Essays of an Information Scientist, Vol:2, p.543-549, 1974-76 Current Contents, #33, p.7-13, August 16, 1976
Reprinted from *The Information Scientist* 10(1):3-10, March 1976

See Introduction on p.541, Essays, Vol:2

Understanding Science by analysing its Literature

A. E. CAWKELL

Institute for Scientific Information®
18 North Common Road, Middlesex UPB8 1PD

A brief description of work using citation networks for the better understanding of the history and structure of Science. The use of a computer is necessary for large scale automatic clustering procedures using the phenomena of bibliographic coupling and co-citation coupling. Results may be used as indicators of changing relationships between specialties, and possible coalescences and for the early detection of possible 'research fronts'.

'The institutional conception of science as part of the public domain is linked with the imperative for communication of findings. Secrecy is the antithesis of the norm, full and open communication is its enactment.'
Robert Merton

'If I have seen farther it is by standing on the shoulders of giants.'
Isaac Newton

Introduction

The above quotations imply that published papers are the end product of much scientific research, and that Science is like an edifice, building upon the past. The socio-scientific aspects of this activity have been discussed by Kuhn[1] and Merton.[2] Since published papers follow a traditional pattern in the main and the structure of the edifice is indicated by references to earlier 'building blocks' the potential for examining Science through its literature obviously exists.

It is fashionable to decry citations and various examples have been put forward purporting to show that they cannot be reliable indicators for a variety of reasons—for example because of excessive self-citations, plagiarism of references, careless or omitted references etc. Some thorough investigations have been carried out, notably by the Cole brothers, and these criticisms have been satisfactorily answered.[3] It turns out that if citation counts to the work of scientific communities are examined, a good correlation exists between numbers of citations received and a range of conventional indicators such as honorific rewards received, occupational position, books and papers published, and so forth.[4] Citation anomalies have little effect—they are like random noise in the presence of strong repetitive signals. However, the use of citations in assessing the quality of the work of an individual scientist requires considerable care. Anomalies having a small effect on the average might have a serious effect in a particular case. It is unwise to draw conclusions from citation counts for individuals unless a detailed examination is made by a person very well versed in citation practices and adequate supportive evidence is available.

It is my purpose here to describe work which has been done using citation networks as aids to a better understanding of the history and structure of

Science. An early suggestion that citations might be used as quality indicators was made by Garfield,[5] and later de Solla Price[6] suggested that citation analysis could be used to investigate research fronts. A detailed examination was carried out by Garfield *et al.*[7] of the history of the Genetic Code in terms of 'nodal articles' and citation inter-connections, and the coalescence of protein chemistry, genetics and nucleic acid chemistry was tracked. A further exercise was carried out by Cawkell[8] who illustrated the controversy surrounding S. R. Ovshinsky's work on Amorphous Semi-conductors with a citation network. A portion of this network is displayed in Fig. 1. Articles are shown as circles, ordered chronologically from top to bottom; an arrow between articles indicates a citation.

