Scientometrics, Vol. 20. No. 3 (1991) 439-449

FRACTAL GEOMETRY OF INFORMATION SPACE AS REPRESENTED BY CO-CITATION CLUSTERING

A.F.J. VAN RAAN

Centre for Science and Technology Studies, University of Leiden
Wassenaarseweg 52, P.O.Box 9555, 2300 RB Leiden (The Netherlands)

(Received June 25, 1990)

In this paper we discuss geometrical properties of 'information space' as represented by the phenomenon of co-citation clustering. More specifically, the size distribution of co-citation clusters is studied and interpreted in terms of fractal dimensions.

1. Introduction

Like many natural phenomena, the growth of scientific knowledge appears to be cluster-like. This seems to be true in a physical sense. On a spatial scale, scientific discoveries mainly 'cluster' around important universities, governmental and industrial research institutes. On a temporal scale, scientific discoveries often occur in a relatively short period of time, since an important breakthrough makes new advancements possible. However, these spatial and temporal clusterings are not the subject of this paper. We here focus on the 'non-physical', i.e. abstract structure of scientific information, a structure in which pieces of information are grouped together according to specific rules which govern the 'aggregation' of these pieces of information. More specifically, we shall discuss geometrical properties of co-citation clustering. In particular, this paper addresses the size distribution of these clusters in terms of fractal dimensions.

The 'fractal dimensions' is a geometrical factor providing a global description of scale-invariant irregularities and fragmentation. With scale-invariant we mean a (statistical) self-similarity: in a structural hierarchy, each level is an up-sized or down-sized version of the level below or above it. After the introduction of fractals into statistical physics in the beginning of this decade, fractals are now an important and growing field in the whole discipline of physics, and in other disciplines and fields as well (chemistry, biology, even medicine and, for instance here, information science). The reasons is clear: fractals are observed in many branches of science, in particular

Elsevier, Amsterdam – Oxford – New York – Tokyo
Akadémiai Kiadó

with respect to aggregation phenomena (dendritic growth, gelation, polymerization, percolation, tumor growth, epidemics, forest fires, cloud formation, etcetera).

The structure of this paper is as follows. We first give a short introduction on the basic principles of co-citation clustering. Important empirical findings on the cluster size distribution are presented. Next, a short introduction on fractals and fractal dimensions is given. Finally, the fractal model is applied to co-citation cluster size distribution and some preliminary conclusions with respect to structural aspects of 'information space' are drawn.

2. Co-citation clustering: principles and empirical findings

When a scientific paper A1 cites two earlier papers b1 and b2, these latter papers are 'co-cited'. The strength of such a co-citation relation is determined by the number of citing papers (A1, A2, A3,...) having the above pair (b1, b2) in their lists of cited papers (references). But paper b2 can also form a co-citation pair with a third paper b3, etcetera. Thus one can analyse all citing papers of a specific field in a specific publication year with such a co-citation algorithm. In this way, clusters of (co-) cited papers emerge, i.e. structures of interlinked co-cited pairs, and a 'map' of that field, at least in terms of citation practices, can be created. In this paper we will not discuss the many methodological and techniques details of co-citation analysis (citation and co-citation thresholds, calculation of strengths, the display ('mapping') of the emerging clusters by multivariate analysis techniques, etcetera). Pioneering work in co-citation analysis was done by Small and co-workers at the Institute of Scientific Information (ISI, Philadelphia) with the *Science Citation Index*. For thorough presentations and discussions of the basic co-citation methods and techniques we refer to the work of *Small* and colleagues.[1-3]

In the ISI co-citation analysis, an iterative clustering procedure is involved. The above described co-citation clustering is in fact the first step (C1): input are cited papers, and the output are clusters of cited papers. In the C2-step, the C1 clusters are input in a further clustering process, the output thus are clusters of C1 clusters. With each step, the units of analysis become more highly aggregated. To give an idea of the data involved, we use the example given by *Weingart* et al.[4]

