Jointly published by Elsevier Science Ltd, Oxford and Akadémiai Kiadó, Budapest

Scientometrics. Vol. 43, No. I (1998) 69-76

RANDOM THOUGHTS ON CITATIONOLOGY. ITS THEORY AND PRACTICE*

E. GARFIELD

Institute for Scientific Information, 3.501 Market Street, Philadelphia, PA 19104 (USA)
and
The Scientist, 3600 Market Street, Philadelphia, PA 19104 (USA)
e-mail.. garfieId@codex.cis.upenn.edu

(Received April 9, 1998)

Theories of citation are as elusive as theories of information science, which have been debated for decades. But as a basis for discussion I offer the term citationology as the theory and practice of citation, including its derivative disciplines citation analysis and bibliometrics. Several maxims, commandments if you will, have been enunciated. References are the result of a specialized symbolic language with a citation syntax and grammar. References, like words, have multiple meanings which are related to the aposteriori quality of citation indexes. Therefore, citation relevance cannot be predicted. Mathematical microtheories in bibliometrics abound, including the apposite laws of scattering and concentration. Citation behavior is a vast sub-set of citation theory, which like citation typology, can never be complete. Deviant citation behavior preoccupies certain authors but it is rarely significant in well-designed citation analyses, where proper cohorts are defined. Myths about uncitedness and the determinants of impact are discussed, as well as journal impact factors as surrogates and observations on scientists of Nobel Class.

After two years at Johns Hopkins investigating "machine documentation," and another year as a student of library science, I became, fortuitously, a documentation consultant By 1954, I called myself an infomation engineer, which was an apt description of my professional consulting activities However, Pennsylvania licensing law requires that engineers be graduates of engineering schools So I became an information scientist! I've never thought of myself as an information theoretician and have been skeptical about a need for a theory of information science I've practiced information science and engineering without explicit theoretical support. But undoubtedly there are underlying principles which can guide information scientists who, like myself. could be called "citationists" or "citationologists." If there is a theory and practice of citation, it should probably be called citationology.

* Comments on Theories of Citation? L. LEYDESDORFF *Scientometrics,* 43 (1998) No. 1

0138-9130/97/US *$15 00*

I myself did not begin as a bibliometrician, scientometrician, or citation analyst. There was no guiding "theory" which led to my interest in citation indexing. My basic dissatisfaction with traditional methods of indexing, cataloging and classification was a major stimulus. Then my serendipitous encounter with the American system of legal documentation led to citation indexing. Citators, as they were called, provided the framework for my earlier parallel interest in natural language and "simple" systems for dissemination. The latter were manifested in the creation of Current Contents – which eventually led to title-word, and permuterm indexing.[1] Nevertheless, a formalized description of citation indexes might be regarded as essential to citation theory building. In their printed form, citation indexes are two-dimensional displays of the linkages between document addresses. These addresses are sometimes called references, but also citations, depending upon the direction of reference-to-reference links. Papers and books cite references, that is, earlier papers and books and other documents (patents, letters, etc.) but occasionally cite into the future, for exaniple, papers and books in press. Citations to earlier work provide backward links while citation indexes provide forward links.

From the outset, we recognized that references or citations reflect a natural international language of science and scholarship. The "grammar" of bibliographic citations was described by compiling a dictionary of many dozens of citation formats.[2] Their symbolic role was formally described by *Small* in his now seminal paper.[3] It is relevant to mention my earliest lecture tours. To illustrate the symbolic role of citations, I often quoted Lewis Carroll's "Humpty Dumpty" to express the ambiguous nature of words and citations. "When I use a term it means just what I want it to mean – nothing more or less." And so when you use a cited reference (citation), it also means what you want it to mean. A citation is generally more precise than words, but its meaning is ambiguous nevertheless. We all use citations with slightly different intentions and meanings. In some contexts citations can be extremely precise symbols. If you see a reference to Oliver Lowry's 1951 classic on protein determination,[4] you can be 99.9% certain that his method. is being used, but you cannot predict whether or when a slight modification is reported by the 250,001st paper that has cited it. From the beginning, citation indexes were characterized as *aposteriori* indexes, in contrast to *apriori* traditional indexes. *Wouters* has recently reminded us of this quality of citation indexing.[5]

