Can RANK Be Used to Generate a Reliable Author List for Cocitation Studies?

Jeff Alger

This study investigates the possibility of using DIALOG's RANK command to generate a list of prominent authors for use in cocitation studies. The emerging field of biodiversity is used primarily because it represents a new and rapidly expanding field of study. The results indicate that RANK does not effectively retrieve a quality set of prominent authors for use in cocitation studies. Highly cited authors of general texts on biodiversity cause the derived author map to present a misaligned picture of specialization within the field. By limiting citations to only journal articles, a clearer and more accurate picture of the field should emerge.

Biodiversity—The variety of organisms considered at all levels, from genetic variants belonging to the same species through arrays of species to arrays of genera, families, and still higher taxonomic levels; including the variety of ecosystems, which comprise both the communities of organisms within particular habitats and the physical conditions under which they live.

—E. O. Wilson[1]

Science has often been thought of in terms of the totality of the literature that comprises the body of scholarship inherent in scientific research. In the past two decades, efforts have been made to examine the specialties and subdisciplines within broader fields of interest in both the natural and social sciences.

Derek J. de Solla Price postulated that scientific literature is "knitted" together by the associations that citing papers make to previous works within a field. He ascertained that highly cited papers can be arranged in a matrix so that a "research front" is apparent. This front comprises the area within the discipline in which the most active research is taking place.[2] By analyzing the scientific literature for relationships within disciplines, a descriptive map can be derived that represents the areas of specialization within those disciplines. Cocitation analysis provides a means of defining these relationships.

Cocitation: An Introduction

Cocitation is the frequency in which two documents or authors are cited together by more recent papers. Henry G. Small and Belver C. Griffith state:

Jeff Alger is the Life Sciences Reference Librarian at Kansas State University; e-mail: jalger@ksuvm.ksu.edu.

> Cocitation is a relationship which is recognized and maintained by current researchers. This dependence on the population of current citing authors is intrinsic to cocitation, the patterns of which change with time as new discoveries are made and introduced through the literature.[3]

The basic assumption behind cocitation is that documents that are frequently cited together by succeeding works are related in subject matter. The strength of this relationship is viewed as correspondingly intensifying as the cocitation frequency of the two documents increases. The strength of the cocitation link between papers provides a measure by which a researcher can quantitatively construct descriptive maps of subject specialties.

The fundamental steps in cocitation are the selection of documents or authors for analysis; retrieval of cocitation frequencies, usually by searching the Science Citation Index (SCI) or the Social Sciences Citation Index (SSCI); compilation of a raw cocitation matrix; conversion of the raw data matrix to a correlation matrix; analysis of the correlation matrix through nonmetric multidimensional scaling (MDS); and finally, interpretation and validation of the results.[4]

Some of the problems inherent in cocitation studies are similar to those in other citation methodologies, such as a lack of consistency in name authority of cited authors and selection of prominent documents or authors within a given field.

The selection of the author set to be used in the study is, quite obviously, the key component in the outcome of the analysis. In most cases, the author set is selected through consultation with a panel of experts in the field being studied.[5] By soliciting the opinions of experts in the field in order to construct the author set, the researcher introduces an element of bias in the study. Although this bias may be reduced by consulting a wider panel of experts, there are limitations on how many experts a researcher can realistically expect to receive feedback from. The question that arises is: Can an unobtrusive method be derived for generating the prominent author list that relies upon the citation patterns themselves?

This study explores using the DIALOG RANK command in SCISEARCH (the online version of SCI) for deriving a list of prominent authors for cocitation analysis in the research area of biodiversity. Biodiversity represents an incredibly active field of research that has emerged in the literature only fairly recently. Using biodiversity as the field of study presents an opportunity to examine a rapidly expanding body of literature with a variety of subspecialties.

The primary purpose of this study is to test the effectiveness of using this method of retrieval to generate a prominent author list. If the methodology is effective, it will allow exploration of the literature in fields in which the researcher may have limited experience or background. It also will help to eliminate one of the basic problems in cocitation studies—the selection of the prominent authors.

As academic librarians begin to move into more active roles with research faculty, they will be expected to manage the copious flow of information in the field being studied, whether it is biological diversity or nineteenth-century poets. This methodology for cocitation research will provide assistance for the librarian conducting bibliometric research in fields where they may not know the prominent authors or have access to a panel of experts. Librarians can quickly identify important researchers and target these authors as starting points for further research.

