Back to Introduction

Mapping the social sciences: the contribution of technology to information retrieval

EUGENE GARFIELD *ISI (USA)*
ROBERT KIMBERLEY *ISI (UK)*
DAVID A. PENDLEBURY *ISI (USA)*

INTRODUCTION

By virtue of the computer's storage capacity, its powers of speed and specificity in retrieval and, above all, its economy, technology has reshaped knowledge classification.

At least since the time of Plato and Aristotle—even before their era if we wish to consider mythographers—humans have been ardent classificationists. It is obvious, however, that human subjective judgment produces taxonomies that partially reflect objective reality and partially the mind of the taxonomer. John H. Finley, Jr., in writing about how the early Greeks ordered their world, observed that "thought proceeds by scheme and sequence; it manipulates, puts things where it wants them, makes different designs from any that the eyes see."[1] (p. 8) Human classification schemes, such as subject heading categories, are, then, inherently subjective, owing to the perceptions upon which they are based.

The alternative is an objective or natural system of classification in which the attributes of objects (their similarities or differences) are the defining elements. Such a system of classification, while theoretically possible, was not a practical pursuit without computer technology.

It is assuredly not the aim of this essay to describe the manifold ways in which information technology (IT) is being exploited today to aid researchers in the social and behavioral sciences. Nor do we intend to comment on how this IT has changed the nature and type of research projects undertaken by social scientists. (It is plain, however, that quantification has been a hallmark of the social sciences since the Second World War, and it is no coincidence that researchers became increasingly interested in quantitative studies at the same time that the introduction of computers made such activities feasible.) Rather, this chapter focuses on the efforts of the Institute for Scientific Information® (ISI®), a producer of computer-based information products for researchers in the sciences, social sciences, and humanities, to create a natural system of classifying knowledge (or, more narrowly, research activity) through the use of citation indexing and, more recently, "geographic" maps of research through co-citation clustering.

CITATION INDEXING

E. Garfield applied the principle of citation indexing to the academic literature.[2] Citation indexing was first used in *Shepard's Citations*, an index for the legal profession to precedents of the Federal and State courts. In drawing an analogy between the progression of legal decisions based on precedents, and scientific research based on previously published results, Garfield imagined the utility of citation indexing in the scientific literature.[3]

The principle underlying citation indexing is as follows: if one paper cites an earlier publication, they bear a conceptual relationship to one another. The references given in a publication thus serve to link that publication to earlier knowledge. Implicit in these linkages is a relatedness of intellectual content. In reordering the literature by works cited, we obtain a citation index. Citation indexing is a natural or automatic system of classification: the material to be classified orders itself through its conceptual links.[4]

After succeeding in developing a citation index to the scientific literature—the *Science Citation Index®* (*SCI®*)—Garfield applied the technique to the literature of the social sciences.[5] Since 1966 ISI has published the *Social Sciences Citation Index®* (*SSCI®*). In 1985 the *SSCI* fully covered about 1,500 journals and selectively covered some 3,300 more, for a total of about 4,800 journals representing over twenty-five different fields. In 1985 alone over 120,000 articles, reviews, let-

ters, editorials, abstracts, etc., and nearly 1.5 million references from these items were indexed. The *SSCI* has become an important tool for researchers in the social sciences. Since a citation index gives access not only to the publications indexed, but also to cited works, the *SSCI* is multidisciplinary in scope. Moreover, the user of a citation index is not limited to retrospective searching. The *SSCI* reveals what current publications have cited an older work. Searching forward in time is a chief strength of citation indexes.

A significant by-product of producing the *SCI* and the *SSCI* is the enormous data base ISI creates. This data base contains the citations given by all the articles indexed. The file can be sorted in various ways to reveal the networks of publications on specific subjects. ISI's data bases have been an important source for information scientists and others working in the field of scientometrics or quantitative studies of the history and sociology of science. H. Small, D. Crane and B.C. Griffith demonstrated that citation data could also reveal the structure of research in the social sciences as well.[6,7] The methodology for manipulating ISI's citation data base to reveal these structures is known as co-citation analysis.