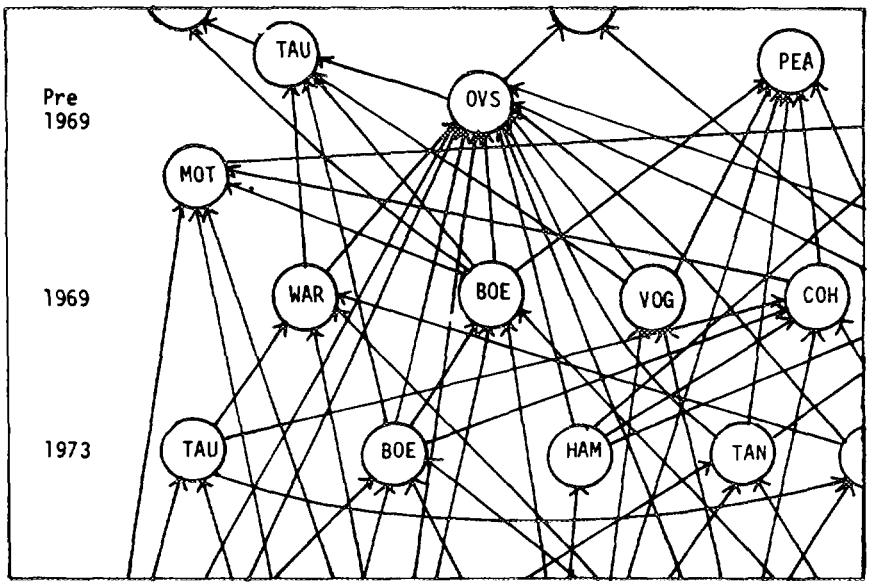

Fig. 1. Part of amorphous semiconductor citation network

Article coupling

Work is in progress for the interactive plotting of individual diagrams by computer, and to look at whole areas of science it is essential to use automatic methods. It will be helpful to understand two forms of article coupling before describing these methods. In Fig. 2 articles A and B are said to be 'bibliographically coupled' by their common references to earlier articles. An extreme case of bibliographic coupling occurred recently[9] when six references in each of two articles, each of which contained a total of eight references, were to the same earlier articles. These two articles covered the same subject with a remarkable degree of similarity. Another form of subject-similarity indicator is provided by 'co-citation coupling'.[10] In Fig. 3 articles

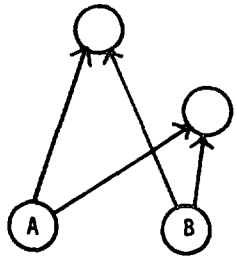

Fig. 2. Bibliographic coupling

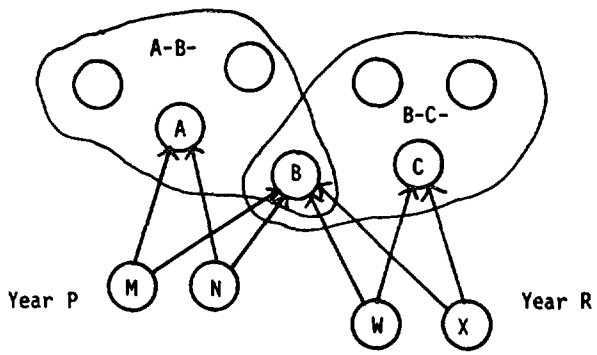

Fig. 3. Co-citation coupling

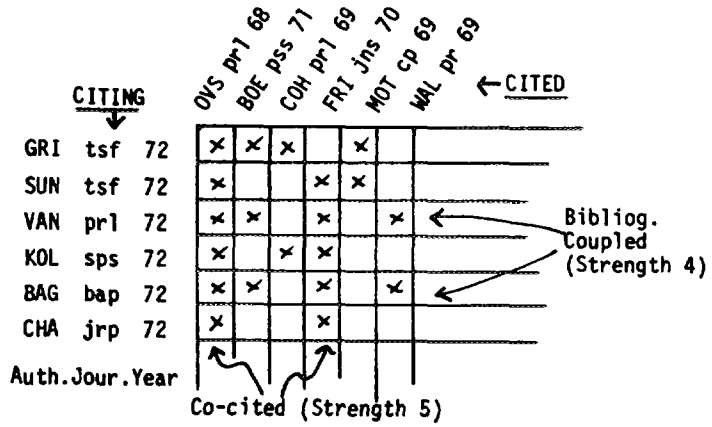

Fig. 4. Citation matrix

A and B are perceived as being about the same subject by M and N which co-cite them in year P. Later, an allied subject emerges and W and X perceive that B and C are related. The subject areas containing A, B and B, C, presumably are related or would be if the number of co-citations was larger. Bibliographic coupling fixes the relationship between the citing papers A, B, for all time, but dynamic relationships are exposed by co-citation activity as authors perceive changing subject relationships. These functions (Fig. 4) can be shown as a matrix (taken from reference 8). Bibliographic coupling between two 1972 articles is shown, whilst several 1972 authors perceive the relationship between a 1968 and a 1970 article (the other symbols denote author and journal).

A method has been evolved by Small and Griffith[11,12] to use co-citations on a grand scale as part of an automatic clustering procedure. This method pre-supposes two effects—firstly that heavily cited articles are of greater interest than less cited articles, and secondly that heavily cited articles which are co-cited by many pairs, triples, quads, etc. of later articles are significant and subject related.

Clustering

ISI's annual machine-readable file covering all major Science and Technology—say the 1973 file—lists about 400,000 1973 articles together with the 4·5 million earlier articles which they cite. Some of these articles date back to antiquity (Aristotle and Herodotus received a number of citations from 1973 articles), but a high proportion are concentrated within the last 20 years. In the clustering procedure, firstly only those articles cited p or more times are listed. This new file is substantially smaller than 4·5 million articles.