Seminal Studies

Henry G. Small was among the first to explore the use of cocitation as a means of examining the relationships between two documents.[6] Small and Griffith used cocitation to "identify clusters of highly interactive documents in science," which they contended represented active specialties.[7] They concluded that the mapping of specialties through cocitation was important in order to

show their internal structure as well as their relationships to one another.

Griffith successfully applied cocitation analysis to the social and behavioral sciences' literature. Previously, cocitation had been used only in the analysis of the natural and physical sciences literature.[8]

Howard D. White joined Griffith in conducting several cocitation studies involving cocited authors as the unit of analysis rather than cocited documents. The first of these studies was conducted on authors in the field of information science.[9]

The assumption that two documents that are cocited frequently are related in content applies to authors as well; those who are cocited frequently exhibit a closer relationship. White and Griffith maintained that working with cocited authors had many advantages over working with cocited documents. Their primary reason was that much less information was needed in order to search either SCI or SSCI than is necessary with document searching. With a limited knowledge of a given field (all one needs is a list of prominent authors), a map can be generated for any field or specialty, no matter how small. A final advantage they mentioned is that thirty to forty authors represent a much wider sample of a field than do thirty to forty documents.[10]

In a series of papers, H. P. F. Peters and A. F. J. van Raan mapped the chemical engineering field by studying top scientists in the field. They identified the set of scientists to be used in the study based on the quantity of publications produced by each, utilizing the assumption that those who publish the most produce the best research.[11,12] In a later paper, Peters and van Raan compared the top scientists in chemical engineering as defined by their methodology with a group of "average scientists." They presented a bibliometric profile of these top scientists based on their findings, which lends support to their hypothesis that quantity equals quality.[13]

Sean B. Eom used author cocitation analysis coupled with factor analysis to study decision support systems (DSS) research. By using factor analysis, Eom was able to identify seven clusters of subspecialties in DSS research.[14]

Methods

Author Selection

The author conducted an online search in SCISEARCH (DIALOG, File 34, 1988–present) to generate a list of potential authors to be used in the study. The search statement "s biodiversity or biological(w)diversity and py>1989" retrieved 716 documents. This statement searched for two common variations of the term *biodiversity* and limited retrievals to documents published after 1989. The RANK command ("rank ca cont") generated a ranked list of authors cited in the retrieved documents and provided for continuous output of the results.[15] The

Librarians can quickly identify important researchers and target these authors as starting points for further research.

RANK command considers the number of documents that cite an author, not the number of actual citations within those documents when generating a ranked list. (For instance, if an author had three different papers cited in a single document, it would only count as one citation instance when using "rank ca.") To keep the number of cocited pairs to a manageable number, this study examined only authors cited in twenty or more documents.

Data Collection

The author searched all author pair combinations in SCISEARCH to determine raw cocitation frequencies using a statement similar to "s cr=wilson-e? and cr=myers-n?." This statement searched for references by the cited authors' name.

TABLE 1
List of Authors Retrieved by RANK

Rank Number	Items Retrieved	Author
1	89	Wilson, E. O.
2	67	Soule, M. E.
3	49	May, R. M.
4	48	MacArthur, R.
5	47	Ehrlich, P. R.
6	46	Myers, N.
7	44	Simberloff, D.
8	34	Terborgh, J.
9	31	Lande, R.
10	31	McNeely, J. A.
11	30	Harris, L. D.
12	30	Wilcove, D. S.
13	28	Janzen, D. H.
14	27	Erwin, T. L.
15	27	Frankel, O. H.
16	27	Vitousek, P. M.
17	24	Diamond, J. M.
18	24	Pimm, S. L.
19	24	Reid, W. V.
20	23	Noss, R. F.
21	22	Gilpin, M. E.
22	20	Scott, J. M.
23	20	Peters, R. L.

Truncation of the authors' names accounted for variation in citation styles. The author constructed a raw data matrix from the resulting cocitation frequencies. Calculation of the diagonals followed White and Griffith's procedure of taking the three highest intersections and dividing by two.[16] Conversion of this raw data matrix to a correlation matrix provided a mechanism for subsequent MDS analysis.

Data Analysis

All data analysis used SAS version 6.09. The CORRELATIONS procedure generated the correlation matrix. Converting the raw cocitation frequencies to Pearson's *r* correlation coefficient removes the differences between highly cited authors and less frequently cited authors. The correlation coefficient also provides a measure of similarity between cocited author pairs in addition to the raw cocited frequencies.[17]

The CLUSTER program (Ward's method) clustered the authors using the correlation coefficients. Clustering continued until approximately 85 percent of the variation could be accounted for. Using the correlation matrix, a nonmetric multidimensional scaling program generated a proximity map. The author drew outlines around the resulting clusters to facilitate interpretation of the proximity map.

