Essays of an Information Scientist, Vol:10, p.380, 1987 Current Contents, #51-52, p.3, December 21-28, 1987

Current Comments®

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

The R&D Mission at ISI: Basic and Applied Research, for Us and for You

Numbers 51-52 December 21-28, 1987

Virtually every company of any significant size—whatever its business—supports a program of research and development. R&D departments help management to improve existing products and often suggest new ones. Until relatively recently, labor or capital resources were the dominant inputs in manufacturing. Today, knowledge, especially in the form of technology, is increasingly becoming the key element in a company's success. More and more, know-how yields that all important competitive edge. And what holds true for a company also holds true for a nation's economy: this year's Nobel laureate for economics, MIT's Robert M. Solow, demonstrated in the late 1950s that technology is the engine that drives an economy's growth.

For companies like ISI® that provide information-based products, knowledge is doubly important. Companies in the information industry are constantly seeking new knowledge about knowledge—how it can be manipulated, divided, or joined together for a variety of purposes and end-users. It is not an exaggeration to say that ISI's expenditure on R&D is one of its most important investments in the company's future, in its ability to serve customers and thereby ensure its own prosperity and longevity.

ISI's Corporate Research Department, under the leadership of Dr. Henry Small, is engaged in a program of basic, strategic, and applied research both on behalf of in-house personnel and also, on a contract basis, on behalf of government agencies and institutions, industries worldwide, as well as members of the general scholarly and scientific community. The department develops innovative analytical methods and procedures that, when used in conjunction with ISI's databases, further information or knowledge retrieval and contribute materially to greater understanding of the process and development of science.

In basic research, the department investigates the structure of the knowledge embodied in the academic literature. In strategic research, it nourishes discoveries from the basic level that offer potential for new product development in the medium term (two to four years in the future). In applied research, the department is involved in improving current products and services and in conducting specialized research for clients.

That last function—research sponsored by extramural agencies and institutions—has not been discussed much in *Current Contents®*. Rather than listing the almost innumerable types of research that the department can undertake, I have asked Henry to describe the general capabilities and recent activities of his team. In this way, readers can imagine how our resources might be applied to their specialized information needs.

As Henry mentions at the end of his summary, one of our long-term goals is the development of a knowledge database that draws on the most advanced techniques of artificial intelligence (AI). Through the use of AI, we will aim for the ultimate combination of objective and subjective, or expert, analysis. The availability of full texts in machine-readable form will enable this process to become fully algorithmic, the possibility of which Henry touched upon in a recent paper.[1] He was too modest to mention that

article, which recently brought him the American Society for Information Science's annual award for the best paper published in its journal. It was also recently announced that Henry and Professor V.V. Nalimov of the USSR will be the next recipients of the Derek de Solla Price Award, sponsored by *Scientometrics* for outstanding contributions to the quantitative study of science.

REFERENCES

1. **Small H.** The synthesis of specialty narratives from co-citation clusters. *J. Amer. Soc. Inform. Sci.* 37:97-110, 1986.

Report on Citation Analysis Research at ISI

Henry Small

Few readers will be surprised to learn that ISI® is actively engaged in citation analysis research. ISI's major databases—the *Science Citation Index®* (*SCI®*), *Social Sciences Citation Index®* (*SSCI®*), and *Arts & Humanities Citation Index™*, which now extend back significant periods of time—are the most comprehensive resources available for citation-based studies. When I first came to ISI 15 years ago, I thought of working with title-word data, but when I saw the unique power of citation data, I quickly changed my plans.

