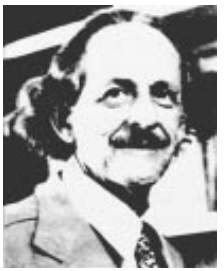
Essays of an Information Scientist: Creativity, Delayed Recognition, and other Essays, Vol:12, p.123, 1989 Current Contents, #18, p.3-8, May 1, 1989



Current Comments®

EUGENE GARFIELD
INSTITUTE FOR SCIENTIFIC INFORMATION®
3501 MARKET ST., PHILADELPHIA, PA 19104

Citation Behavior—An Aid or a Hindrance to Information Retrieval?

Number 18 May 1, 1989

Citation behavior has many meanings. How well we acknowledge our intellectual debts is determined by a variety of motivations—professional, connectional, documental, applicational, and confirmational, among others. Tacit citation norms and the roles of referees are important, too. Average citation behavior produces enough redundancy to ensure retrieval of relevant information through citation indexes or by bibliographic coupling.

The Hungarian Academy of Sciences, Budapest, has become one of the worldwide centers of scientometric research.[1] Tibor Braun, as both editor of the journal *Scientometrics* and an active citation analyst, has been quite involved in promoting the use of the *Science Citation Index*® (*SCI*®) as a source for various studies.[2,3] One of his colleagues at the Central Research Institute for Chemistry (CRIC) of the Hungarian Academy of Sciences, P. Vinkler, published an interesting paper[4] that strikes a responsive chord in me. However, its title, "A quasi-quantitative citation model," does not tell the average reader that it deals with citation behavior—a vast topic about which there is relatively little systematic knowledge. Terrence A. Brooks, School of Library and Information Science, University of Iowa, Iowa City, also recognizes the need for more systematic study of this problem. The following, from one of his papers on citation analysis, is worth quoting: "Evaluative citation analysis has been employed without a clear understanding of why authors give references and in the absence of any empirical work investigating citer motivations."[5]

Citation behavior has many meanings. As the late Manfred Kochen, University of Michigan, Ann Arbor, indicated, it can mean how well we acknowledge our intellectual debts.[6] One of the earliest and continuing criticisms of the *SCI* as a tool for information retrieval and for measuring the "real" impact or influence of authors or papers is that citation behavior is uneven, unpredictable, and biased. Subjective evidence is abundant. Most authors have at least one anecdotal example of a paper that should have cited their work but did not. Over 30 years ago, I pointed out that it was the job of patent examiners to refresh the memories of inventors.[7] I can't recall how often I've said the referee's job is similar—to remind authors when they overlook or perhaps deliberately omit relevant references.

Vinkler's Research

Vinkler provides a commendable analysis of the motivations for citing a paper or book and proposes a model for citation behavior. He also provides some interesting data on a small group of his colleagues (20 authors at CRIC). I summarize some of his results in what follows. Vinkler analyzes 484 references in 20 chemistry articles written by 20 selected authors at CRIC. His goal was to assess the authors' motivations for citing and also to find out why some papers are

cited while others are not. Vinkler also characterizes (quantitatively) the strength of cognitive pressure to cite a paper. He uses the term *citation threshold* to define the lowest value of cognitive pressure.

As Vinkler sees it, authors' motivations fall into two groups. The first group, "professional motivations," refers to the relationship of the citing author to his or her own research (that is, it refers to the question of knowledge or authority). These motivations are linked to the theoretical and practical aspects of research. The second group, "connectional motivations," stems from the relationship of the citing author to the cited authors. Personal, social, or external factors all play a role in connectional motivations.

Of the 484 references analyzed by Vinkler, 81 percent are used for "exclusively professional reasons." Seventeen percent result from a combination of professional and connectional motivations, and only 2 percent are made for "exclusively connectional reasons."

Of the professional motivations, citing a paper for "documentary" reasons (that is, the citing article provides a review of the relevant literature, and the cited article is part of that review) is the most frequent. "Applicational" motivations, where citing authors use part or all of the cited article in their paper, is the second most common professional motivation. Vinkler also finds that "confirmative" motivations play an important role in the citation strategies of authors. Here, authors cite a paper because the results confirm their own.

