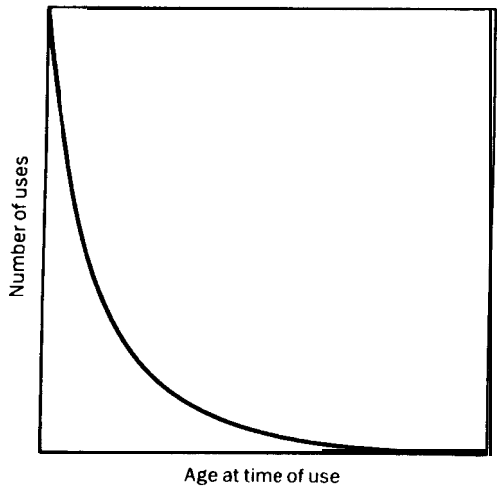

Figure 2.3. The obsolescence curve

qualities of entire literatures, it has been suggested that "what librarians need to know is how long they need to keep individual journals. For this, *item half-life* figures for each of the journals are ideally required."[20]

Obsolescence studies have been conducted of diverse materials in a variety of environments, including periodical use in a small biomedical library,[21] book circulation in a major medical library,[22] the date distribution of periodicals circulated in a large public library,[23] the circulation of books in a university library,[24] the use of physics[25] and biomedical[26] journals in academic libraries, and photocopy requests in a pharmaceutical library.[27] Although the shape of the curve is dependent upon the specific circumstances of the study, the general nature of the obsolescence curve shown in Figure 2.3 appears to be universal.

20. M. B. Line, "The 'Half-Life' of Periodical Literature: Apparent and Real Obsolescence," *Journal of Documentation* 26:47 (March 1970).

21. Judith Wallen Hunt, "Periodicals for the Small Bio-Medical and Clinical Library," *Library Quarterly* 7:121–140 (January 1937).

22. Frederick G. Kilgour, "Recorded Use of Books in the Yale Medical Library," *American Documentation* 12:266–269 (October 1961).

23. Peter Spyers-Duran, "The Use of Periodicals in a Large Public Library," *Wilson Library Bulletin* 36:299–300 (December 1961).

24. Kent and others, 16–18.

25. Ching-Chih Chen, "The Use Patterns of Physics Journals in a Large Academic Research Library," *Journal of the American Society for Information Science* 23:254–265 (July–August 1972).

26. Michael V. Sullivan and others, "Obsolescence in Biomedical Journals: Not an Artifact of Literature Growth," *Library Research* 2:29–45 (Spring 1980–81).

27. Victor A. Basile and Reginald W. Smith, "Evolving the 90% Pharmaceutical Library," *Special Libraries* 61:81–86 (February 1970).

Citation Studies: The Paper Trail in Scholarly Publication

The practice of giving credit to the sources of information used by an author in preparing a publication has a long history and is considered an integral and fundamental part of scholarly activity. Because it is such a basic activity, it has also become the subject of considerable study. The first citation study was probably that of Gross and Gross in 1927, in which the authors explored the use of citation counts as a selection tool for journals in chemistry.[28] Some of the questions addressed by citation studies are:

- What motivates an author to cite a particular work?
- What is the relationship between a citing work and the works cited by it?
- Why are some works cited long after their publication while others are cited only when relatively new?
- Why are some works heavily cited while others are cited infrequently or not at all?
- How do citation practices and patterns differ among disciplines or families of disciplines?
- How can citation practices and patterns be used in the evaluation of information sources?
- How can citation practices and patterns be used to enhance information retrieval systems?

The fundamental assumptions of most citation studies include:

- The citing author has actually used the cited work and has cited all works used.
- Citation of an information source is an indicator of its quality.
- The citing author has provided references to the best possible works.
- Content of the citing work is significantly related to the content of the cited works.
- All citations are of equal value.

Smith has provided a good summary of the rationales for and problems of these assumptions as part of an overview of the nature and use of citation analysis.[29] The overall conclusion that can be drawn is that none of the assumptions is universally true, although each may be true under certain circumstances.

