Fractal dimension of co-citations

SIR—Like many natural phenomena, the growth of scientific knowledge seems to be cluster-like. On a spatial scale, scientific discoveries mainly 'cluster' around important research institutes. On a temporal scale, scientific discoveries often occur in relatively short periods of time as an important breakthrough makes new advances possible. Here I focus on a non-physical abstract structure in which pieces of scientific information are clustered according to specific aggregation rules. I discuss geometrical properties of co-citation clustering and in particular, the size distribution of these clusters in terms of fractal dimensions.

When a scientific paper cites two earlier papers, these latter papers are 'co-cited'. The strength of such a co-citation relation is determined by the number of citing papers having the above pair in their reference list. One of these co-cited papers can also form a co-citation pair with a third paper. In this way, clusters of (co-)cited papers emerge, and a 'map' of the citation field can be created[1]. Looking at papers published in 1984, about six million cited papers are reduced to some 70,000 highly cited papers, and with these papers nearly 10,000 clusters are formed. These C1 clusters are used as input for a second-step clustering resulting in about 1,400 C2 clusters, each of them containing 2 to 60 clusters of the first (C1) cluster generation. The iteration procedure is then performed twice again, the 1,400 C2 clusters being input for about 180 C3 clusters, and these latter clusters being input for the final C4 clustering, yielding 21 C4 clusters. Each iteration reduces the number of clusters by an order of magnitude. In general, one may say that at the C1 level science is structured in terms of (small) research specialities, whereas at higher levels co-citation clusters become increasingly extensive in size, representing higher hierarchical structures such as subfields, fields and disciplines. 'Cluster size' refers to the number of citing papers involved.

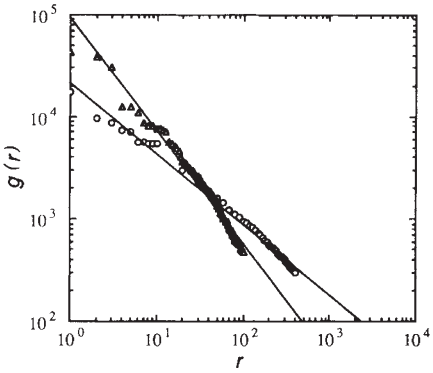
In the figure, I present data on the size–rank distribution of C2 and C3 co-citation clusters. For the largest part of the ranking scale, over about two orders of magnitude, the distribution is a power law $g(r) = k\, r^{-\gamma}$ where $g(r)$ is the size (number of citing papers) of the cluster with rank r, k is a value determined empirically, and γ is the slope of the line. I found that $\gamma(C3) = 1.09$ and $\gamma(C2) = 0.71$, indicating a clear difference between the cluster systems. The next step is the conversion of the size–rank distribution function into the 'usual' size distribution function: the number of clusters with size g, $n(g)$, as a function of g

$$n(g) = -\mathrm{d}r/\mathrm{d}g \sim g^{-(1/\gamma+1)} \qquad (1)$$

Using fractal theory[2,3], the size distribution of fractal 'islands' (fragments of area A from fractal fragmentation) is

$$n(A) \sim A^{-(D/D_c+1)} \qquad (2)$$

where D is the fractal dimension describing the degree of fragmentation of the set of islands, and D_c is the fractal dimension of the 'coastline' of individual islands or fragments.



Size-rank distribution of C2 and C3 Co-citation clusters. Circles, distribution for the power-law part of the (total) 1,371 C2 clusters[5]; triangles, distribution for the power-law part of the (total) 179 C3 clusters (C3 data was collected using the online version of the Science Citation Index (SCISEARCH, at the host Deutsches Institut für Medizinische Dokumentation und Information, Cologne)).

The 'island area' can be compared with size or 'volume' of the co-citation clusters, that is $g = A$. The citing papers constitute the cluster size and in geometrical terms may be considered as unit surface or unit volume elements. Comparing equations (1) and (2) gives $1/\gamma = D/D_c$. I am primarily interested in the fractal fragmentation dimension D of the co-citation cluster distribution. Assuming, to a first approximation, that the form of the co-citation clusters is regular (the clusters have a smooth 'coastline'), I put $D_c = 1$, and therefore $\gamma = 1/D$. This agrees with the Mandelbrot[2] generalization of the Zipf frequency distribution recently used to describe the size distribution of species in ecosystems[4]. For the C2 clusters, $\gamma(C2) = 0.71$, so that $D(C2) = 1.41$, and for the C3 clusters, $\gamma(C3) = 1.09$, which gives $D(C3) = 0.92$. The fractal dimension of C2 clusters is significantly higher than for C3 clusters. These results agree qualitatively with results for 'fractal landscapes' (which mimic those generated from brownian motion[2]), in particular the relation between degree of fragmentation and the fractal dimensions of island size distributions[2]. A striking characteristic of these fractal landscapes is that the higher the fractal dimension $1 \leqslant D \leqslant 2$, the more fragmentation of larger islands occurs.

What is the meaning of a fractal dimension of information space as represented by co-citation clustering? The size distribution of the co-citation clusters is a snapshot of a dynamical process, reflecting the presence of established fields and the emergence of new specialities. Like fractal distributions in ecological systems[4] one may consider co-citation clusters as *a* representation of the ecosystem of scientists.

In this model, the structure of the ecosystem is strongly related to some optimal distribution of energy, mass and information. If co-citation clusters represent 'species of scientists', then the fractal cluster distribution gives a measure of the diversity of the research community, that is, the distribution of individuals among species owing to the optimization of flows of 'mass' (scientists, budgets) and information. For the small C2 and C3 clusters there is a deviation from the power-law behaviour and for such small clusters, the parameters that determine their structure and relations (such as flow of people, money and information) are much more subject to a random process, whereas for the larger clusters, the underlying dynamics follows particular patterns yielding fractal distributions. Fractal geometry is known to be closely related to the problem of describing the propagation of order in non-equilibrium systems. Therefore, the fractal model of co-citation clustering is an interesting starting point for further modelling of scientific ecosystems.

Very recently we determined the size distribution of the nearly 10,000 C1 clusters, the 'finest' fragmentation of science (in terms of co-citations). A fractal dimension of $D = 2.0$ is found, which is in line with the findings for the C2 and C3 clusters.

A. F. J. VAN RAAN
Centre for Science and Technology Studies,
University of Leiden,
Wassenaarseweg 52,
PO Box 9555,
2300 RB Leiden,
The Netherlands

1. Small, H. & Garfield, E. *J. Inf. Sci.* **11**, 147–159 (1985).
2. Mandelbrot, B.B. *The Fractal Geometry of Nature* (Freeman, San Francisco, 1982).
3. Voss, R. F., Laibowitz, R. B. & Allessandrini, E. I. in *Scaling Phenomena in Disordered Systems* (eds Pynn, R. & Skjeltorp, A.) NATO ASI Ser., B**133**, 279–288 (Plenum, London, 1985).
4. Frontier, S. in *Developments in Numerical Ecology* (eds Legendre, P. & Legendre, L.) NATO ASI Ser., G**14**, 335–378 (Springer, Berlin, 1987).
5. Weingart, P., Sehringer, R. & Winterhager, M. in *Handbook of Quantitative Studies of Science and Technology* (ed. Van Raan, A. F. J.) (North Holland-Elsevier, Amsterdam, 1988).

Scientific Correspondence

Scientific Correspondence is intended to provide a forum in which readers may raise points of a scientific character. They need not arise out of anything published in *Nature*. In any case, priority will be given to letters of fewer than 500 words and five references. □