Vinkler next reports that connectional motivations play a much smaller role than professional motivations in determining citation behavior. Forty percent of the authors that Vinkler studied indicated that they have or will have a professional relationship with the cited author. This, however, does not appear to represent a primary motivation. Michael J. Moravcsik, Institute of Theoretical Science, University of Oregon, Eugene, suggests that Vinkler's number of situations that combine professional and connectional motivations is too low, "probably because what a 'combination' is remains vague." Like Vinkler, he points out the higher probability of citing an author one has met face-to-face (in conferences, etc.), contending that this is among the factors working against scientists in developing countries.[8]

As for reasons for omitting citations, it appears that professional reasons (such as the work was not relevant enough to be cited) exert the most influence. Other professional reasons for omitting a citation include (1) authors taking over commonly known information as their own, (2) authors using one large review as a reference instead of the original papers discussed in the review, and (3) the artificial restriction of the number of references in an article. (Related to the third of these is the inverse situation, when an author may add "perfunctory" references to pacify an editor or referee—a situation that Moravcsik again sees as illustrating the ambiguities involved in the professional-connectional dichotomy.[8])

Vinkler's data also illustrate that the "*citation threshold depends primarily on the professional relevance of the work* potentially citable in the given paper."[4] On the other hand, when connectional motivations play a role, the citation threshold (strength of the motivation to cite) is less related to the relevance of the research.

Human behavior being as complex and varied as it is, I doubt that Vinkler or anyone else has adequately cataloged the full range of citation behaviors. But Vinkler has made significant progress. Additional surveys will be needed in fields other than chemistry to determine whether his findings are representative. I suspect that the analysis would be far more difficult in the social sciences and easier, for example, in mathematics.

Brooks, in the paper mentioned earlier, comes up with a somewhat different, though corroborative, classification of motivators. His list of motivators, from the most to the least prevalent, is as follows: persuasiveness, positive credit, currency, reader alert,

operational information, social consensus, and negative credit. Interestingly, when he further divides his sample into science and humanities subsets, the ranking changes significantly—with currency and social consensus, for example, ranking several places higher in the science subset.[5]

The motivations Vinkler describes all imply, in one way or another, professionally responsible citation behavior. A recent paper by Geoffrey K. Pullum, Cowell College, University of California, Santa Cruz, comes to a very different conclusion. He discusses evidence of a new citation-behavior pattern that is replacing the traditional courtesies of scholarship: what he terms a "me-first," exclusionary behavior norm.[9] Another point communicated by Moravcsik is relevant here. He suggests that what referencing system we consider "just" or "satisfactory" depends on whether we see it from the point of view of science or scientists. In other words, "systems which may allow personal injustices...may nevertheless be quite adequate for the progress of...science." He believes this distinction should be made clear in order to assess behaviors and systems in terms of whether "they are nice to scientists or good for science."[8]

Citation Behavior and Human Judgment

What we "ought" to cite is still a largely subjective matter. Vinkler may not have realized that I had once reported on a rather extensive "experiment" in citation behavior or citation expectancies.[10] By accident, one of my papers, published in the inaugural issue of the *Journal of Chemical Documentation* (now the *Journal of Chemical Information and Computer Sciences*), appeared *without* its list of 41 cited references.[11] The enthusiastic editor, Herman Skolnik, had asked for a copy of my speech at an American Documentation Institute meeting immediately after I had delivered it. He did not realize the version for oral presentation omitted the bibliography contained in the full manuscript. Speakers ordinarily do not read their supporting references to their audiences even though, in those days, we still read entire papers.

Since I was teaching a graduate course in information retrieval at the University of Pennsylvania in Philadelphia, I used this fortuitous event to conduct an exercise in citation behavior. Students were asked to read the paper and indicate any place in the text where they thought a reference "ought" to be provided. The experiment was continued for several years. (I suggest that you try a similar experiment with any group of readers for one of your own papers.)

The experiment proved to be an interesting exercise in refereeing. As it turned out, the number of "expected" references ranged from 15 to 75. The average proved to be almost exactly the 41 I eventually published in a "correction" note.[10]

In reviewing the present essay, Blaise Cronin, Department of Information Science, University of Strathclyde, Glasgow, UK, sent me the following comments, which seem to corroborate the results of my "experiment," described above. He writes:

> I carried out a similar exercise in my doctoral research, when I obtained four prepublication articles from editors of major psychology journals from which I removed all the authors' citations before asking groups of US and UK psychologists (academic and professional) to indicate where they felt citations were needed. I then matched their suggestions with the references provided by the authors and looked for evidence of a consensus among my panelists. I concluded that there was broad agreement among various groups, though, as one would expect, wide variation at the individual level. I never got around to publishing these results, though they are alluded to in my book, *The Citation Process*.[12] My general feeling was that experts in a given field have a tacit understanding as to what constitutes acceptable/required citation behaviour in that field (the details can be found in my dissertation which is cited in the book).[13]

The subjective element in both citing behavior and citing expectation cannot be ignored. However, the general agreement on the average number of citations in this reported "experiment" suggests that there may be a "norm" of citation behavior.[14] The rest of this essay discusses the concepts of citation cycling and bibliographic coupling as further "controls." These, together with norms of citation behavior, ensure a relatively reliable system of information retrieval.