One major problem of citation analysis is that many factors can motivate an author to cite another work, and determining the true relationship between citing and cited publications may require an understanding of the specific motives for a given act of citation. These factors can include a desire to give the appearance of being in touch with the most recent literature, the need to provide support for a methodology or tool, attempts to persuade the reader of the correctness and importance of the ideas being presented, providing appropriate credit for the

28. P. L. K. Gross and E. M. Gross, "College Libraries and Chemical Education," *Science* 66:385–389 (October 28, 1927).

29. Linda C. Smith, "Citation Analysis," *Library Trends* 30:83–106 (Summer 1981).

origin of ideas, alerting the reader to important publications, establishing evidence of a consensus of opinion among researchers, and refutation of the claims of other researchers. Brooks found that the balance among these factors is quite variable and that motivations may represent a complex combination of factors.[30] Other research has suggested that the point within a publication at which a reference is given cannot be successfully used as an indicator of the motivation for citation.[31] Despite these limitations, the notion that citation represents a rather constant indication of the relationship between one information source and another lies at the heart of most citation studies, and plays a key role in the practical application of citation analysis.

Citation analysis has also been extensively explored as a means for evaluating the work of institutions and individuals. The data provided by *Science Citation Index*, *Social Sciences Citation Index*, and *Arts & Humanities Citation Index*, particularly in their online forms, can readily be exploited for

> obtaining lists of publications by a given author, for determining centers of certain types of research, for comparative evaluations of academic departments, and for evaluation of peers in tenure and promotion considerations.[32]

These applications of citation analysis have been the subject of considerable controversy and have been regarded with a mixture of acceptance, trepidation, and scorn. A listing of publications obtained from such a source may not represent all of an author's output, and the assumption that it does may result in a faulty assessment of the author's contribution to his or her field. Similarly, citation counts will represent only citations from the journals covered by the citation index and cannot honestly be assumed to represent all possible citations to an author's work. There is also a substantial potential for error in determining whether citations should really be attributed to a given author, since the citation indexes provide only the author's last name and initials and are subject to virtually no authority control. The uneducated use of citation counts for evaluative purposes of any kind can have disastrous results, and a very real problem of citation analysis is application of results by individuals who are not capable of effectively interpreting them.

The Theoretical Basis of Bibliometrics and Citation Analysis

Bibliometrics and citation analysis both provide ways of examining the patterns of activity related to information sources. A potential limitation of bibliometrics and citation analysis is the lack of a well-developed unified theoretical base to

30. Terrence A. Brooks, "Evidence of Complex Citer Motivations," *Journal of the American Society for Information Science* 37:34–36 (January 1986).

31. Susan Bonzi, "Characteristics of a Literature as Predictors of Relatedness between Cited and Citing Works," *Journal of the American Society for Information Science* 33:208–216 (July 1982).

32. Barbara A. Rice and Tony Stankus, "Publication Quality Indicators for Tenure or Promotion Decisions: What Can the Librarian Ethically Report?" *College and Research Libraries* 44:173 (March 1983).

explain and predict the patterns that have been observed. One proposal that has been presented and explored as an explanation for bibliometric functions is the cumulative advantage theory. This theory, which was developed by Derek de Solla Price, invokes the "Matthew Principle":

> For whosoever has, to him shall more be given, and he shall have an abundance; but whoever does not have, even what he has shall be taken away from him.[33]

In more practical terms, the cumulative advantage theory suggests that all information sources begin with an equal probability of use. Each time an information source is used, however, its likelihood of use increases, while the likelihood of an as-yet-unused information source being used remains constant. By extension, the cumulative advantage theory can be used as a tentative explanation of the "success breeds success" nature of all the bibliometric laws:

> A paper which has been cited many times is more likely to be cited again than one which has been little cited. An author of many papers is more likely to publish again than one who has been less prolific. A journal which has been frequently consulted for some purpose is more likely to be turned to again than one of previously infrequent use. Words become common or remain rare. A millionaire gets extra income faster and easier than a beggar.[34]

The cumulative advantage theory makes it possible to understand that the various bibliometric laws really represent one phenomenon that is closely related to other statistical distributions, including the size distribution for islands described by Mandelbrot and the Pareto Law of Income Distribution. Price's theory has been criticized by others, but the criticisms have generally taken the form of quibbles regarding the exact formulation of equations related to the theory rather than rejections of the theory itself. Although the complete ramifications of the cumulative advantage theory as an explanation of bibliometric phenomena have not yet been revealed, it does serve to provide a preliminary approach to understanding a complex set of related distributions.