The MDS procedure (formerly ALSCAL) produced the author map. The resulting map places the authors according to their similarity (or dissimilarity) to the other authors in the matrix. Authors with high similarities are placed close together on the map whereas those with high dissimilarities are placed farther apart.[18]

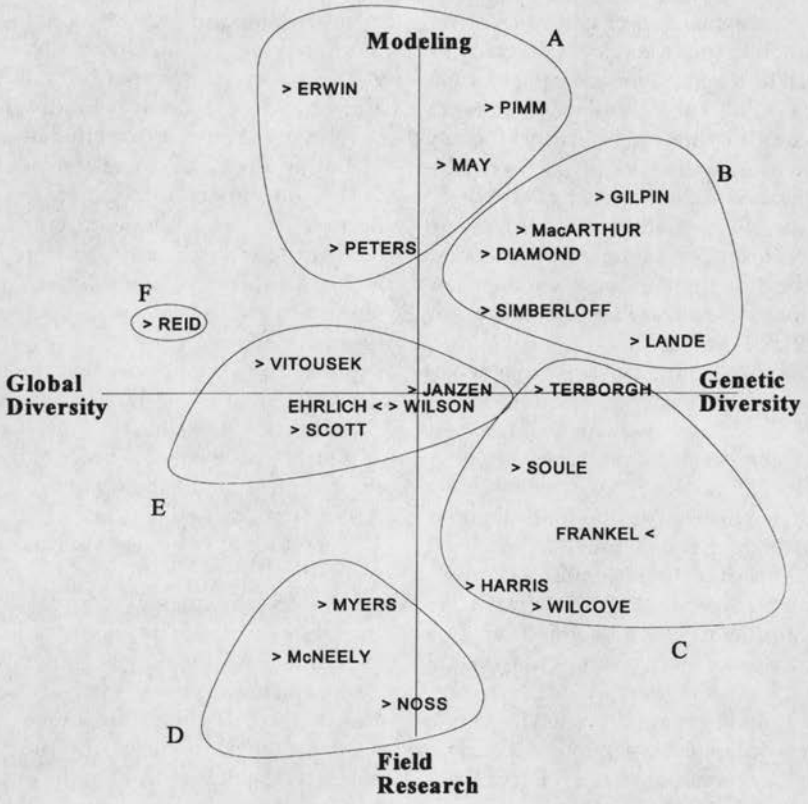
Feedback from current researchers in biodiversity through posting of the resulting clusters on the BIODIV-L list provided for a measure of the validation of the results. The author asked respondents to label individual author interests as well as the clusters. Several respondents provided comments on the makeup of the clusters as well.

Results

Table 1 is a list of the authors generated from the SCISEARCH database using the RANK command. To limit the matrix to a manageable size, table 1 includes only those authors cited by twenty or more documents. The number of citations of authors in this study ranged from the minimum threshold of twenty to a high of eighty-nine.

Raw cocitation frequencies ranged from a high of 247 to a low of zero. Each cocitation instance represents a document in which both of the authors were cited at

FIGURE 1
Cocitation Map of Prominent Authors in the Field of Biodiversity



least once. R. H. MacArthur was cocited the most often with the other authors in the study (1,617 times) and W. V. Reid was cocited the least often (57 times).

The author generated the correlation matrix using Pearson's *r* correlation coefficients for each cocitation pair across all author pairs, exclusive of the two being compared. The correlation coefficient thus functions as a measure of similarity. Higher positive correlations indicate authors perceived to be more similar by citing authors. Conversely, lower correlations indicate that citing authors perceive the two authors to be less related because of either research interest or some other factor, such as research methodology. For example, individual authors may tend to conduct theoretical or modeling research whereas others may engage in field work to achieve their results. Thus, authors may be cocited together because of related subject emphasis, but on the MDS map they may be separated either horizontally or vertically as a result of differing methodologies.