CO-CITATION ANALYSIS AND CLUSTERING

Co-citation analysis measures the frequency with which two documents are cited together. Highly co-cited publications are almost always closely related in content or context of use. Co-citation analysis is the inverse of M.M. Kessler's idea of bibliographic coupling: the number of references a given pair of documents have in common is a measure of their proximity of subject.[8] Small, who pioneered co-citation analysis,[9] has demonstrated how a group of co-cited papers can be organized into discrete and meaningful units, called clusters.[10,11] Clusters are networks of interrelated, co-cited publications. When the data base is sorted for a certain year, research fronts (active areas of current research), consisting of related and highly cited articles of a given year and the group of core, co-cited documents they share, can be identified. Co-citation strength is indicative of strength of intellectual connections. Co-citation analysis, therefore, has revealed the speciality structure of knowledge. Some specialities that are identified are new, owing to the automatic or natural organizing process that the citation linkages permit.

A brief explanation of how ISI uses its citation data base and the technique of co-citation analysis to identify clusters of core documents in speciality areas follows.

To begin, the data files of the *SCI* and the *SSCI* covering a single year are combined and sorted for works cited above a certain threshold (typically, five citations). This process, which focuses attention on only relatively active research, greatly reduces the number of publications to be considered. To ensure a balanced representation across

Table 1: List of cited core documents in 1984 C1 cluster #4940.

Cross M. *New firm formation and regional development.* Farnborough, UK: Gower, 1981.
Fothergill S & Gudgin G. *Unequal growth: urban and regional employment change in the U.K.* London: Heinemann, 1982.
Freeman C, Clark J & Soete L. *Unemployment and technical innovation: a study of long waves and economic development.* Westport, CT: Greenwood, 1982.
Granger C W J. *Spectral analysis of economic time series.* Princeton, NJ: Princeton University Press, 1964.
Gudgin G. *Industrial location processes & employment growth.* Farnborough, UK: Saxon House, 1978.
Lewis W A. *Growth and fluctuations, 1879-1913.* London: Allen & Unwin, 1978.
Lloyd P E. *Regional statistics. No. 16.* London: Her Majesty's Stationery Office, 1982.
Lotka A J. *Elements of mathematical biology.* New York: Dover, 1956.
Mandel E. *Late capitalism.* London: NLB, 1975.
Massey D B & Meegan R. *The anatomy of job loss: the how, why, and where of employment decline.* London: Methuen, 1982.
Mensch G. *Stalemate in technology: innovation during the depression.* Cambridge, MA: Ballinger, 1979.
Rostow W W. *The world economy: history and prospect.* Austin, TX: University of Texas Press, 1978.
Rothwell R & Zegveld W. *Industrial innovation and public policy: preparing for the 1980s and the 1990s.* Westport, CT: Greenwood, 1981.
Rothwell R & Zegveld W. *Innovation and the small and medium sized firm: their role in employment and economic change.* Hingham, MA: Kluwer-Nijhoff, 1982.
Rothwell R & Zegveld W. *Technical change and employment.* New York: St. Martin's Press, 1979.
Schumpeter J A. *Business cycles: a theoretical, historical & statistical analysis of the capitalist process.* New York: McGraw-Hill, 1939.
Slutzky E. The summation of random causes as the source of cyclic processes. *Econometrica* 5:105-46, 1937.
Storey D J. *Entrepreneurship and the new firm.* Beckenhem, UK: Croom Helm, 1982.
Van Duijn J J. *The long wave in economic life.* London: Allen & Unwin, 1983.

disciplines, ISI employs a weighting technique known as fractional citation counting, which entails assigning a unit of strength to each current year based on the number of references it lists. After meeting the integer threshold and that of fractional citation weight, every pair of papers left in the set is measured for co-citation strength.

The foregoing process reduces the original data file of 6 million cited documents to a group of roughly 70,000 and results in a giant network of co-cited papers linking all fields. To break this giant cluster into smaller clusters, the co-citation strength threshold is raised. A cluster that is meaningful as a discrete unit usually contains no more than sixty core papers. At this level about 9,000 clusters emerge, each one corresponding to a subspeciality. The group of current year papers and the core documents co-cited (the cluster) make up a single research front. A subject specialist then examines the research front and, with the help of an index of frequently occurring words in the citing and core publications, names the unit.

For example, a cluster named "Regional growth and economic development in the UK due to technological innovation and formation of firms" (#84-4940) was identified in the 1984 *SCI/SSCI* file. This cluster contains 130 citing articles from 1984 and the nineteen core documents co-cited by them. A few of the core documents are quite old and most are monographs rather than articles (Table 1). This group of core documents illustrates the chief differences Small and Griffith observed in their comparative study of the structure of science and social science research: "in contrast to the natural sciences, the social and behavioral sciences utilize older documents and place greater emphasis on scholarly monographs."[7] (p. 4)

At this stage, clusters are clustered together. The lowest level, representing research fronts composed of individual publications, is known as the C1-level. The first iteration of the computer, creating a cluster of clusters, is the C2-level. There are five iterations in all. At the C5-level, a global view of research is obtained. In other words, the C5-level represents one giant cluster of knowledge.