I would like to enunciate some commandments or maxims of a theory of citation. A first commandment might be – there is no way to predict whether a particular citation (use of a reference by a new author) will be "relevant." *Cleveron's* great contribution to information retrieval theory[6] was built around the notions of precision, recall, and

relevance. But none of his or related studies accounts for or measures relevance in retrieval by citation. As in so many things, relevance is in the eye of the beholder. The *main theme* of most papers that cite the Lowry method have little to do with the method itself. Relevance studies concern themselves primarily with main theme indexing.

In this respect, related records, a variant on bibliographic coupling,[7] produces a type of relevance ranking which might be compared with ranking by permuterm indexing, that is co-word occurrence.

Another maxim – if author *X* cites the work of author *Y,* regardless of the reason, then this fact alone makes the citing paper relevant to author *Y* and, furthermore, author *X* may be interested in other papers that cite *Y.* Undoubtedly, Professor Lowry long ago gave up trying to assess the thousands of papers that cited his work each year, but each occurrence was highly "relevant" to others interested in protein determination methods and reagents.

I have experienced the greatest excitement in finding references to my own work in the least likely of places, as e.g., Kevin Kelly's discussion of citation analysis[8] and, more recently, in Candace Pert's reflection on – *Current Contents* – in The Molecules of Emotion.[9] Conventional subject headings or title word indexing would not call out these connections.

Full-text searching will make this possible. While the use of bibliographic coupling (related records) can give some degree of relevance ranking to the papers that cite a particular paper I have written, it is often the "onesies" that are the most interesting. In other words, the degree of linkage between my work and the newly discovered author is simply that we both have cited one or two other authors in common.

Putting aside the anomaly of Oliver Lowry and other super citation stars, one can say, for the sake of theory building, that whenever author *X* is cited, he will regard the citing work as initially relevant, even though on closer inspection it may not prove interesting. However, in laboratory-based scientific research, it is highly likely that high degrees of bibliographic coupling (related records) will produce high degrees of relevance.

Is there a body of laws which govern the citation world? The literature of bibliometrics provides all sorts of mathematical descriptions of citation distributions. These go back quite far, and I will not attempt to recapitulate these microtheories. We all know or have heard about Lotka's Law, Pareto's Law, Zipf's Law, or whatever, but also especially Bradford's Law of Scattering. The latter has been discussed in countless papers by Brookes, Bookstein, Price, Leimkuhler, Rousseau and others too numerous to mention. Bibliometricians are fascinated by these mathematical exercises which permit them to display their admirable mathematical and probabilistic insights.

My limited contribution to this mathematical microtheory world includes my counterpoint to Bradford's Law of Scattering,[10] namely Garfield's Law of Concentration.[11] It seems remarkable to me after so many decades that so few people really appreciate the economic consequences of these phenomena. The most productive journals have remarkable stability and impact. This has been demonstrated periodically every decade since we began our first experiments with the Genetics Citation Index. When we created, restrospectively, the 1945-54 *Science Citation Index,* – we selected the core group of 600 journals by a purely algorithmic procedure – that is citation frequency. The effectiveness of this list is demonstrated daily. This group of journals identifies a high percentage of the post-war literature which was then and still is cited regularly.

Another modest contribution I made to the microtheory of citation is Garfield's constant.[12] Actually, we know that this "constant" is really a ratio. That ratio is remarkably "stable" considering how much the literature has grown. Due to continuous growth of source journal coverage and increasing references cited per paper, the ratio of citations to published papers increased about 75% from 1945 to 1995 – from 1.33 to 2.25 over the past 50 years. It is the inflation of the literature which increases the ratio each year.