First, what is citation analysis? On the one hand, it is a prominent subarea of bibliometrics, long practiced by librarians, bibliographers, and others. Alan Pritchard, then at Northwestern Polytechnic, London, UK, defined bibliometrics as "the application of mathematics and statistical methods to books and other media of communication."[1] On the other hand, citation analysis is also a subarea within scientometrics, a field of research (as well as the name of a journal) concerned with the quantitative study of science. This is not the full story, however, because citation analysis is a derivative of citation indexing, which was originally and primarily concerned with methods of information retrieval. Hence, it has strong ties to information science.[2]

I will begin this overview of citation analysis research at ISI with this last-named application, information retrieval. One might wonder: How can a quantitative study of reference patterns in scientific papers aid in retrieval? A good example is bibliographic coupling, a statistically defined measure of association based on the number of references cited in common by different papers. Early work by MIT's Michael M. Kessler, using the physics literature, showed how this measure could be effective.[3] At ISI we have returned to some of those earlier ideas[4] and used CD-ROM technology to design a powerful "see also," or cross-referencing, tool—as you might find in a thesaurus. Given almost any paper having a reference list, we can quickly show the user the most closely related papers based on bibliographic coupling strength.

To illustrate the subtlety and power of this method, Figure 1 shows a paper on "instantaneous plumes in the atmosphere." The ISI data point us to a paper on "insect sex pheromones in mating behavior," which is coupled to the first paper by two shared references. At first this appears to be a spurious linkage, but on closer examination (and even an inquiry to the author) we found a very well-founded scientific basis for this connection, namely, the structure of the air currents that carry pheromones. It is very unlikely that these two articles would have been indexed together by conventional methods.

In the near future, with the aid of new parallel processors such as the Connection Machine,[5] we expect to be able to extend the concept of bibliographic coupling to other attributes of the document in addition to its references. The idea is to match any or all aspects of the bibliographic description of two papers, including title words, coauthors, addresses, and so on. This becomes practical only when the full records of thousands

Figure 1: Bibliographic coupling: articles related to atmospheric plumes ranked by coupling strength. Cross-referencing is independent of title words.

JONES CD
On the structure of instantaneous plumes in the atmosphere
JOURNAL OF HAZARDOUS MATERIALS (7):87, 1983 21 REFS

STOREBO PB, BJORVATTEN T, HONNASHAGEN K, LILLEGRAVEN A, JONES CD, VANBUIJTEN CJP
Tracer experiments with turbulently dispersed air ions
BOUNDARY-LAYER METEOROLOGY (26):127, 1983 7 REFS
Shared References—3

HIROOKA Y
Role of insect sex-pheromones in mating-behavior. 4. Repeated turnings of male fall webworm, Hyphantria-cunea Drury (Lepidoptera, Arctiidae)
APPLIED ENTOMOLOGY & ZOOLOGY (18):139, 1983 10 REFS
Shared References—2

KENNEDY JS
Zigzagging and casting as a programmed response to wind-borne odour - a review
PHYSIOLOGICAL ENTOMOLOGY (8):109, 1983 53 REFS
Shared References—2

of papers can be matched simultaneously—in parallel—on all data fields. The concept of "data parallelism" of the Connection Machine seems ideally suited to this information retrieval application.[6]

The second area of citation analysis research at ISI is what might be termed "science indicators," which describes any quantitative analysis of bibliographic data that bears on the state of science. Some readers will be familiar with the series of reports from the National Science Foundation (NSF) of the same name,[7] and indeed citation data derived from ISI's files have been among the many indicators used to show the position of US science relative to that of other countries.