MULTIDIMENSIONAL-SCALING MAPS

What is achieved in clustering is a matrix of objects linked together by varying degrees and in different states of aggregation. In order to repre-

Figure 1: The 1984 C1 cluster 4940 "Regional growth and economic development in the UK due to technological innovation and formation of firms." Each node (accompanied by surname and year) represents a core document in the cluster.

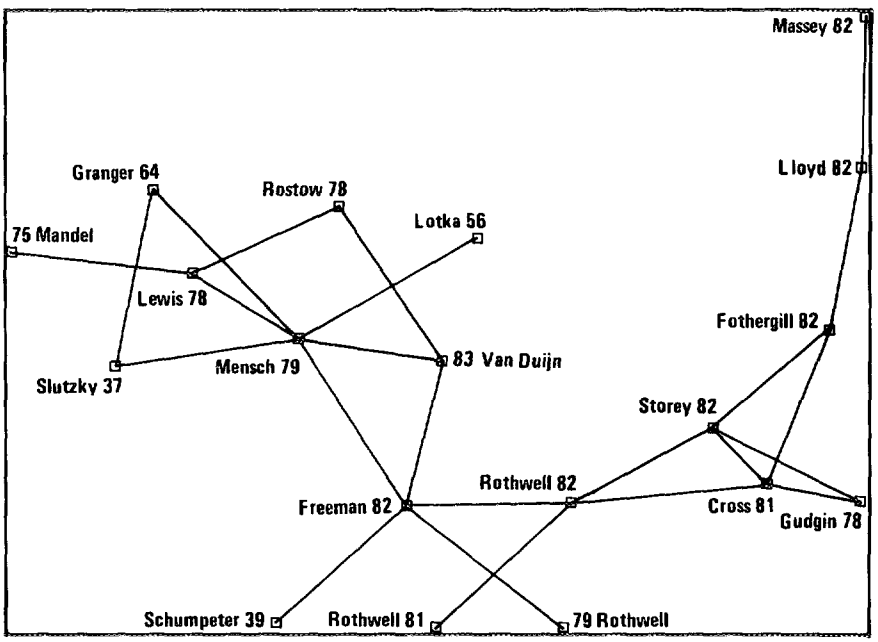
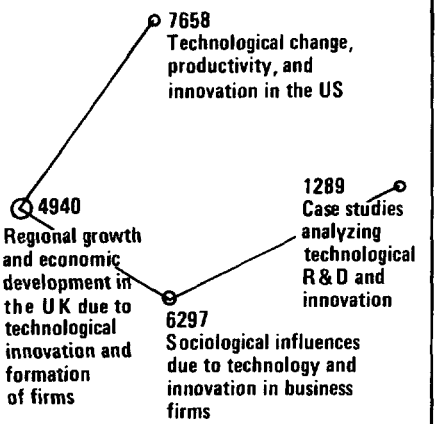


Figure 2: The 1984 C2 cluster 516 "Sociological repercussions of technology and innovation in the UK, the US, and other countries."



sent these relationships graphically, ISI uses multidimensional-scaling mapping,[12,13] also known as similarity mapping.[14] Small and Garfield used the analogy of the relation between a road map's table of distances and the map itself to describe the process of multidimensional-scaling:

> Imagine taking a map of the United States and constructing a table showing the distances between all major cities. Our problem is the reverse. We have the table of distances (or actually degrees of closeness) but lack the map embodying those distances. This is what the scaling technique provides in, of course, an approximation.[15]

Such a map has no absolute axis, but is only a representation of related things, in two dimensions, wherein distance signals the degree of relatedness. Those (at the C1-level) or clusters (C2–C5 levels) lying closest to the center of the map are the most highly co-cited, while those at the margins are least co-cited and, therefore, weakest in their relatedness of subject content.

Figure 1 is a multidimensional-scaling map at the C1-level of the links between individual core publications in research front #84-4940. In Figure 2 the cluster has been clustered with three others: "Case studies analyzing technological research and development and innovation" (#84-1289), "Sociological influences due to technology and innovation in business firms" (#84-6297) and "Technological change, productivity, and innovation in the United States" (#84-7658). The four clusters taken together make up the aggregate

Figure 3: The 1984 C3 cluster 76 "Sociology."