Citation behavior

If description is part of citation theory building, the characterization of *citation behavior* must be part of it. A considerable literature discusses citation "behaviors." Such behavior was discussed in the earliest days of experimenting with citation index building which I recently reviewed in "When to Cite"[13] There probably never will be a complete typology of citation behavior. There always will be new reasons why people cite. My 1979 text on *Citation Indexing,*[14] contained a somewhat limited typology which is sufficient for most purposes. Within each branch of science, there will be more specific types of citation.

For example, in chemistry, an author may cite a paper simply because it reports the melting point of a solid chemical. One can never know apriori whether or how that citing paper will interest another chemist. The latter might cite the same paper because it mentions the failure to make a similar compound. That is why citation indexes can be useful to the compilers of specialized chemical databases or handbooks which provide precise information of this kind in a more condensed form. It also highlights the dilemma faced by citation index compilers who regularly encounter pageless documentation.[15] The latter simply means that a book or paper has been cited but the

specific page or chapter has been omitted from the reference or is buried in the text of the citing paper. The dilemma lies in whether or not these pageless citations should be unified with page specific citations so that all citations to the main work are combined into a single entry. Otherwise, scientometricians obtain distorted citation counts especially when they rely on electronic searches that do not unify the variations.

Deviant citation behavior preoccupies critics of citation analysis. It is fashionable for them to mention that authors are lax in their scholarship even with good refereeing. Anecdotal accounts would have us believe that most documentation is fraudulent. But I have never seen a documented case of a citation analysis gone astray due to accumulated deviant citation behavior. A common flaw in appraising citation analysis, is that author citation frequency is used out of historical context. Using the current year's literature to ascertain the most influential authors of the past is naive indeed.

Such an exercise ignores another maxim of citation analysis. Always measure citation links in the appropriate period of literature. Cawkell and I addressed these issues in discussing the long-term influence of Albert Einstein and other authors. [16,17] It is absurd to use only recent literature for such purposes. Longitudinal data is essential to evaluate the historical influence of a particular paper, author or book. It is often forgotten how rapidly the scientific literature changes and how new discoveries are superseded. In short, the modern scientific article makes no pretense at being historically comprehensive and stresses the most recent literature. Modem bibliographies do not recapitulate the entire literature of a topic unless that is the stated purpose.

A theory of citation might include a set of commandments of citation analysis. Another commandment that pertains to the evaluation of people, journals, and institutions – always compare or judge equivalent or truly comparable cohorts. Naive administrators, uninformed in citation analysis, will make the mistake of using citation data without regard to the discipline or invisible college involved. Cross-disciplinary comparisons are usually inappropriate. Even in large disciplines, it can be difficult to establish perfect cohort groups of authors or journals. To approach ideal cohorts, co-word and/or co-citation clusters can be used. Thus, when we wanted to identify the pioneers of apoptosis we created a database of papers whose titles contained the word apoptosis or its equivalent, programmed cell death. [18] And even then the omission of a paper using the hyphen in cell-death adversely affected the original choice of core papers. [19]

A theory of citation may also need some maxims concerning the many myths about citation studies. For example, it is repeatedly asserted that the size of a discipline is the primary determinant of the impact factor for people, papers, or journals. But in

discussing Garfield's constant I demonstrated[12] that it is the average number of references cited per paper (R/S) citation half life, and utilization factors that determine impact – and not the size of the literature.

The size of the literature will determine the number of papers that can exceed a particular citation threshold. Thus, biochemistry will produce a large number of papers whose citation frequency exceeds 400 or 500 citations, what I've called citation classics.[20] But the higher impact of average biochemistry papers is due to the field's high R/S per paper rather than the number of papers in the field.

The considerable variation in citation frequency of articles in large fields is overlooked simply because we are conscious of the many highly-cited classics. We are unable to be fully conscious of the thousands of papers that are less frequently cited. The large number of less frequently cited papers is difficult to visualize, unless one examines an article-by-article ranked listing for a specific journal.

But even a paper or book in a small field that attracts cross-disciplinary citation may break out of the expected impact of that field. Kuhn's *Structure of Scientific Revolution* is a superclassic even though it represents a relatively small field of scholarship.