Practical Applications of Bibliometrics and Citation Analysis

A common theme of bibliometric studies and citation analyses is that the results of such studies can be of practical assistance in library collection management and the development of new information retrieval systems. The appeal of methods based on bibliometrics lies largely in their emphasis on quantification. Such methods may allow for more scientific approaches to making decisions regarding the selection, retention, and location of bibliographic items in library collections. Similarly, it is possible to apply bibliometric techniques to the process of determining what information sources should be covered by an indexing or abstracting service. Citation analysis has been proposed as a means of identifying high-quality publications and has also been used in the development of alternatives to traditional subject indexing.

33. Matthew 13:12 NASB.

34. Derek de Solla Price, "A General Theory of Bibliometric and Other Cumulative Advantage Processes," *Journal of the American Society for Information Science* 27:292 (September–October 1976).

Some bibliometric principles appear to have no direct applicability to the solution of practical problems. Zipf's Law, for instance, is of interest to linguists and is useful in characterizing the differences among literatures,[35] but has as yet had no direct impact on library practice and has had only a limited effect on the design of information retrieval systems. As fulltext databases become more common, however, Zipf's Law may be applied to the processing of large documents in electronic form, and it has already had some impact on the design of natural language interfaces for information retrieval systems. Similarly, Lotka's Law provides insight into the nature of authorship and can be used in comparing disciplines and their literatures, but it appears to have little potential for explicitly aiding in the design or operation of information systems. The two bibliometric concepts that have been most frequently proposed as potential aids to collection management and information system design are scatter and obsolescence.

Application of the Bradford distribution has frequently been proposed as a means for identifying the journals that are most important to the study of a specific topic, based on the assumption that the most productive journals are also in some way the most valuable. A limitation of this assumption is the lack of sound empirical evidence to support it. Very few studies have been conducted of the relationship between journal productivity and journal quality, and the results of those that have been conducted are mixed. Wallace and Bonzi, in a study of the literature of bibliometrics and citation analysis, found that there was a correlation between journal productivity and the frequency with which journals were cited, but this approach requires acceptance of the unproven assumption that frequency of citation can be used as a measure of quality.[36] Using a different methodology, Lamb found a similar relationship between quality and quantity for the literature of mathematics.[37] Pontigo-Martinez, on the other hand, found no significant relationship between journal productivity and the evaluation of journal articles by a panel of judges,[38] and Boyce and Pollens found no significant correlation between a ranking of mathematics journals by citation and a ranking produced by a Bradford analysis.[39]

Brookes has listed a set of information system design questions that could be answered through proper application of Bradford's Law:

35. Ronald E. Wyllys, "The Measurement of Jargon Standardization in Scientific Writing Using Rank-Frequency ('Zipf') Curves" (Ph.D. dissertation, University of Wisconsin, 1974).

36. Danny P. Wallace and Susan Bonzi, "The Relationship between Journal Productivity and Quality," *Proceedings of the American Society for Information Science* 22:193–196 (1985).

37. Gertrude House Lamb, "The Coincidence of Quality and Quantity in the Literature of Mathematics" (Ph.D. dissertation, Case Western Reserve University, 1971).

38. Jaime Pontigo-Martinez, "Qualitative Attributes and the Bradford Distribution" (Ph.D. dissertation, University of Illinois, 1984).

39. Bert R. Boyce and Janet Sue Pollens, "Citation-Based Impact Measures and the Bradfordian Selection Criteria," *Collection Management* 4:29–36 (Fall 1982).