We start with all publications in the 1984 *Science Citation Index* (*SCI*) and *Social Science Citation Index* (*SSCI*). These are some 660000 papers with nearly 10^7 citations to about 6×10^6 unique earlier papers. First a citation threshold is introduced: only papers cited more than 5 times are selected. Another threshold

relates to the different citation practices in different fields: scientific field differ considerably in length of reference lists. ISI (*Small* and *Sweeney*[2]) therefore introduced 'fractional counting': each citing paper has a total 'voting' of one point, and that single vote is divided equally among all references. Only cited documents with a fractional count of 1.5 are selected. The following technical details are not essential to understand the main points of this paper, but they give the reader who is not familiair with co-citation analysis an idea about the clustering procedure.

The above mentioned threshold reduce the source of 6×10^6 cited papers to about 7×10^4 selected cited papers (in simple terms: the highly cited papers). Now the algorithm, as described above, of identifying co-cited papers starts. First single pairs are identified, and then those pairs are linked together having one common paper. Thus structures of papers interlinked by co-citation, the co-citation clusters, are created. The co-cited papers in the cluster are often called 'core papers'. Membership of a cluster is determined by the co-citation threshold: the 'raw' co-citation count divided by the square root of the product of the individual citation counts for the cited papers in each co-cited pair ('co-citation strength') must be $> =$ 0.17. This means that from the 7×10^4 highly cited papers about 5×10^4 participate in the co-citation clustering. Further, a maximum cluster size of 60 (co-cited) papers is introduced. If a created cluster exceeds this limit of linking 60 co-cited papers together, the above co-citation threshold is increased in order to keep the cluster at the maximum size of 60 core papers. This, however, is a relatively rare event, as can be proved simply by listing the co-citation clusters as a function of the number of core papers. What we sketched sofar, is the earlier mentioned C1 step: a first clustering of the selected original source (*SCI* plus *SSCI*). About six million cited papers have been reduced to some 7×10^4 highly cited papers, and with these papers nearly 10^4 C1-clusters are formed with altogether some 5×10^4 highly cited (and co-cited) papers.

These 10^4 C1-clusters can then be used as input for a second step clustering: the creation of 'superclusters' composed of these C1-clusters. The result is about 1400 C2-clusters, each of them having 2 to 60 clusters of the first (C1) cluster generation. The iteration procedure in then performed twice again, the 1400 C2-clusters being input for about 180 C3-clusters, and these latter clusters being input for the final C4-clustering, yielding 21 C4-clusters.

Roughly, each iteration step reduces the number of (super)clusters with an order of magnitude. In general, one may say that at the C1 level science is structured in terms of (small) research specialties, whereas at higher levels the co-citation clusters

become increasingly extensive in size and therefore represent more and more higher hierarchical structures like subfields, fields and disciplines. But because of the artificial character of the clustering procedures, one has to be careful in the explanation of the meaning of the clusters at different levels. In this paper we do not discuss these important cognitive aspects, but focus on the phenomenon of clustering scientific literature as a well-described aggregation process with specific elements and specific rules.

We finish this general introduction to co-citation clustering with two additional points. So far, we discussed the clustering of *cited* 'core' papers. For each cluster, however, we can identify the papers *citing* these clustered core papers. These citing papers form, so to speak, a 'cloud' around the cluster core, and therefore all co-citation clusters can also be described in terms of citing papers. Throughout this paper the concept of 'cluster size' refers to the number of *citing* papers involved. Citing papers actually reflect the 'research front' of current research. Cited papers constitute the 'older', common base of these citing papers. Therefore, the number of citing papers gives a better measure of the size of the specialties concerned (as far as pictured by co-citation structures) than the number of cited papers.