There is also the uncitedness myth. *Hamilton*[21] helped to perpetuate this myth by misinterpretation of citation data which was undifferentiated with respect to types of editorial material. *Pendlebury* effectively rebutted his data[22] but Hamilton's claims persist due to the high visibility of the journal in which it was published and because the theme has popular appeal. *Wade* repeated the myth in a recent *New York Times* story.[23] While I myself have often reported that a large number of papers do indeed remain uncited, they are primarily published in low impact journals. Thus, another commandment in citation studies is: Thou shalt compare items in equivalent editorial categories such as original research papers, reviews, letters, etc.

Another commandment, concerns the journal impact factor. This ratio, which was created to facilitate the comparison of journals regardless of size. has lately been used as a surrogate for actual citation data on individuals. There has been a spate of articles deploring this use but many of the authors have used this particular malpractice as an excuse to malign all quantitative studies. Seglen has justifiably criticized evaluation exercises which blindly use journal impact factors as surrogates. It is well known that there is considerable variation in citation frequency of articles within the same journal.[24,25]

The current impact factors reported in ISI's *Journal Citation Reports* – are useful in their appropriate place. Long-term cumulative impact data can also be used. Various techniques for combining current impact with half-life to produce estimates of long-term impact have been made. The recent availability of ISI's *Journal Performance*

Indicator s makes estimation unnecessary in most cases. The calculation of long-term impacts will, however, not produce exceptional results unless one is making cross-disciplinary comparisons.[26] The rankings of journals within their appropriate category does not change all that much but generalizations are to be avoided.

As a concluding note in these miscellaneous comments, let me refer to some data on Nobel Class scientists which was reviewed at the AAAS meeting in Philadelphia.[27] As *Sher* and I reported in 1965, the average Nobel Prize winner published five times the average author and were cited 30 times the average.[28] Since that long-ago study, we also found that over 95% of Nobel Prize winners are authors of one or more Citation Classics. By extension, over 50% of the 1,000 most-cited scientists are members of the U.S. National Academy of Sciences. Further, these scientists "of Nobel Class" are among the most-cited decile for their discipline. The general figure of 50% fits well with the anecdotal account of a former Academy president who told me "for every scientist elected to the Academy, there is another equally qualified who is not elected." As *Zuckerma n* expressed it,[29] in characterizing the 40 member limit of the French Academy, "Who shall occupy the 41st chair?"

Conclusion: a citation-ranked list of scientists will identify 50% or more present and future members of the Academy. Therefore, while citation frequency alone does not warrant election, the nominating group should at least consider as candidates all those who achieve a given threshold of citation frequency.

References

1. GARFIELD, E., *The Permu ter mSubje ct Ind ex:* An Autobiographical Review, Journal of the American *Society of Information Science* 27(5-6):288-91 (1976). Reprmted in *Essa ys of an Information Sc ien ti st*, Volume 7. Philadelphia: ISI Press, p. 546-550, 1985.
2. GARFIELD, E., SHER, I., Progress Report of Citation Index Project, October 9, 1962 (unpublished).
3. SMALL, H. G. Cited Documents as Concept Symbols, *Social Studies of Science* 8(3):327-340 (1978).
4. LOWRY, 0. H., ROSEBROUGH, N. J., FARR, A. L., RANDALL, R. J., Protein Measurement with the Folin Phenol Reagent, *Journal of Biological Chemistry 193:265-27 5* (195 1).
5. WOUTERS, P., The Signs of Science, *Scientometrics, 41(l-2):225-241* (Jan-Feb 1998).
6. CLEVERDON, C. W., Cranfield Tests on Index Language Devices, *ASLIB Proceedings,* 19: 173 (1967).
7. KESSLER, M. M., Bibliographic Coupling Between Scientific Papers, *American Documentation,* 14(1):10-25 (1963).
8. KELLY ,K., Out of Control: The New Biology of Machines, *Social Systems and the Economic World,* Addison-Wesley, 520 p., 1995.
9. PERT, C., *Molec ul e of Emotio n Why you Feel the Wa you Feel* ,Scribner, 304 p.. 1997.
10 a. BRADFORD, S. C., Sources of Information on Specific Subjects, *Engineering 137:85-86* , 1934.
   b. BRADFORD, S. C., Documentation, Washington, DC: Public Affairs Press. 1950, 156 p.
11. GARFIELD. E., The mystery of the transposed journal lists – wherein Bradford's law of scattering is generalized according to Garfield's law of concentration, *Current Con te nt No. 7:5* (August 4 I971) Reprinted in *Ess ays of an In forma ti o Scienti t* Volume 1 Philadelphia: ISI Press, p. 222-223, 1977