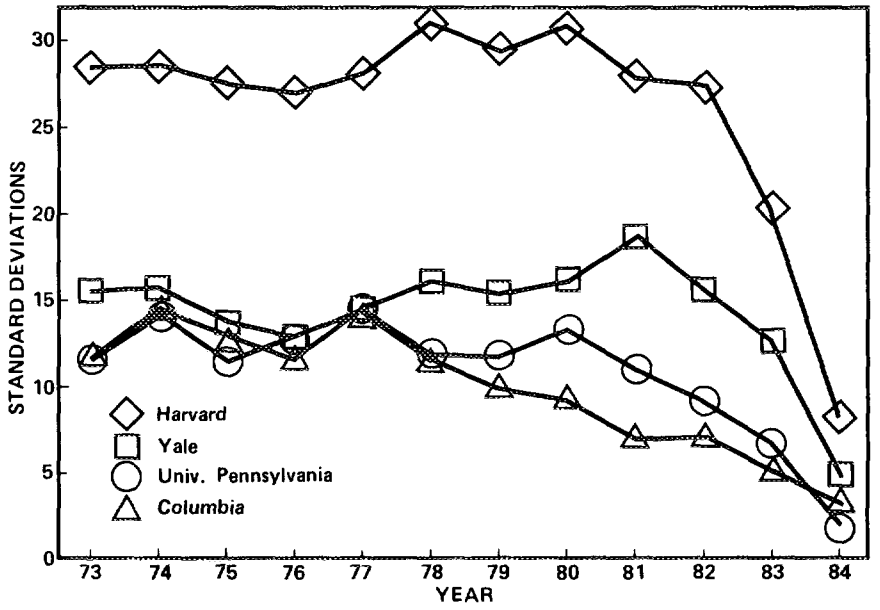
The NSF effort has focused on one of many ways the ISI data could be used to indicate the state of science. Our approach is to expand the kinds of units of analysis studied, for example, from countries to research institutions and from disciplines to scientific specialties. These analyses can also exploit the growing longitudinal dimension of ISI files (the *SCI* will soon extend back to 1945) for time-series studies.

At ISI we have been preparing our data archives for long time-series analyses of science indicators. This involves, in part, matching our extensive multiyear files of citation counts for specific papers with the detailed source description of each article, including all coauthors and institutional addresses. A prototype file we have built covers the years 1973 to 1984. Figure 2 summarizes data for four Ivy League institutions over this period. The plot shows the number of standard deviations, in citation terms, that an institution is from the overall behavior of the file represented by the horizontal line at zero. As one approaches the present, papers have less and less time to be cited, so the magnitude of the differences decreases. In effect, all papers are created equal (at the time of their publication). As the papers age, however, the curves diverge and become increasingly stable. This shows clearly how access to a long time series of citation data is essential to the evaluation of scientific productivity.

In the near future ISI will be offering a full range of international science indicator services based on such multiyear files. These services will enable users to make assessments of countries, institutions, laboratories, disciplines, and specialties, both comparatively and longitudinally.

A second kind of evaluative study is represented by ISI's recent study of the TOXLINE database under contract from the National Library of Medicine (NLM).[8] TOXLINE is an online bibliographic database maintained by NLM that covers the litera-

Figure 2: Time-series analysis of the citedness of papers from Harvard University, Yale University, University of Pennsylvania, and Columbia University expressed as the number of standard deviations from the norm for the *SCI*® as a whole.