We now present data on the size-rank distribution of C2 and C3 co-citation clusters. The C3 data were collected with help of the online version of the *Science Citation Index* (SCISEARCH, at the host DIMDI, Cologne). A special part of SCISEARCH provides data, in particular number of citing papers, for the co-citation clusters as calculated by ISI. We used the 1984 clustering, and the 1986 citing papers to these 1984 core papers. There are 179 of these C3-clusters. The data for the C2 clusters can be retrieved in the same way. At the C2 level there are 1371 clusters, which is quite numerous for online retrieval. The C2 statistics, however, was kindly provided to us in collaboration with *Weingart* and his colleagues, since these data[5] were readily available in the context of their co-citation analysis project.[4]

In Figures 1 and 2 the empirical results of the size-rank distribution are presented on a log-log scale. Figure 1 shows this distribution for the C2 co-citation clusters, and Fig. 2 for the C3 clusters. For the largest part of the ranking scale, over about two orders of magnitude, we find a power-law dependence and the distribution can be represented by an exponential expression of the general form:

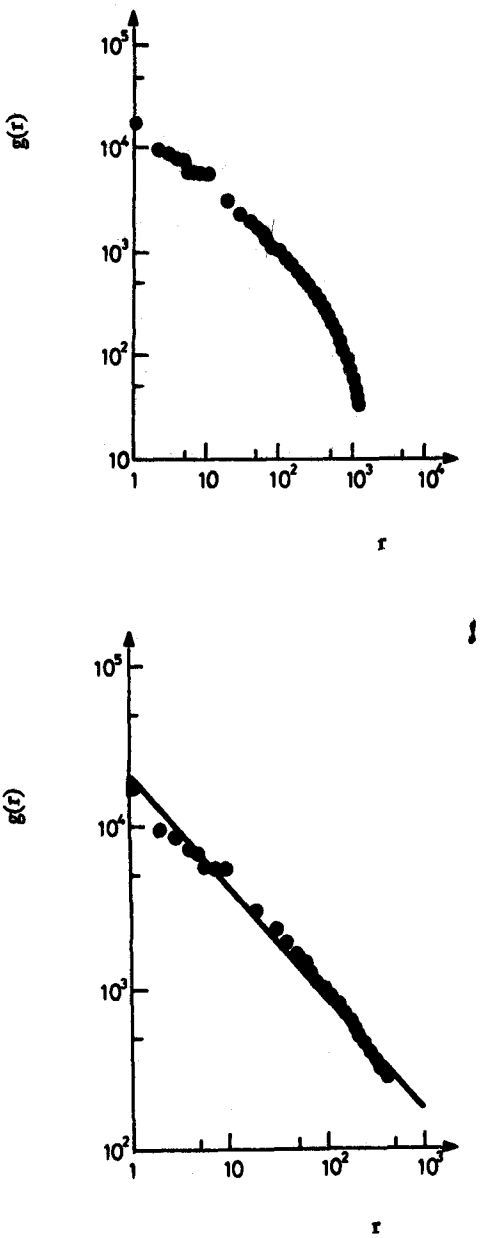

Fig. 1. Size-rank distribution of C2 cocitation clusters. Upper part: distribution for all 1371 C2 clusters; lower part: distribution for the power-law part

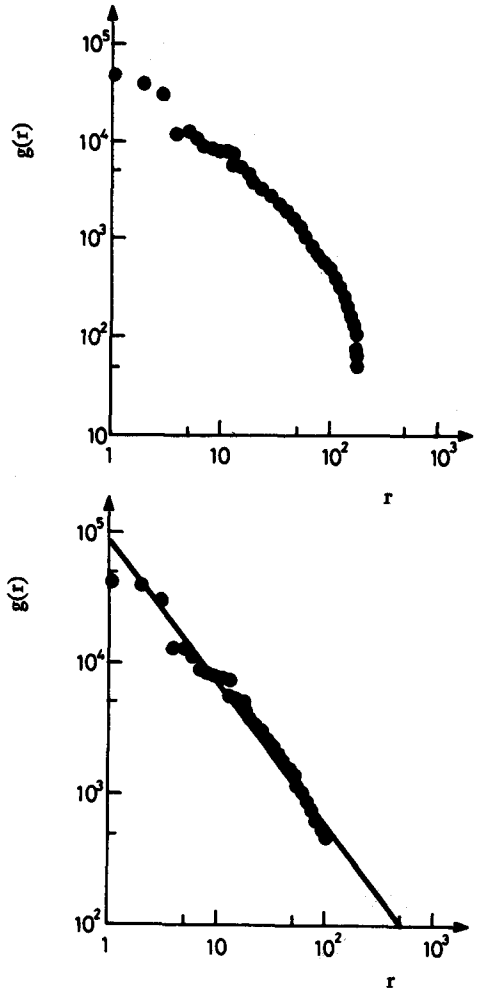

Fig. 2. Size-rank distribution of C3 cocitation clusters. Upper part: distribution for all 179 C3 clusters; lower part: distribution for the power-law part