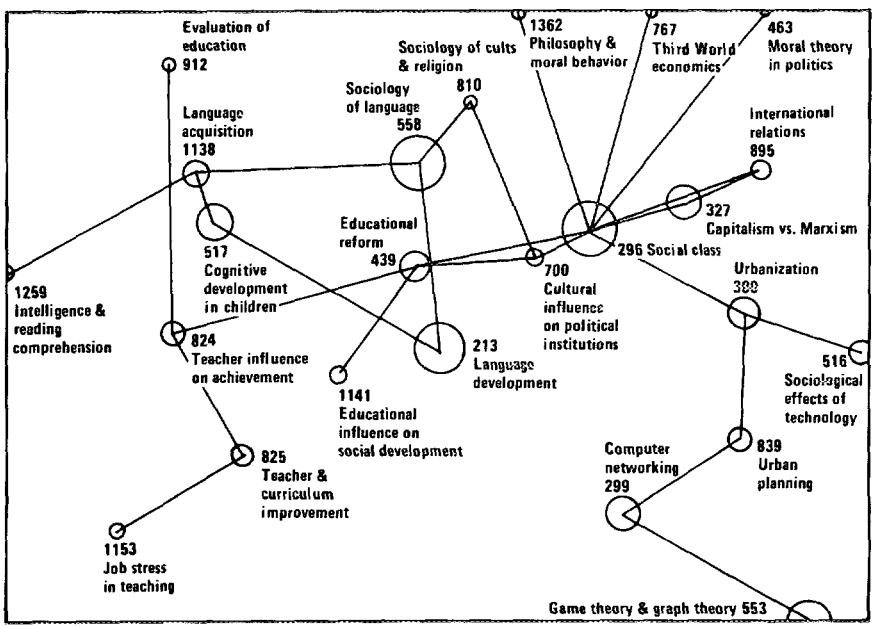


Figure 4: The 1984 C4 cluster 1 "Natural and social sciences."

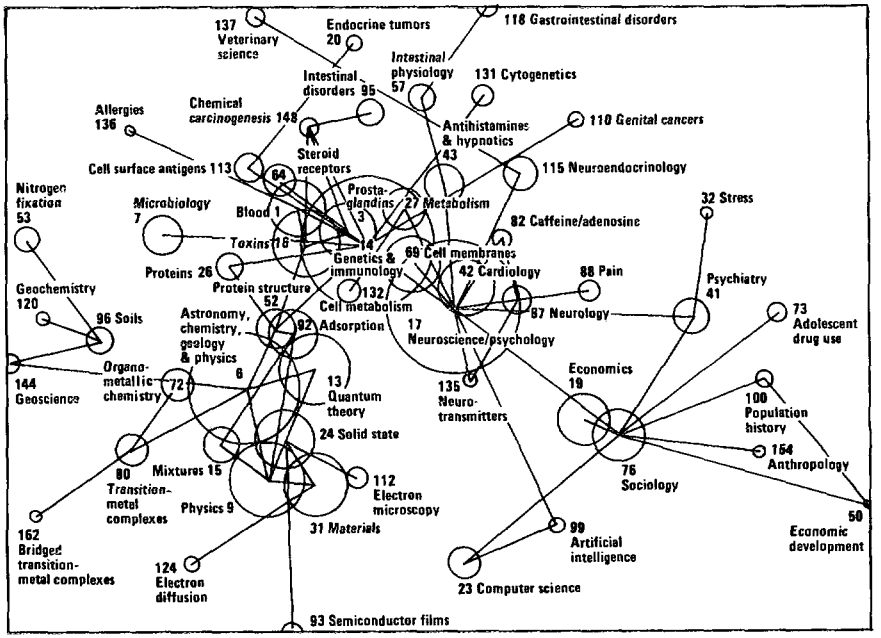
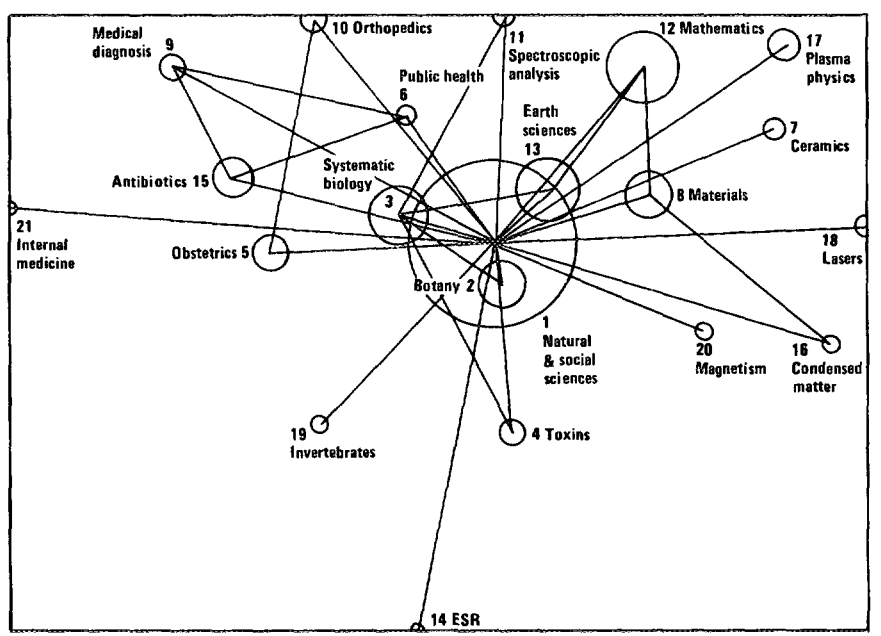