1. What would be the cost of collecting *all* the journals relevant to a given topic?
2. What fraction of the total coverage would be available at any specified limit of cost?
3. What is the optimum distribution of journal collections between a central reference point and satellite departments or regional collections?
4. How can a given collection best be subdivided into collections of primary, secondary, and tertiary relevance or into stores requiring frequent, occasional, or only rare access?[40]

The model for the use of Bradford's Law that is usually presented involves first conducting a Bradford analysis, then making decisions regarding the level of coverage to be provided as determined by available space, cost of acquiring journals, or some related set of factors. This set of decisions is used to define a cutting point in the distribution to be used in collection management. If, for instance, a fixed amount of money is available for purchasing journals in a specific subject area, it is possible to use a Bradford analysis to identify the most productive journals, and then adopt the strategy of acquiring as many of the most productive journals as the budget will allow. This strategy is particularly appealing in that it will result in the acquisition of those journals that make the greatest number of contributions to the subject literature, and it does not require that explicit decisions be made regarding the relative quality or usefulness of the journals. The strategy is effective even if the subject literature does not conform very well to the Bradford distribution, since

> it is not the zones of productivity, or the exponential nature of the decrease in productivity, but the simple decrease itself that leads to the selection strategy.[41]

The major limiting factor of the strategy is the effort required to gather the data for Bradford analysis. A fairly large body of articles is necessary for an accurate analysis, and it is necessary to conduct the study not just once, but continuously, since journal rankings may change over time. It is conceivable that an automated procedure for evaluating the contents of machine-readable databases could be developed, but no such procedures exist at the present time.

The Bradford distribution also appears to apply to the frequency with which materials are circulated in libraries, and in this area the potential for developing automated procedures is high. As more and more libraries adopt automated circulation systems, it is virtually inevitable that the ability to automatically collect and analyze circulation statistics will become increasingly sophisticated. One augmentation that would require a relatively minimal effort is a procedure for cumulating such statistics in a Bradford manner. Such

40. B. C. Brookes, "The Derivation and Application of the Bradford-Zipf Distribution," *Journal of Documentation* 24:249 (December 1968).

41. Bert R. Boyce and Mark Funk, "Bradford's Law and the Selection of High Quality Papers," *Library Resources & Technical Services* 22:391 (Fall 1978).

an analysis could rather easily be used in making decisions concerning such factors as discarding materials that are infrequently circulated, acquiring additional copies of heavily used materials, and moving materials to remote storage areas.

The principles underlying the study of obsolescence are also of potential use in library collection management:

> If documents are being considered, the interest is probably a practical one in the probability that an item will be required, as a guide to such questions as when to discard older volumes, how long to keep new ones, what sort of retrospective storage and access an information retrieval system should provide, and so on.[42]

The idea that older materials may legitimately be either discarded or relegated to some remote storage area is familiar to all librarians. The problem of employing strategies for discarding or moving to secondary storage lies in determining when an item is old enough to be removed from the primary collection, and decisions are frequently made on the basis of *ad hoc* rules of thumb or vague guesses. Determination of the actual patterns with which use of the collection declines over time can help make it possible to make more informed decisions and reduce the potential for making incorrect decisions. The use of systematic obsolescence studies in collection management, like the use of Bradford analysis studies, is at present hampered by the difficulty of gathering appropriate data.

The simplest application of citation analysis to collection management involves obtaining citation counts for a body of publications and using the counts to rank the sources. If, for instance, the object is to determine the most important journals in a particular subject area, it is possible to (1) identify all the journals that appear to be relevant to the subject; (2) consult the "Journal Citation Reports" section of *Science Citation Index* or *Social Sciences Citation Index* for citation counts; and (3) rank the journals accordingly.

This approach has the appeal of simplicity, but it may produce results that are biased in favor of journals that publish relatively large numbers of articles. This bias can be reduced through the use of "impact factor" rather than gross citation as a measure of importance. An impact factor divides the number of citations to a journal over a fixed period of time by the number of articles available for citation. "Journal Citation Reports" includes impact factors as well as raw citation counts. Similar approaches can be taken to assessing the relative value of monographs or the contributions of different publishers, although gathering data for such studies is at present more difficult than gathering data for ranking journals. Although the relationship between citation and quality remains uncertain, citation is an indicator of use, and a citation-based approach to identifying important information resources does have the potential for producing a listing of the most-used sources.