$$g(r) = k.r^{-\gamma} \qquad (1)$$

with g(r) is the size (i.e., the number of citing papers) of the cluster with ranking number r; k is a value which can be determined from the empirical results, and γ is the slope of the line as given by Eq.(1).

The empirical value for γ in the case of C3 clusters is γ (C3) = 1.09, and for the C2 clusters γ(C2) = 0.71. Thus, a first clear finding is the obvious difference between both cluster systems with respect to this parameter γ.

The next step is the conversion of our size-ranking distribution function (Eq. 1) into a 'usual' size distribution function: the number of clusters with size g, n(g), as a function of g. The relation between both distributions is as follows:

$$n(g) = -dr/dg = (k^{1/\gamma} / \gamma) \cdot g^{-(1/\gamma + 1)} = k' \, g^{-(1/\gamma + 1)} \qquad (2)$$

This normal size distribution function allows for convenient comparison with other size distributions. In particular, size distributions involved in fractal structures will be discussed in the next section of this paper.

3. Fractals, scaling, and size distributions

Consider a wrinkled line, consisting of smaller wrinkles which, on their turn, as revealed by further magnification, consist of even smaller wrinkles. The most striking natural example is a coastline. Try to measure the length of the costline. We'll find that the smaller the length scale we use as unit of measure, the longer the coastline appears to be. For a thorough discussion of this remarkable phenomenon we refer to Mandelbrot's book.[6,7] If λ is the unit of measure, then the number of units required to measure the length of the wrinkled line is $N(\lambda) \sim \lambda^{-D}$, and the measured length is $L(\lambda) = N(\lambda) \cdot \lambda \sim \lambda^{1-D}$. The parameter D is the dimension of the wrinkled line, and its value lies between one (the dimension of a smooth line) and two (the dimension of a uniformly surface filling from). In the case of a coastline geometry, D has a non-integer value. Therefore it is called a *fractal* dimension, and the object involved (in this case the wrinkled line or coastline) is called a *fractal*. Since a natural coastline is not a mathematical, systematic structure, we could further specify it as a *statistical fractal*. A simple example of a mathematical, systematic, artificially constructed fractal is the Koch curve. Starting with a line of unit length (the 'initiator'), one divides this line in three segments of equal length and replaces the middle segment by an equilateral triangle with side length $l = 1/3$. This new curve is called the 'generator' with total length 4/3. By successively replacing each line segment of this generator by its scaled-down version (in the n^{th} iteration the triangle side length is $l = 1/3^n$), a *triadic Koch curve* with fractal dimension D = ln4/ln3 is constructed. We again refer to *Mandelbrot*[6,7] and to the recent work of *Feder*[8] for detailed discussions.

The above means that by scaling the size of a fractal system (such as a Koch curve or a natural coastline) by a factor α all geometrical quantities – such as the contour length – will change by α^D, for any value of α. In other words: a fractal system is self-similar with respect to the scaling factor α.

The next step is from fractal curves to fractal surfaces or 'clusters'. Instead of a straight line as a Koch curve initiator, one begins, for instance, with a equilateral triangle and applies similar generators to each side of the triangle as in the case of the Koch curve. In this way, we construct a (triadic) Koch island with a fractal coastline of the same dimension as calculated above. The curve length is now replaced by the perimeter of the island. It is not too difficult to calculate the relation between the perimeter (P) and the area (A) of the fractal island. We find $P \sim A^{D/2}$ (Refs 8,10-12).