Figure 5: The 1984 C5 Global Map.



cluster "Sociological repercussions of technology and innovation in the UK, the US, and other countries" (#84-0516) at the C2-level.

Cluster #84-0516 becomes a member of the sociology cluster (#84-0076) at the C3-level (Figure 3). This cluster lies on the right-hand margin of the map and is clearly not as frequently co-cited as the clusters in the center of the map.

Figure 4 represents the C4-level, cluster #84-0001. Cluster #84-0076 appears in the lower right corner of the map, close to the realms of economics, psychiatry, population history, and anthropology.

Finally, at the C5-level, an overall structure of knowledge appears (Figure 5).

The accurate interpretation of multidimensional-scaling maps requires the reader to keep two pieces of visual data separate: the lines, on the one hand, and the circles drawn around each cluster, on the other.

First, the length of the line is inversely proportional to the relatedness of papers or clusters. Short links denote closely related subjects and longer lines denote research areas that are more distant intellectually from one another. Second, the size of the circles around each cluster indicates only the relative size of the citing literature for each. Overlapping circles are not to be interpreted as graphic representations of the percentage of the literature two clusters share in common, although if they are linked, such sharing does exist.

Multidimensional-scaling maps derived from co-citation clustering are, as extensions of citation indexing, natural organizations of the structure of knowledge in both the sciences and the social sciences. Furthermore, they can be useful tools in aiding researchers. Those unfamiliar with a subject can locate the area in question on a lower level map and obtain a list of highly cited and core documents for that subject. Even the expert may be led to a related field by the unsuspected proximity of his or her area to another revealed by the map. It is notable that the creation of a research tool, the citation index, led to new understandings of the structure of knowledge, which, in turn, produced new tools in the form of maps.

REFERENCES

1. **Finley J H.** *Four stages of Greek thought.* Stanford, CA: Stanford University Press, 1966.
2. **Garfield E.** *Citation indexing: its theory and application in science, technology, and humanities.* New York: Wiley, 1979.
3. ----------. Citation indexes for science. *Science* 122:108-11, 1955. (Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1984. Vol. 6. p. 468-71.)
4. **Garfield E, Malin M V & Small H.** A system for automatic classification of scientific literature. *J. Indian Inst. Sci.* 57:61-74, 1975. (Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1977. Vol. 2. p. 354-65.)
5. **Garfield E.** Citation indexing: a natural science literature retrieval system for the social sciences. *Amer. Behav. Sci.* 7(10):58-61, 1964. (Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1984. Vol. 6. p. 499-502.)
6. **Small H & Crane D.** Specialties and disciplines in science and social science: an examination of their structure using citation indexes. *Scientometrics* 1:445-61, 1979.
7. **Griffith B C & Small H.** *The structure of the social and behavioral sciences literature.* Stockholm, Sweden: Royal Institute of Technology Library, 1983.
8. **Kessler M M.** Bibliographic coupling between scientific papers. *Amer. Doc.* 14:10-25, 1963.
9. **Small H.** Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Amer. Soc. Inform. Sci.* 24:265-9, 1973. (Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1977. Vol. 2. p. 26-31.)
10. **Small H & Sweeney E.** Clustering the *Science Citation Index* using co-citations. I. A comparison of methods. *Scientometrics* 7:391-409, 1985.
11. **Small H, Sweeney E & Greenlee E.** Clustering the *Science Citation Index* using co-citations. II. Mapping science. *Scientometrics* 8:321-40, 1985.
12. **Kruskal J B.** Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1-27, 1964.
13. **Schiffman S S, Reynolds M L & Young F L.** *Introduction to multidimensional scaling.* New York: Academic Press, 1981.
14. **Spenser R.** Similarity mapping. *BYTE* 11(8):85-92, 1986.
15. **Small H & Garfield E.** The geography of science: disciplinary and national mappings. *J. Inform. Sci.* 11:147-59, 1985. (Reprinted in: **Garfield E.** *Essays of an information scientist: towards scientography.* Philadelphia: ISI Press, 1988. Vol. 9. p. 324-35.)