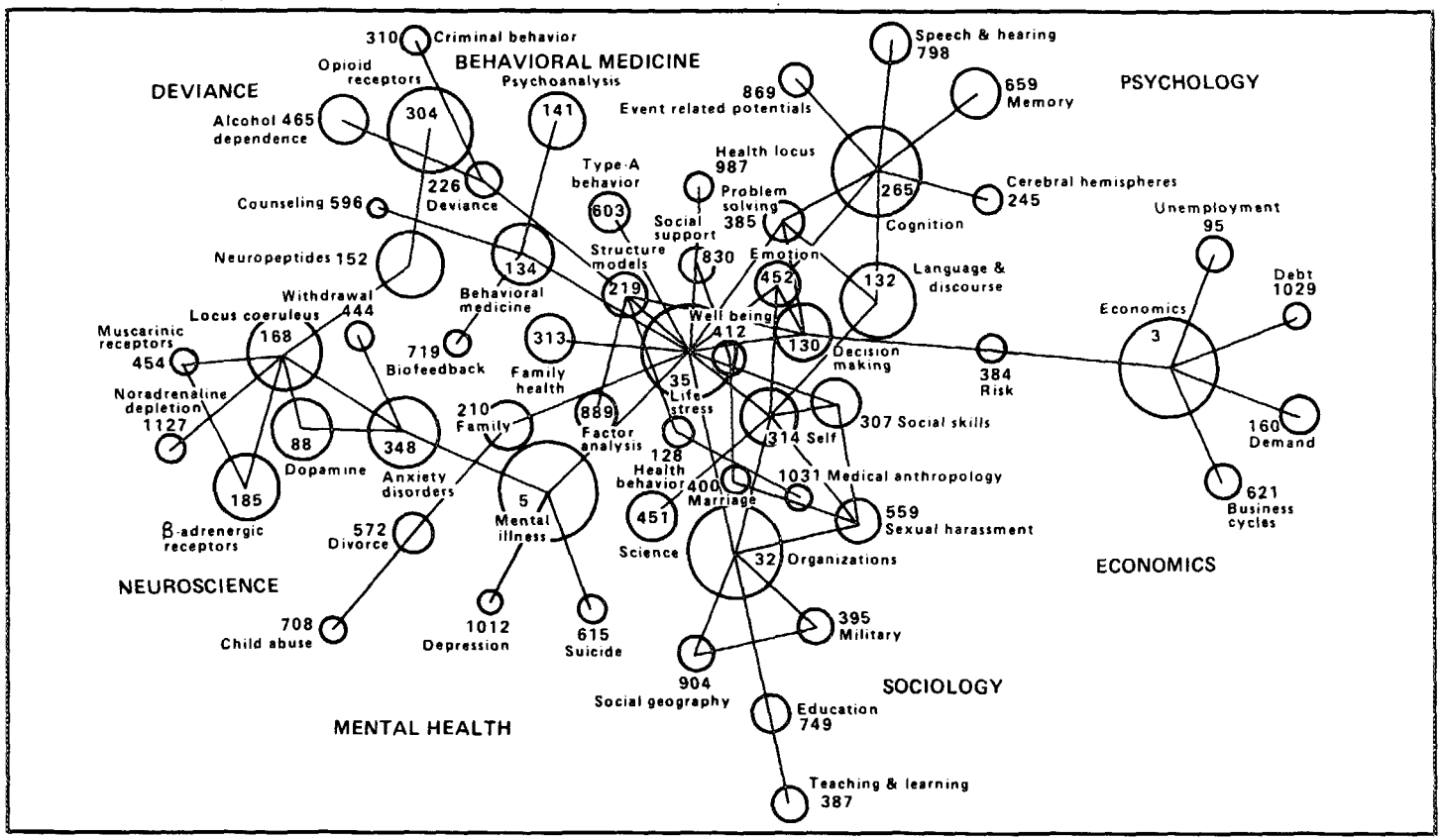
ture of toxicology. Since ISI's database is uniquely multidisciplinary, we were able to examine how TOXLINE's journal coverage compared with ISI's *Journal Citation Reports*® data and reveal some possible weak areas of coverage in TOXLINE. Assessing the journal coverage of specialized or disciplinary databases is only one way ISI's citation database can be of utility to other database producers, both large and small.

The last research topic I will discuss, but by no means the least in terms of our interest, is what might be called the taxonomy of scientific literature. Most classification systems we are familiar with in the bibliographic world, such as the Dewey decimal system, MeSH tree structure, and the UDC, are subjectively (manually) constructed. At ISI the philosophy of a "natural language" of science has been the dominant approach, and this especially applies to the references scientists cite in the construction of their papers. These references constitute one dialect of the language of science. Numerical taxonomists have long sought to derive classes of living things based on shared characteristics or common ancestry.

The analogous pursuit in the bibliographic world is the cluster analysis of papers based on citation patterns. Two ways this can proceed are by the sharing of references cited by papers (the previously mentioned bibliographic coupling concept), or, conversely, by the "co-citation" of earlier papers in the reference lists of later papers.[9] The latter measure is particularly useful when focusing on highly cited papers that often represent key concepts and discoveries. In either case, what we achieve with a cluster analysis is an organization of many units of information into a smaller number of manageable units.