Although geometrical concepts such as clusters perimeters of cluster areas could be applied, in principle, also in the case of co-citation clusters (by using appropriate data-analytical techniques), we here prefer to focus on the size of a clusters in terms of the number of citing papers. In fact, we could see this number as the 'volume' of a co-citation cluster, with each citing paper as a unit element. We know that in 3-dimensional Euclidean space the number of unit (volume) elements (for a ball of radius R) is given by $N(R) \sim R^3$. The 2-dimensional case gives $N(R) \sim R^2$, here N(R) is the number of unit (surface) elements. *Mandelbrot*[7] proved that the rule $N(R) \sim R^D$ applies to fractal, self-similar structures, with fractal dimension D.

Now we proceed with a very important element of fractal theory: the size distribution of fractal structures. We refer again to *Mandelbrot*[7] for his original approach to this problem. By splitting the Koch fractal curve generator into disconnected portions, a coastline generator and an island generator is constructed. After some algebra we find (*Mandelbrot*, Ref.7, p. 117-118) a size (i.e., area) distribution function of the form

$$N(A>A_o) \sim A_o^{-D/D_c} \qquad (3)$$

which indicates the number of islands with area A larger than A_o, with D is the fractal dimension of the complete set of islands, and D_c is the fractal dimension of the coastline of the individual islands. This means that D is a geometric indicator of the *fragmentation* of the whole distribution, whereas D_c is a geometrical indicator of the *irregularity* of the form of individual islands. It is remarkable that the above size distribution not only applies to the described strictly mathematical (Koch) procedure,

but also to the 'statistical' fractal islands created by procedures based on Brownian motion (the 'Brownian fractal landscapes').

We now transform this distribution function into a distribution function for islands having (precisely) area A:

$$n(A) = d/dA\ N(A>A_o) \sim A^{-(D/D_c+1)} \qquad (4)$$

The concept of island area can be compared with the size or 'volume' of the co-citations clusters, i.e. g=A. In fact, the citing papers constitute the cluster size, and in geometrical terms we may consider them as unit surface or unit volume elements. The size distribution of co-citation clusters was given in Eq.(2), and we find by comparison of Eq.(2) and (4):

$$1/\gamma = D/D_c \qquad (5)$$

We are primarily interested in the fractal dimension D of the co-citation cluster distribution. We now assume, in first approximation, that the form of the co-citation clusters is regular, i.e. the clusters have a rather smooth 'coastline', which means $D_c=1$. As a consequence, we find with help of Eq.(5) the simple relation:

$$\gamma = 1/D \qquad (6)$$

This finding is in agreement with the *Mandelbrot* generalization of the Zipf frequency distribution (*Mandelbrot*,[7] p. 344-348), recently used to describe species size distribution in ecosystems (*Frontier*[9]).

4. Results and discussion

With the above outlined fractal model we can determine the fractal dimension of the co-citation cluster size distribution.

For the C2 clusters we found (see Fig. 1) $\gamma(C2) = 0.71$, so that $D(C2)=1/\gamma(C2)=1.41$. In the case of the C3 clusters (Fig. 2), $\gamma(C3)=1.09$, which gives $D(C3)=0.92$.

We observe that the fractal dimension of the C2 clusters is significantly higher than for the C3 clusters. In the introduction to co-citation analysis, we pointed out that C2 clusters are of a lower level of aggregation, i.e. a multitude of smaller

'subfields' of science are formed, whereas at the C3 level a further aggregation to less but larger fields of science takes place. These results are in qualitive agreement with the findings of *Mandelbrot*[7] on the earlier mentioned 'Brownian fractal landscapes', in particular the relation between degree of fragmentation and the fractal dimensions of island size distributions. A striking characteristic of these fractal landscapes in that the higher the fractal dimensions is ($1 \leq D \leq 2$), the more fragmentation of larger islands occurs.

What is the meaning of a fractal dimensions of information space as represented by co-citation clustering? The fractal distribution looks, at first sight, some sort of static distribution, but it is in fact a snapshot of a dynamical process, reflecting the presence of older, established fields and the emergence of new specialties.

