"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.