42. Maurice B. Line and A. Sandison, " 'Obsolescence' and Changes in the Use of Literature with Time," *Journal of Documentation* 30:283 (September 1974).

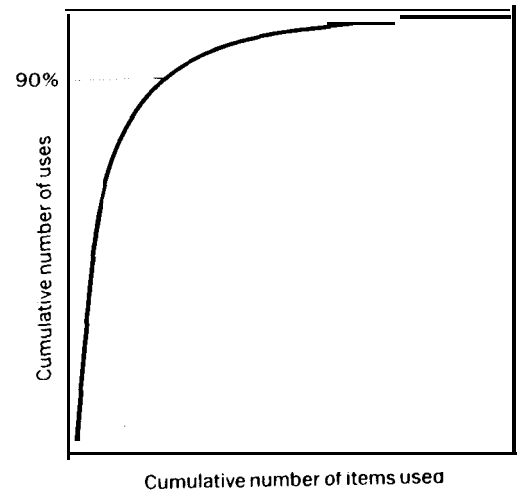

Figure 2.4. The 90 percent library

The application of bibliometric and citation analysis methods to collection management has sometimes been described in terms of the "90 percent library."[43] The principle of the 90 percent library is based upon the intuitively sound notion that no library can expect to meet all the demands of its users. It may, however, be possible to predict the proportion of needs that can be met, and to plan accordingly. In other words, it is usually not practical to provide service at a *maximum* level, but providing service at some carefully defined *optimum* level may be a reasonable goal. Figure 2.4 provides graphic evidence that there is some point of diminishing returns beyond which major additions of resources are required to achieve minor additional value. The 90 percent figure is arbitrary; the desired performance level for a given library should be based on local needs and requirements.

To a certain extent, the concepts of bibliometrics and citation analysis are ways of systematically looking at situations that are obvious to anyone who has been involved in collection management. Information sources are not all of equal value. In terms of the scatter of a periodical literature, different journals contribute different numbers of articles. In terms of library circulation, some items are heavily used while others are not used at all. In terms of obsolescence, newer items receive more use than older items. In terms of citation, some publications are cited frequently and for a long time, others are cited rarely and only while they are relatively new. The model of the 90 percent library adds a systematic basis to these intuitive conclusions. By graphing the distribution of articles over

43. Charles P. Bourne, "Some User Requirements Stated Quantitatively in Terms of the 90 Percent Library," in Allen Kent and Orrin E. Taulbee, eds., *Electronic Information Handling* (Washington: Spartan Books, 1965), 93–110.

the journals that contribute them in a Bradfordian manner, it is possible to identify those periodicals that should be held in order to achieve some selected percentage of coverage of the literature of a particular field. Similarly, by graphing the distribution of items according to their age at the time they are used, it is possible to determine which items should be held in order to achieve a certain percentage of retrospective coverage. The advantage in conducting such systematic analyses is the removal of guesswork from activities that are fundamental to collection management.

The Future of Bibliometrics and Citation Analysis

Bibliometrics and citation analysis have been the subject of study for three-quarters of a century, but do not appear to have been incorporated into the literature or the practice of collection management. Although collection management journals have included some articles on the use of bibliometrics, they do not seem to have treated bibliometrics as a serious alternative to more traditional methods. Textbooks dealing with the selection, acquisition, and management of publications in libraries generally carry at best a passing reference to bibliometrics and citation analysis. The journals in which bibliometric studies and citation analyses most frequently appear are probably not the journals most frequently read by librarians with collection management responsibilities, and this may contribute to the paucity of published reports on the application of bibliometrics and citation analysis. This does not imply that bibliometrics and citation analysis are not useful as collection management tools, but it does seem to be the case that such methods are not frequently applied to "real world" collection management problems.