Indicator s makes estimation unnecessary in most cases. The calculation of long-term impacts will, however, not produce exceptional results unless one is making cross-disciplinary comparisons.[26] The rankings of journals within their appropriate category does not change all that much but generalizations are to be avoided.

As a concluding note in these miscellaneous comments, let me refer to some data on Nobel Class scientists which was reviewed at the AAAS meeting in Philadelphia.[27] As *Sher* and I reported in 1965, the average Nobel Prize winner published five times the average author and were cited 30 times the average.[28] Since that long-ago study, we also found that over 95% of Nobel Prize winners are authors of one or more Citation Classics. By extension, over 50% of the 1,000 most-cited scientists are members of the U.S. National Academy of Sciences. Further, these scientists "of Nobel Class" are among the most-cited decile for their discipline. The general figure of 50% fits well with the anecdotal account of a former Academy president who told me "for every scientist elected to the Academy, there is another equally qualified who is not elected." As *Zuckerma n* expressed it,[29] in characterizing the 40 member limit of the French Academy, "Who shall occupy the 41st chair?"

Conclusion: a citation-ranked list of scientists will identify 50% or more present and future members of the Academy. Therefore, while citation frequency alone does not warrant election, the nominating group should at least consider as candidates all those who achieve a given threshold of citation frequency.

References

1. GARFIELD, E., *The Permu ter mSubje ct Ind ex:* An Autobiographical Review, Journal of the American *Society of Information Science* 27(5-6):288-91 (1976). Reprmted in *Essa ys of an Information Sc ien ti st*, Volume 7. Philadelphia: ISI Press, p. 546-550, 1985.
2. GARFIELD, E., SHER, I., Progress Report of Citation Index Project, October 9, 1962 (unpublished).
3. SMALL, H. G. Cited Documents as Concept Symbols, *Social Studies of Science* 8(3):327-340 (1978).
4. LOWRY, 0. H., ROSEBROUGH, N. J., FARR, A. L., RANDALL, R. J., Protein Measurement with the Folin Phenol Reagent, *Journal of Biological Chemistry 193:265-27 5* (195 1).
5. WOUTERS, P., The Signs of Science, *Scientometrics, 41(l-2):225-241* (Jan-Feb 1998).
6. CLEVERDON, C. W., Cranfield Tests on Index Language Devices, *ASLIB Proceedings,* 19: 173 (1967).
7. KESSLER, M. M., Bibliographic Coupling Between Scientific Papers, *American Documentation,* 14(1):10-25 (1963).
8. KELLY ,K., Out of Control: The New Biology of Machines, *Social Systems and the Economic World,* Addison-Wesley, 520 p., 1995.
9. PERT, C., *Molec ul e of Emotio n Why you Feel the Wa you Feel* ,Scribner, 304 p.. 1997.
10 a. BRADFORD, S. C., Sources of Information on Specific Subjects, *Engineering 137:85-86* , 1934.
   b. BRADFORD, S. C., Documentation, Washington, DC: Public Affairs Press. 1950, 156 p.
11. GARFIELD. E., The mystery of the transposed journal lists – wherein Bradford's law of scattering is generalized according to Garfield's law of concentration, *Current Con te nt No. 7:5* (August 4 I971) Reprinted in *Ess ays of an In forma ti o Scienti t* Volume 1 Philadelphia: ISI Press, p. 222-223, 1977