The question now is, what property of the scientific enterprise is pointed out by the fractal geometry of co-citation cluster size distribution? Like fractal distributions in ecological systems (*Frontier*[9]), we could consider co-citation clusters as *a* representation (and of course not *the* representation) of the 'ecosystem of scientists'. In this model, the structure of the ecosystem is strongly related to some sort of optimal distribution of energy, mass, and information. If the co-citation clusters represent 'species of scientists', then the fractal cluster distribution gives a measure of the diversity of the research community, i.e. the distribution of individuals among species. This fractal distribution, in analogy with ecosystems, could be the result of an optimization of flows of 'mass' (scientists, budgets) and information. It is important to understand why for the small C2 and C3 clusters the fractal distribution is not valid (see Figs 1 and 2). We conjecture that for the small clusters the parameters which determine their structure and their relations (such as flow of people, money and information) are much more subjected to a random process, whereas for the larger clusters apparently the underlying dynamics follows particular patterns which give rise to fractal distributions. Perhaps one could say that a sort of phase transition, toward a specific 'crystallization' of scientific information takes place. From physics we know that fractal geometry is closely related to the problem of describing the propagation of order in non-equilibrium (irreversible aggregation processes) systems. Therefore, although being not more than a global property, the fractal model of co-citation clustering is a very interesting starting point for further modelling of scientific 'ecosystems'.

Work is in progress to study the meaning of fractal structures in science more extensively. Very recent preliminary results on the fractal dimension of C1-clusters are in line with our above results.[13]

References

1. H. SMALL, E. GARFIELD, The geography of science: Disciplinary and national mappings, *Journal of Information Science*, 11 (1985) 147-159.
2. H. SMALL, E. SWEENEY, Clustering the Science Citation Index using co-citations, I: A comparison of methods, *Scientometrics*, 7 (1985) 391-409.
3. H. SMALL, E. SWEENEY, E. GREENLEE, Clustering the Science Citation Index using co-citations, II: Mapping science, *Scientometrics*, 8 (1985) 321-340.
4. P. WEINGART, R. SEHRINGER, M. WINTERHAGER, Bibliometric indicators for assessing strengths and weaknesses of West German science, In: *Handbook of Quantitative Studies of Science and Technology*, A.F.J. VAN RAAN (Ed.), Amsterdam, North Holland/Elsevier Science Publishers, 1988.
5. As in the case of the C1 data, the C2 data are also extracted from ISI's *Science Citation Index*.
6. B.B. MANDELBROT, Fractals. Form, chance and dimension, San Francisco, W.H. Freeman & Co., 1977.
7. B.B. MANDELBROT, The fractal geometry of nature, San Francisco, W.H. Freeman & Co., 1982.
8. J. FEDER, *Fractals*, New York / London, Plenum Press, 1988.
9. S. FRONTIER, Applications of fractal theory of ecology. In: *Developments in Numerical Ecology*, P. LEGENDRE, L. LEGENDRE (Eds), NATO ASI Series, Vol. G14, Berlin/Heidelberg: Springer-Verlag, 1987, pp. 335-378.
10. F.S. RYS, A. WALDVOGEL, In: *Fractals in Physics*, L. PIETRONERO, E. TOSATTI (Eds), Amsterdam, North Holland/Elsevier Science Publishers, 1986, pp. 461-464.
11. R.F. VOSS, 1985. Random fractals: Characterization and measurement, In: *Scaling Phenomena in Disordered Systems*, R. PYNN, A. SKJELTORP (Eds), NATO ASI Series, Vol. B133, New York/London; Plenum Press, 1985, pp. 1-11.
12. R.F. VOSS, R.B. LAIBOWITZ, E.I. ALESSANDRINI, Fractal geometry of percolation in thin gold films, In: *Scaling Phenomena in Disordered Systems*, R. PYNN, A. SKJELTORP (Eds), NATO ASI Series, Vol. B133, New York/London, Plenum Press, 1985, pp. 279-288.
13. A.F.J. VAN RAAN, Fractal dimension of co-citation *Nature*, 347 (1990) 626.