The Significance of Citation Behavior for Information Retrieval

Vinkler's study did not address the crucial question of how citation behaviors affect either information retrieval or impact. And rightly so, it can be argued: a typographical error may cause a relevant citation to be "lost" just as readily as an inadvertent (or deliberate) omission.

However, the retrieval of information through the *SCI* is not exclusively dependent upon the citation of a single specific work or author. The average scientific paper contains dozens of references. The probability of missing a paper in an *SCI* search becomes reasonably small if the researcher is diligent, especially when pursuing the complete citation network. I say this although many people complain that they frequently retrieve too much too fast rather than miss an "obvious" paper.

Bibliographic Coupling

A fundamental notion in citation theory, one that I have always taken for granted, is that two documents are essentially "identical" if their reference lists are identical. However, in the real world, two papers on the "same" topic rarely cite the identical list of articles. Authors have imperfect memories, so their recollections of the prior literature differ. Their motivations also may differ. Perhaps one author cites 5 or 10 papers that another does not. Each, however, may cite about the same 50 percent of the references. More than likely an even higher percentage of the core papers or books in the field will be co-cited.

The theory behind information retrieval by citation indexing is based on this key assumption—papers are linked together by what they cite. An *SCI* search usually starts with one of the papers that are cited by either of two authors (whose papers are "coupled" on the basis of the number of citations they share). You will be led to either of the two citing papers through the *Citation Index* section of the *SCI*. By "cycling," that is, by examining the *SCI* entry for another reference cited in the first paper retrieved, you will quickly be led to the second. We all do this kind of cycling in our normal reading of papers but are probably not conscious of it. This idea of bibliographic coupling was tested by M.M. Kessler, Massachusetts Institute of Technology, Cambridge, more than 25 years ago.[15]

Recently we implemented a type of retrieval, based on "bibliographic coupling," in the new compact disc (CD-ROM) version of the *SCI*. The new *SCI Compact Disc Edition* employs bibliographic coupling to make instantly available the papers most closely related to the paper you've retrieved—thus rendering "cycling" in some cases an almost superfluous step. For any paper you find in your first step, you are shown the "related records." These are papers that have cited the most references *in common with* your starting paper. In the new *SCI Compact Disc Edition*, 20 such related papers can be traced in order of their coupling strength or relevance.[16] This is the same kind of "relevance" weight we provide in a search of our database of research fronts. However, in the latter case, the citing papers are arranged by the number of core papers cited instead of a computed coupling strength for a particular paper. Indeed, the bibliography of your starting paper defines the initial "core" literature of your search.

I hasten to add that some of the most "interesting," unexpected papers may be those with the *least* coupling. Indeed, some purists argue that the most interesting papers deal with *analogous* problems, whose authors somehow do not cite any references in common. Presumably, it is these kinds of disjunctive connections that make for great discoveries. In a recent paper, Donald R. Swanson, Graduate Library School, University of Chicago, Illinois, questions whether such connections can be facilitated through citation indexing.[17]

No one I know has the time to analyze all the serendipitous connections that could be made through the *SCI*, much less through semantic or other techniques. Artificial intelligence (AI) aficionados promise us all kinds of superconnections, especially when parallel processors are available; but, for the moment, I have my hands full dealing with the riches uncovered by bibliographic coupling. Undoubtedly, once we have implemented these associative processes on a large scale, we can compare the results retrieved by word coupling and then the combinations of both—word and citation coupling. Citation analysts in The Netherlands, such as A.F.J. Van Raan and D. Hartmann, LISBON Institute, University of Leiden, have investigated some of those possibilities.[18] Already in the *SCI Compact Disc Edition* you can start with a combination of words in the *Permuterm® Subject Index* and immediately proceed through bibliographic coupling across the literature of the current year. J.T. Sharabchiev, a Soviet scientist at the Public Health Ministry of Byelorussia, Minsk, USSR, has recently compared the results of clustering by bibliographic coupling and by co-citation in a remarkable study on the history of immunology.[19]

Uncited Work

There are times when we are happy to learn that a particular work has never been cited. In our *SCI* database, covering the years 1955-1987, more than 56 percent of the source items are uncited—not even self-cited. (Many of these source items are abstracts, letters, and editorials, of limited interest; nevertheless, a huge number of papers go uncited.) More knowledge about uncited papers is important for many reasons, not the least of these being that there are so many scientists who believe they are latter-day Mendels. How many of these uncited papers contain identifiable works of unrecognized premature genius? Unfortunately, the problem in finding such examples involves more than just searching our files for the papers that have not been cited. There are, of course, millions of these. However, they may be even less interesting than those that are cited a few times, either by their own authors or by others who apparently could not convince the world of their significance. In reality, most uncitedness is probably due to the fact that our earlier papers are superseded by those we publish later. Eventually our own review papers, or those by others, make it superfluous to cite earlier papers.