12. GARFIELD, E., Is the ratio between number of citations and publications cited a true constant? *Current* Contents. No. 8: 5-7 (February 9, 1976). Reprinted in *Essays of an Information Scientist,* Volume 2. Philadelphia: ISI Press, p. 4 19-42 1, 1977.
13. GARFIELD, E., When to Cite, *Library, Quarterly,* 66(4): 499-558 (October, 1996).
14 GARFIELD, E., *Citation Indexing – Its Theory and Application in Science, Technology and Humanities,* Philadelphia: ISI Press, 1979. 274 p.
15 GARFIELD, E., Pageless documentation; or what a difference a page makes, Current *Contents, No.* 17:3-6 (April 29. 1985). Reprinted in *Essays of an Information Scientist,* Volume 8, Philadelphia: IS1 Press, p. 160-163. 1986.
16 GARFIELD, , E., The Einstein Centennial and Citation Analysis, *Current Contents, No. 17: 5-9* (April 27, 198 I). Reprinted in *Essays of an Information Scientist,* Volume 5. Philadelphia: IS1 Press, p. 91-95, 1983.
17. CAWKELL A. E., GARFIELD, E., Assessing Einstein's impact on today's science by citation analysis, (GOLDSMITH M, MACKAY A., WOUDHUYSEN J., (Eds), Einstein: the First Hundred Years, Oxford: Pergamon Press, Pg. 3 I-40, 1980. Reprinted in *Essays of an Information Scientist,* Volume 5. Philadelphia: IS1 Press, p. 9 l-95, 1983.
18 GARFIELD, E., MELINO G., The growth of the cell death field: An analysis from the IS1 Science Citation Index, Cell *Death and Differentiation, 4(5):352-361* (July, 1997).
19 G5.74, E.. MELINO G., The growth of the cell death field: An analysis from the IS1 Science Citation Index-erratum, Cell *Death und Differentiation 5(* 1): 127 (January, 1998).
20 GARFIELD, E., Introducing Citation Classics: the Human Side of Scientific Reports, *Current Contents, No. 1.* p. *5-7* (January 3, 1977). Reprinted in *Essays of an Information Scientist,* Volume 3. Philadelphia, ISI Press, p. 1-2, 1980
21 HAMILTON, D. P., Publishing by – and for? – the numbers, Science, 250(4986): 133 l-1332 (December 7, 1990).
22 PENDLEBURY, D. A., "Uncrtedness". Letter to the Editor, *Science, 25 l(5000): 14* 10- 14 I! (March 22, 1991).
23 WADE, N., No Nobel Prize This Year? Try Footnote Counting, NW York *Times,* p. F4 (October 7, 1997).
24 SEGLEN, P. O., Why the Impact factor of journals should not be used for evaluating research, *British Medical Journal 314(7079).498-502* (February 15, 1997).
25 SEGLEN, P. O., Evaluation of Scientists by Journal Impact, Representations of Science and Technology; *Proceedings of the International Conference on Science and Technology Indicators,* Bielefeld, Bielefeld, 10-I 2 June, 1990. P. WEINGART. R SEHRINGER, M. WINTERHAGER, (Eds), Leiden, DSWO Press, p. 240-252, 1992.
26. GARFIELD, E., Long-Term Vs Short-Term Journal Impact: Does It Matter? *The Scientist, 12(3).* 1 I-13 (February 2, 1998).
27 GARFIELD, E., Mapping the World of Sctence. Paper presented at the 150th anniversary meeting of the AAAS, Philadelphia, PA, February 14, 1998 (see http:l/www.the-
28 SHER, 1. H., GARFIELD, E.. New tools for Improving and Evaluating the Effectiveness of Research. *Research Program Effectiveness,* M C. YOVITS, D. M. GILFORD, R. H. W5.74 E. STAVELEY. H. D. LEMER (Eds), New York: Gordon and Breach, p. 135-146. 1966.
29 ZUCKERMAN. H., *The Scientific Elite. Nobel Laureates in the United States,* New Brunswick and London: Transaction Press, 1996