The absence of bibliometrics and citation analysis from the toolboxes of collection management librarians may in large part reflect satisfaction with the tools already in use. Quantitative methods require an orientation that is quite different from that necessary for the application of more traditional qualitative methods and require the acquisition of new and different skills. It has also been suggested that "the qualitative methods currently used by librarians allow for a degree of personal control and enjoyment that would not be present if purely quantitative methods were employed."[44] The adoption of new methods inherently involves some element of risk; the limited resources of most libraries and the resultant heavy penalty associated with error may make the risk of exploring new collection management methods intimidating. External forces may in the future provide an increased impetus for the employment of quantitative tools as augmentations to, if not as replacements for, qualitative methods. As Warr has pointed out, "the transition from the affluence of the 1960s to the austerity of the 1980s" has produced a need for more creative collection management practices and a concomitant demand for greater accountability in library budgeting

44. Danny P. Wallace, "A Solution in Search of a Problem: Bibliometrics and Libraries," *Library Journal* 112:47 (May 1, 1987).

and spending.[45] Warr has developed a "favorability index" that adds quantification to the traditional practice of using reviews to judge the quality of publications. This index represents a very positive step in the process of incorporating bibliometrics and citation analysis into the collection management process.

It is possible that the major contributions of bibliometrics and citation analysis lie in areas other than providing direct input into collection management decisions. A major goal of information science is expanding understanding of the ways in which information resources are produced and used, and the ways in which production and use differ among different groups of people. Bibliometrics and citation analysis provide a great deal of potential for accomplishing this goal. The priorities and objectives of a particular set of scholars are surely reflected in their publication activities, and it may therefore be possible to use bibliometric techniques to summarize the character of a discipline and compare it to others. In addition to increasing general understanding of the sociology of scholarly production, bibliometrics and citation analysis may provide a useful indirect contribution to library collection management and other information system functions by replacing poorly formed speculation and traditional wisdom concerning the differences among disciplines with objectively verified data.

At any rate, it is likely that bibliometrics and citation analysis will continue to grow in interest and in scope. There are many problems in both areas that have not yet been adequately addressed. An area that is as yet largely unexplored is the interrelationships among different bibliometric phenomena. Bibliometric principles have not yet been extensively applied to the arts and humanities, or to nonscholarly literatures. The development of programs for incorporating bibliometric procedures into automated systems for libraries and other information systems seems likely. As these and other areas are explored, old questions will be answered and new ones will arise. As answers to new questions are provided, the overall impact and implications of bibliometrics and citation analysis will be better understood.

45. Richard Bruce Warr, "Bibliometrics: A Model for Judging Quality," *Collection Building* 5:29 (Summer 1983).

CHAPTER

3

Linguistics and Information Science

Chingkwei Adrienne Lee
and
John N. Olsgaard

The disciplines of linguistics and information science are closely related to each other. Linguistics can be defined as the study of human language as a system for communication, whereas information science is concerned with the communication of information for which language is the primary medium. The study of twentieth-century linguistics has been concerned primarily with oral communication, while information science has focused on written documentation. However, these formalistic delimiters between the two disciplines have increasingly become blurred, and interdisciplinary examinations have recently been started.

The purpose of this chapter is to give a general introduction to linguistics and to examine the manner in which linguistic concepts can be applied to information science.

A Brief Overview of Linguistics

The word "linguistics" was first used in the nineteenth century to emphasize the difference between a newer approach to the study of language that was then developing and the more traditional approach of "philology." The differences were and are largely matters of attitude, emphasis, and purpose. The philologist is concerned primarily with the historical development of languages as it is manifested in written texts and in the context of the associated literature and culture. The linguist tends to give priority to spoken languages and to their structure as they operate at a given point in time, without reference to their history. The linguist, in principle, is interested in all languages and not merely in the great literary languages of the world.

The field of linguistics may be viewed from two different perspectives; the first is in terms of dichotomies, the second is in terms of branches or subdisciplines.

Linguistic Dichotomies

The field of linguistics can be divided into three dichotomies: synchronic versus diachronic, theoretical versus applied, and microlinguistics versus macrolinguis-