Conclusion

Individual citation behavior, like other behavior, is quite varied. However, the collective behavior of small and large groups of scholars produces reasonably predictable results. There are some groups of scientifically trained professionals, for example, computer scientists and engineers, who have not adopted the norms of citation behavior. These people need to be made aware of the importance of documentation. Their works require a special form of refereeing that includes mandatory literature searching.

Those of you who feel that there is a significant lack of acknowledgment of previous work in average citation behavior today must ask whether it is likely that we can automate the intellectual process of providing bibliographies in papers in support of conscientious refereeing. In the early

1960s, the possibility of artificially intelligent documentation was considered.[20] After nearly 30 years, we are still not significantly closer to realistic, unaided, automatic provision of pertinent references. We can visualize computer-aided documentation, but completely unaided AI analysis of scientific texts would result in choices that would barely approach the intelligent choices that could be made by a student. Expert systems using syntactic analysis might provide clues to where references would be appropriate. However, the actual selection of prior references would be limited to modeling text vocabularies, when and if full texts of the journal literature become available.

Whether and when we can imitate human citation behavior is problematic. The very attempt to do so makes us all the more conscious of the special human intelligence it requires.

* * * * *

My thanks to Elizabeth Fuseler-McDowell and Sanaa Sharnoubi for their help in the preparation of this essay.

REFERENCES

1. **Garfield E.** Library of the Hungarian Academy of Sciences builds computerized information services on ISI's data base. *Essays of an information scientist.* Philadelphia: ISI Press, 1983. Vol. 5. p. 4-6.
2. **Braun T, Glänzel W & Schubert A.** *Scientometric indicators.* Singapore: World Scientific, 1985. 424 p.
3. **Braun T, Bujdosó E & Schubert A.** *Literature of analytical chemistry: a scientific evaluation.* Boca Raton, FL: CRC Press, 1987. 259 p.
4. **Vinkler P.** A quasi-quantitative citation model. *Scientometrics* 12:47-72, 1987.
5. **Brooks T A.** Private acts and public objects—an investigation of citer motivations. *J. Amer. Soc. Inform. Sci.* 36:223-9, 1985.
6. **Kochen M.** How well do we acknowledge intellectual debts? *J. Doc.* 43:54-64, 1987.
7. **Garfield E.** Breaking the subject index barrier—a citation index for chemical patents. *J. Patent Office Soc.* 39:583-95, 1957. (Reprinted in: *Essays of an information scientist.* Philadelphia: ISI Press, 1984. Vol. 6. p. 472-84.)
8. **Moravcsik M J.** Personal communication. 6 March 1989.
9. **Pullum G K.** Citation etiquette beyond Thunderdome. *Natur. Lang. Linguist. Theor.* 6:579-88, 1988.
10. **Garfield E.** Information theory and all that jazz: a lost reference list leads to a pragmatic assignment for students. *Essays of an information scientist.* Philadelphia: ISI Press, 1980. Vol. 3. p. 271-3.
11. ----------. Information theory and other quantitative factors in code design for document card systems. *J. Chem. Doc.* 1:70-5, 1961. (Reprinted in: *Essays of an information scientist.* Philadelphia: ISI Press, 1980. Vol. 3. p. 274-85.)
12. **Cronin B.** *The citation process: the role and significance in scientific communication.* London: Graham, 1984. 103 p.
13. ----------. Personal communication. 10 March 1989.
14. **Kaplan N.** The norms of citation behavior: prolegomena to the footnote. *Amer. Doc.* 16:179-84, 1965.
15. **Kessler M M.** Bibliographic coupling between scientific papers. *Amer. Doc.* 14:10-25, 1963.
16. **Garfield E.** Announcing the *SCI Compact Disc Edition*: CD-ROM gigabyte storage technology, novel software, and bibliographic coupling make desktop research and discovery a reality. *Current Contents* (22):3-13, 30 May 1988.
17. **Swanson D R.** Two medical literatures that are logically but not bibliographically connected. *J. Amer. Soc. Inform. Sci.* 38:228-33, 1987.
18. **Van Raan A F J & Hartmann D.** The comparative impact of scientific publications and journals: methods of measurement and graphical display. *Scientometrics* 11:325-31, 1987.
19. **Sharabchiev J T.** Cluster analysis of bibliographic references as a scientometric method. *Scientometrics* 15(1-2):127-37, 1989.
20. **Garfield E.** Can citation indexing be automated? (Stevens M E, Giuliano V E & Heilprin L B, eds.) *Statistical association methods for mechanized documentation. Symposium proceedings*, 1964, Washington, DC. Washington, DC: National Bureau of Standards, 15 December 1965. Misc. Pub. No. 269. p. 189-92. (Reprinted in: *Essays of an information scientist.* Philadelphia: ISI Press, 1977. Vol. 1. p. 83-90.)