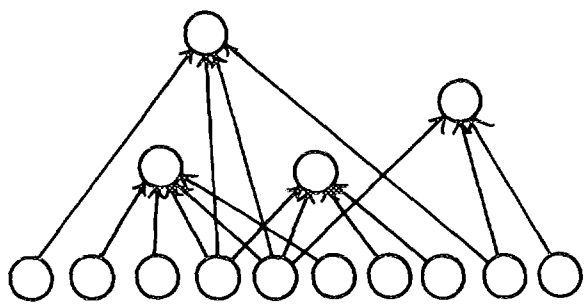

Fig. 5. Cited and co-citing papers

Next, these highly cited papers are clustered by selecting one paper and listing all papers which are co-cited with it. Other papers which are co-cited with these papers are listed in turn until all cited papers linked by co-citations have been identified. The co-cited papers in a cluster depend upon a selected threshold level q, the minimum number of times any paper is co-cited.

In Fig. 5 $p = 3$, $q = 2$. For the experimental work various values are arbitrarily chosen for p and q. As the values are increased clusters separate from the mass, and at high values all but a few, characterized by intense publication activity, disappear. The process is analogous to flood water rising on hilly country; islands, and finally peaks are isolated. At a particular level of p and q, 1600 clusters appeared and these have been individually examined. Subject relationships between papers may be determined by examining each paper in a cluster, or by examining the corpus of co-citing papers, or both. One convenient method is to analyse words in the titles of the co-citing papers. The subject matter, as expressed by significant high frequency words—that is as perceived from a consensus of co-citing authors—may not necessarily be the same as it appears to be from examining the co-cited papers in the cluster. In the event, clusters appear to be definable scientific specialties. Some work has been done to find out whether practising experts agree with ISI's description of a specialty, and whether the articles in the clusters do represent major contributions. The results look very promising. A cluster identified as 'locus of control' is mapped in Fig. 6. Articles are shown as rectangles containing Author, Journal and Year of publication, with the

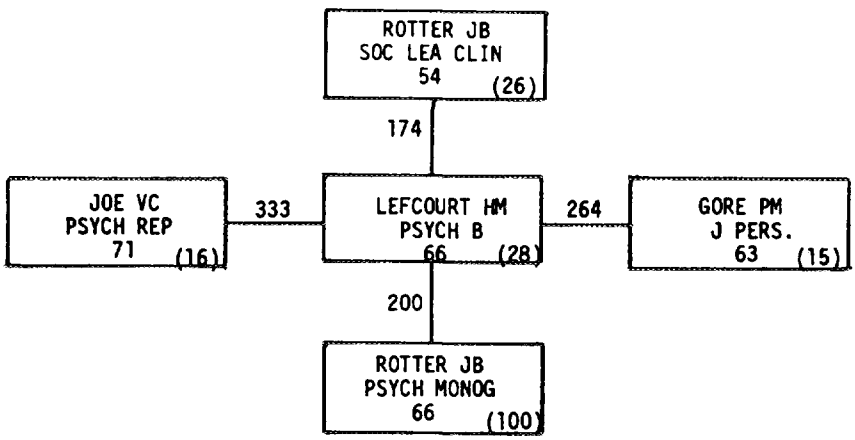

Fig. 6. Specality 'locus of control'

number of times cited in brackets. Figures on connecting lines are a measure of the connectedness of articles, derived from numbers of co-citations received. This particular cluster is not highly interconnected. 100 per cent connectedness would be represented by lines connecting each article with every other article. However a list of some of the co-citing articles, given in Fig. 7 shows that an interesting collection of papers in different fields and journals are drawn together by co-citation characteristics, although the background fields in which 'control' is being examined may be very different.

Description	Number of core articles cited
BLANCHARD LB J SOC PSYCH 89 123 73 "Locus of control and prediction of voting behaviour in college students"	(2)
JOFFE JM SCIENCE 180 1383 73 "Control of their environment reduces emotionality in rates"	(3)
EDWARDS AL ANN R PSYCH 24 241 73 "Measurement of personality traits; theory and technique"	(2)
SADAVA SW CAN J BEH S 5 371 73 "Initiation to cannabis use-longitudinal social psychological study of college freshmen"	(2)
PHARES EJ PSYCHOL REP 32 923 73 "Source and type of wives. Problems as related to responsibility attribution, interpersonal attraction, and understanding"	(2)