Figure 1 presents a two-dimensional map of cocited author positions in relation to other authors in the study. Author clusters derived by the MDS procedure also are outlined on this map. Authors with links to many other authors are placed closer to the center axis of the map. Those whose research is more pe-

ripheral or who have fewer links to other authors due to factors such as recency of publication are placed farther from the center. Placement along the horizontal axis represents subject emphasis of the author, ranging from global diversity on the left to genetic diversity of populations on the right. The vertical axis represents research methodology. Authors using theoretical methodologies such as mathematical modeling appear near the top and those using field research fall toward the bottom. The center of the map can best be thought of as representing a general overview of biodiversity covering all aspects of the field.

The six resulting clusters explained 84 percent of the variance. The authors initially broke into two clusters, basically along the vertical axis. N. Myers, J. A. McNeely, and R. F. Noss broke into a separate cluster with the third iteration. W. V. Reid separated from P. M. Vitousek, P. R. Ehrlich, E. O. Wilson, D. H. Janzen, and J. M. Scott on the sixth iteration.

Nine members of BIODIV-L from three countries offered comments on the makeup of the clusters. Two of the respondents provided an abbreviated list of how they would cluster some of the authors based on perceived research interests.

Discussion

Comments received from active researchers in the field provided interesting feedback on the makeup of the clusters, as well as on the work and emphasis of the authors used in the study. In most cases, the researchers maintained that the inclusion of certain authors in the individual clusters did not mesh with the research interest of other authors in the cluster. The two respondents who offered suggestions on how they would have arranged the clusters simply listed those authors they perceived as belonging together and excluded those they could not fit with the others. None of the respondents could readily place a label on any of the clusters.

As seen in figure 1, letters, rather than labels, are provided for the clusters themselves. This is attributable to the fact that most of the clusters cannot be identified in any meaningful way. Many of the authors making up the clusters have diverse research interests both within and outside the field of biodiversity. Several of the authors are cited because of books they have written dealing with biodiversity in a general sense. Because of the general nature of these books, these authors are clustered together with other authors who have written similar books, and such works sometimes are referred to in an almost token manner. This is not to say that these are not substantial works but, rather, by including such items in the study (many of which are ten or more years old), that the current state of the literature is obscured and diffused.

Robert M. May appears in this study primarily because he wrote the "News and Views" column in the journal *Nature* for many years and was thus called on to comment on a variety of conservation issues, including biodiversity. Much of his cited work is from this source. He is a mathematical modeler who primarily conducts theoretical work at the population level. Terry L. Erwin has written extensively on arboreal insect populations and their use as biodiversity monitors. Robert L. Peters and Stuart L. Pimm write on ecological restoration and community structure, respectively. It is difficult to determine why these authors should cluster together given their widely differing interests.

In cluster B, Daniel Simberloff and Jared M. Diamond cluster together primarily because of their lengthy altercations on various conservation issues. Recently, however, their interests have become far more separated. Michael E. Gilpin and Russell Lande study quantitative population genetics and belong together but should not be grouped with the others in this cluster.

In cluster C, John Terbough, David S. Wilcove, Michael E. Soule, and Larry D. Harris are grouped together primarily because of their concern with fragmentation issues. Terborgh and Harris write primarily about the effect of fragmentation on biodiversity and species survival. Wilcove has more recent book chapters and has written a key paper on experimental fragmentation. Otto H. Frankel is a geneticist who has published extensively on plant genetic resources. Frankel and Soule coauthored a book in 1981 concerning the genetic resources of threatened populations, but otherwise it is hard to place Frankel in the same cluster as these other authors. Based on research interests, Frankel would appear to fit better in a sperate cluster with other geneticists such as Gilpin, Soule, and Lande.

Norman Myers, Jeffrey A. McNeely, and Reed F. Noss have all written on protected areas and their impact on the larger biological diversity picture. These three authors provide the best makeup of any of the clusters given their similar research interests. Clusters C and D could tentatively be labeled "Fragmentation" and "Protected Areas," respectively.

With the exception of J. Michael Scott, cluster E can best be described as the "grand old men" of the biodiversity field. They have all contributed to the general understanding of biodiversity but, other than that, they have little in common. In fact, each has diverse research interests, from butterflies to forest restoration, for example.

Walter V. C. Reid broke off from this central group with the last iteration. He authored a text in 1989 that is widely cited as providing the scientific basis for biodiversity studies. His more recent interests include coastal biodiversity and the effects of development on coastal regions.

Much of this overlap and clustering of nonsimilar authors was apparent after deriving the correlation matrix. None of the author pairs exhibited a high correlation. The highest correlation coefficient was between MacArthur and Diamond at 0.4243. Approximately 50 percent of the correlations were less than 0.1000. This would indicate that subsequent clustering may not provide meaningful results. However, in order to analyze the methodology fully, the data were entered into both the CLUSTER and the MDS programs.