Any citation file can be subjected to cluster analysis using either measure. For example, we have built multiyear files of ISI's data for individual countries. Using co-citation and coupling methods we can show areas of research strength and weakness for each country. Another example is a cluster analysis of our combined annual 1986 *SCI*

Figure 3: Social and Behavioral Sciences, 1986. Co-citation clustering expresses relatedness between specialty areas.



and *SSCI* files. The latter effort is the latest in a long series of cluster analyses of our annual files that extend back to 1970.[10]

To illustrate this latest annual cluster analysis, we have included a map of the social-sciences portion of that file. This map was created at a level of aggregation designed to show the relations between individual research areas (Figure 3). One of the notable features of this map is the visible link between mental-health research and neuroscience. For three years now we have observed the growth of neuroscience and the strengthening of this link between the behavioral and biomedical sciences.

The creation of such maps is a further processing step that uses clusters as input and performs a multidimensional-scaling analysis to obtain a spatial configuration, where each cluster is assigned to its best-fitting point in space.[11] Scaling is a way of summarizing multiple relationships among objects by representing them spatially. To display clusters we use a graphics software system developed by ISI for the IBM PC. Using this system we can browse among four levels of cluster maps starting at a global level. We can then move down successive levels to the highly cited papers that compose the clusters. We believe this cluster graphics system is one of the first purely graphic interfaces for bibliographic searching.

Some clustering algorithms are known to generate loosely connected aggregates of entities that are good for revealing connections between areas (low precision, high recall), while other algorithms create very compact and exclusive categories (high precision but low recall). We are now working to create systems that are not only flexible in terms of the data that are clustered and the criteria that establish their similarity but also that are flexible in terms of the algorithm applied. This means having the ability to change the algorithm depending on the nature of the data and intended application of the results.

Creating a natural classification system for science based on a cluster analysis of empirical data, such as citation patterns, can be applied to the control and retrieval of the literature. It can also aid in more advanced applications such as the selection of review topics, as is currently done in ISI's *Atlas of Science*®,[12] and the development of knowledge bases using artificial intelligence techniques. Unquestionably, the present opportunities for research that combine ISI's databases with new information technologies are more promising than ever before.

REFERENCES

1. **Pritchard A.** Statistical bibliography or bibliometrics? *J. Doc.* 25:348-59, 1969.
2. **Garfield E.** *Citation indexing: its theory and application in science, technology, and humanities.* Philadelphia: ISI Press, 1983. 274 p.
3. **Kessler M M.** Bibliographic coupling between scientific papers. *Amer. Doc.* 14:10-25, 1963.
4. **Vladutz G & Cook J.** Bibliographic coupling and subject relatedness. (Flood B, Witiak J & Hogan T H, comps.) *1984: challenges to an information society. Proceedings of the 47th ASIS Annual Meeting Vol. 21*, 21-25 October 1984, Philadelphia, PA. White Plains, NY: Knowledge Industry, 1984. p. 204-7.
5. **Hillis W D.** *The Connection Machine.* Cambridge, MA: MIT Press, 1985. 190 p.
6. **Waltz D L.** Application of the Connection Machine. *Computer* 20(1):85-97, 1987.
7. **National Science Board.** *Science indicators.* Washington, DC: NSB, 1972-.
8. **Small H.** *TOXLINE evaluation.* October 1987. 46 p. (Unpublished report.)
9. ----------. Co-citation in the scientific literature: a new measure of the relationship between two documents. *J. Amer. Soc. Inform. Sci.* 24:265-9, 1973.
10. **Small H & Garfield E.** The geography of science: disciplinary and national mappings. *J. Inform. Sci.* 11:147-59, 1985.
11. **Kruskal J B.** Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1-27, 1964.
12. **Garfield E.** Launching the *ISI Atlas of Science*: for the new year, a new generation of reviews. *Current Contents* (1):3-8, 5 January 1987.