"Current Comments"

Clusters and Classification

Also see: A system for automatic classification of scientific literature a Indian Inst. Science 57:61-74, 1975. Reprinted on p.356, Essays, Vol:2

October 20, 1975 Number 42

Under whatever name, classification has always been the lodestone of scholarship and reputation in library science. Outside the world of books and documents it is also one of the most interesting and most problematic aspects of scientific inquiry.

At the Third International Conference on Classification held in January 1975 I presented the paper which is reprinted here.[1] This paper describes the use of cluster analysis in classification. Since I plan to deal more extensively with ISI® 's use of cluster analysis in the future, the reprint can be regarded as an introduction to the subject.

Automatic--or more precisely algorithmic--classification has been part of development of the *Science Citation Index®* (*SCI®*) from the beginning. I tried, at the First International Conference on Classification in 1957, to persuade the 'classification establishment' that classification could be automatic. Use of the term *algorithmic* that long ago would only have made my effort more difficult.

Perhaps the main point to be made here is that these bibliographic clusters--these 'self-generating' categories if you will--have been algorithmically identified by the simplest clustering techniques. And they conform remarkably well to what scientists themselves regard as areas 'where the action is.' One can examine data from past years and verify that the data confirm and that the clusters describe where the action *was*. One can examine data over a period of time, and, with some simple extrapolations, discover that it's possible to talk sensibly about where the action looks like it very likely will be.

Probably the best confirmation of this is that scientists often tell us that citation-based cluster analysis gives them a better overview of their own fields than they themselves may have had.

As I have mentioned above, there is to me still surprising resistance in the 'classification establishment' to the concept of automatic or algorithmic classification. Perhaps it should not surprise me considering the intellectual resis-

tance one still encounters also to the concept of automatic or algorithmic *indexing*. This latter, however, fairly floors me whenever I encounter it, especially when I encounter it in the learned journals of the field. A recent article stated: "... there is little hard evidence as to the value of citations in an automated system, particularly as substitutes for other modes of indexing, as opposed to additional keys."[2] With fifteen years' compilation of the *Science Citation Index* on the shelves of large and small academic, industrial, and government libraries around the world, it is difficult to imagine what any rational basis for such a statement can possibly be. I felt constrained to reply, in a letter to the editor of the journal in which the article appeared, that the author "and others persist in ignoring the reality of the *SCI* as the largest extant *automatically,* that is *algorithmically,* indexed collection available... [It is] used every day by thousands of clients who do not require philosophical analysis to measure value received. What theorists should be rigorously seeking is why it does work and what its fundamental implications are for linguistic and other studies."[3]

If the concept of automatic *indexing* and the very existence of automatically--that is, algorithmically--generated indexes can be ignored at this stage of the game, I suppose I must accept the fact that it would indeed be unduly sanguine of me to expect immediate and enthusiastic research into the validity of algorithmic classification.

But if the clustering method of category generation presented here accurately identifies the fields of research that exist in the real world, then surely the indexing terms--the citations--which form the basis of algorithmic classification must reasonably well describe documents. If they did not, then why--despite any doubts about their effectiveness--do they produce such an amazing correspondence to reality?

1. **Garfield E, Malin V M & Small H.** A system for automatic classification of scientific literature. *J. Indian Inst. Sci.* 57(2):61-74. Reprinted in *Current Contents®* No. 42, 20 October 1975, p. 7-16.
2. **Sparck-Jones K.** Progress in documentation: automatic indexing. *J. Documentation* 30(4):393-432, 1974.
3. **Garfield E.** What is automatic indexing? *J. Documentation,* in press.