The main problem in this study is that the resulting clusters basically obscure the methods, breadth, study organisms and areas, and current research interests of the authors in question. The clusters, as they stand, provide little meaningful insight into the state of biodiversity research and serve to lead one astray rather than to help define the field.

The clusters, as they stand, provide little meaningful insight into the state of biodiversity research. . . .

Conclusion

It appears that many citations are to authors who have written general texts on biodiversity. This can possibly be attributed to the relative newness of the field and to researchers seeking a foundation for their studies. The high citation rate of these general texts, coupled with these authors' past and current research interests, tends to produce a misaligned picture of the field using this methodology. Using the RANK command to derive the author list appeared to work well, though it became obvious that many of the most highly cited items were introductory books covering multiple facets of biodiversity rather than journal articles focusing on a specific area. The high citation frequency of these general texts served to obscure the underlying subject specialties of the field.

Limiting citations to only journal articles would produce a clearer picture which would better define the specialties within the field. Future studies us-

ing RANK should limit citations to only journal articles in the author selection procedure. Currently, limiting citations for use in cocitation studies to only journal articles would have to be done manually because no field delimiter is available in SCISEARCH to separate books from articles during the actual search.

The prominent authors in the field are changing rapidly as research in biodiversity continues. One year after the original data were collected, a follow-up search in SCISEARCH produced a substantially different ranked list of authors. Many of the prominent authors in this study were retrieved once again in the follow-up, though not at the same positions. Perfection of this methodology would facilitate the tracking of research fronts in rapidly developing fields such as biodiversity by researchers with little or no previous specific knowledge of the prominent authors in the field.

The question asked in the title of this study remains to be answered. Can RANK be used to generate a reliable author list? Possibly, though refinement of the methodology presented here would have to take place in order to produce an author set that would provide a clear picture of the field being studied.

Notes

1. Edward O. Wilson, *The Diversity of Life* (Cambridge, Mass.: Belknap Pr., 1992).
2. Derek J. de Solla Price, "Networks of Scientific Papers," *Science* 149 (July 30, 1965): 510–15.
3. Henry G. Small and Belver C. Griffith, "The Structure of Scientific Literatures I: Identifying and Graphing Specialties," *Science Studies* 4 (Jan. 1974): 17–40.
4. Katherine W. McCain, "Mapping Authors in Intellectual Space: A Technical Overview," *Journal of the American Society for Information Science* 41 (Sept. 1990): 433–43.
5. Howard D. White and Katherine W. McCain, "Bibliometrics," *Annual Review of Information Science* 24 (1989): 119–86.
6. Henry G. Small, "Cocitation in the Scientific Literature: A New Measure of the Relationship between Two Documents," *Journal of the American Society for Information Science* 24 (Aug. 1973): 265–69.
7. Small and Griffith, "The Structure of Scientific Literatures," 39.
8. Belver C. Griffith, "The Social and Behavioral Sciences' Literature," *Information Choices and Policies: Proceedings of the 1979 ASIS Annual Meeting* 16 (Oct. 1979): 254–62.
9. Howard D. White and Belver C. Griffith, "Author Cocitation: A Literature Measure of Intellectual Structure," *Journal of the American Society for Information Science* 32 (May 1980): 163–71.
10. Ibid., 164.
11. H. B. F. Peters and A. F. J. van Raan, "Co-word Based Science Maps of Chemical Engineering, Part I: Representations by Direct Multidimensional Scaling," *Research Policy* 22 (Feb. 1993): 23–45.
12. ———, "Co-word Based Science Maps of Chemical Engineering, Part II: Representations by Combined Clustering and Multidimensional Scaling," *Research Policy* 22 (Feb. 1993): 47–71.
13. ———, "A Bibliometric Profile of Top-Scientists: A Case Study in Chemical Engineering," *Scientometrics* 29 (Jan. 1994): 115–36.
14. Sean B. Eom, "Decision Support Systems Research: Reference Disciplines and a Cumulative Tradition," *Omega: International Journal of Management Science* 23 (Oct. 1995): 511–23.
15. "RANK in the ISI Citation Indexes," *Chronolog* 21 (Apr. 1993): 116–17.
16. White and Griffith, "Author Cocitation," 165.
17. McCain, "Mapping Authors," 436.
18. Ibid., 438.