Fig. 7. Examples of 'locus of control' co-citing papers

A similar procedure has been used for clustering clusters—that is to show how clusters are perceived as being related to each other. This provides insights to the structure of Science. Space precludes a detailed picture of one of these maps, but part of one (from reference 10) is shown in Fig. 8. The

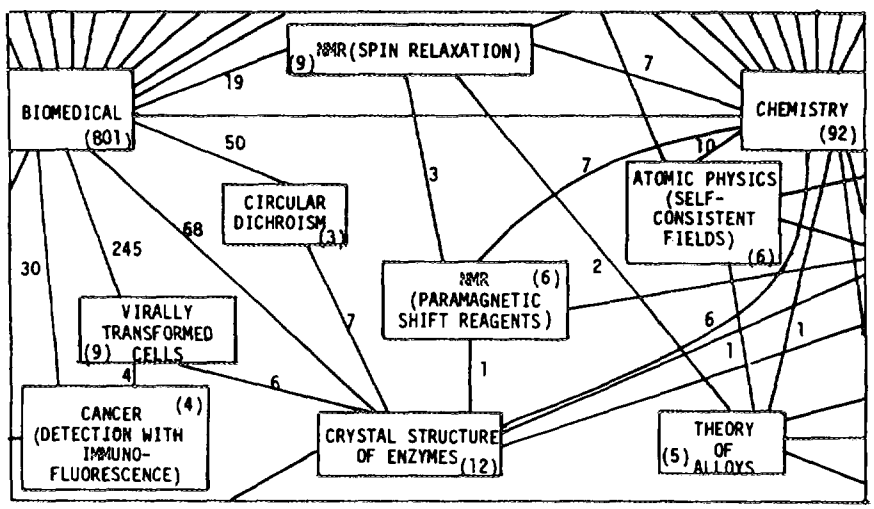

Fig. 8. Part of scientific specialty map

figures within rectangles denote the number of articles in the cluster, and those on connecting lines the relative connectedness of cluster. This map was plotted from a small sample of the 1972 SCI file so weak links may not be statistically significant. Annual analyses produce similar but more meaningful maps.

Application

Work is proceeding towards the interpretation of changes between successive annual maps as possible indicators of changing relationships between specialties and possible coalescences. For example there was a substantial difference between the 1972 and 1973 maps in the biomedical area when a number of clearly distinguishable separate specialties merged into one large specialty which has been entitled 'Viral genetics—reverse transcription and chromosomes'.

Another possibility is the detection of new specialties or possible 'research fronts' at any early stage. These fronts might be detected by virtue of their special characteristics (as has been suggested by Meadows[13]) when analysing citation data.

References

1. **Merton R K.** *The sociology of science.* University of Chicago Press, 1973.
2. **Kuhn T S.** *The structure of scientific revolutions.* 2nd ed. University of Chicago Press, 1970.
3. See letters from **A. Goudsmit *et al.*** and response from **J.R. Cole** and **S. Cole** under the title 'Citation Analysis'. *Science* **183**(4120):28-33, 1974.
4. **Cose J R & Cole S.** *Social stratification in science.* University of Chicago Press, 1973.
5. **Garfield E.** Citation indexes in sociological and historical research. *American Documentation* **14**(4):289-91, 1963.
6. **Price D J D.** Networks of scientific papers. *Science* **149**(3683):510-15, 1965.
7. **Garfield E, Sher I H & Torpie R J.** *The use of citation data for writing the history of science.* Philadelphia: Institute for Scientific Information® , 1964, 86 pp.
8. **Cawkell A E.** Search strategy, construction and use of citation networks with a socio-scientific example; amorphous semiconductors and S.R. Ovshinsky. *J. Amer. Soc. Inform. Sci.* **25**(2):123-30, 1974.
9. **Neville A C & Smith D S.** Airborne organism identified. *Nature* **225**(5228): 199, 1970.
10. **Small H.** Co-citation in the scientific literature; a new measure of the relationships between two documents. *J. Amer. Soc. Inform. Sci.* **24**(4): 265-69, 1973.
11. **Small H & Griffith B C.** The structure of scientific literatures. 1. Identifying and graphing specialties. *Science Studies* **4**:17-40, 1974.
12. **Griffith B C, Small H, Stonehill J A & Dey S.** The structure of scientific literatures. 2. Toward a macro- and micro-structure for science. *Science Studies* **4**:339-65, 1974.
13. **Meadows A J & O'Connor J G.** Bibliographic statistics as a guide to growth points in science. *Science Studies* **1**(1):95-99, 1971.