Scientometrics, Vol. 7. Nos 3–6 (1985) 391–409

CLUSTERING THE *SCIENCE CITATION INDEX*® USING CO-CITATIONS

I. A COMPARISON OF METHODS

H. SMALL, E. SWEENEY

Institute for Scientific Information, 3501 Market Street, Philadelphia, PA 19104 (USA)

(Received June 6, 1984)

Earlier experiments in the use of co-citations to cluster the *Science Citation Indey (SCI)* database are reviewed. Two proposed improvements in the methodology are introduced: fractional citation counting and variable level clustering with a maximum cluster size limit. Results of an experiment using the 1979 *SCI* are described comparing the new methods with those previously employed. It is found that fractional citation counting helps reduce the bias toward high referencing fields such as biomedicine and biochemistry inherent in the use of an integer citation count threshold, and increases the range of subject matters covered by clusters. Variable level clustering, on the other hand, increases recall as measured by the percentage of highly cited items included in clusters. It is concluded that the two new methods used in combination will improve our ability to generate comprehensive maps of science as envisioned by *Derek Price*. This topic will be discussed in a forthcoming paper.

Introduction

This is the first of a two part report on methodological developments in the use of co-citation analysis to cluster the *Science Citation Index*. This first paper examines a new procedure for setting citation frequency thresholds, called fractional citation counting, and a new approach to clustering which we call variable level clustering. These new methods are compared both individually and in combination with the methods which are presently in use at the Institute for Scientific Information (ISI®). The second report will show how these methodological developments, when applied iteratively to successively more aggregated units, can lead to an improved mapping of science. This work can be seen as part of a research program, begun some twelve years ago and inspired by Derek Price, to create annual maps for all of science. Such maps, it was hoped, can indicate the state of science in a particular year, and by their changes from year to year, the overall progress of science. *Price*'s vision is now becoming a reality through ISI's *Atlas of Science* system.[1]

The first experiments to cluster the *Science Citation Index (SCI®)* using co-citation were carried out by Small and Griffith using one-quarter of an *SCI* year.[2] In this early work, integer citation thresholds were used to select highly cited papers, and co-citation was also defined on a integer basis. Clustering thresholds were set in terms of the integer co-citation frequency to perform single-link clustering. Following these experiments, ISI began applying the single-linkage clustering algorithm to entire annual files of the *SCI* from 1970 to 1974.[3]

The use of integer co-citation counts to set clustering thresholds, however, presented some difficulties. First, it was necessary to delete very highly cited papers, such as biomedical methodology papers, since they tended to tie together large portions of the file. Second, integer co-citation counts introduced a size dependency–highly cited papers also tended to be highly co-cited–which biased the analysis against smaller research areas.

Beginning in 1975, therefore, co-citation normalization was introduced as a way to partially overcome the problems of highly cited method papers and size dependency.[4] The 1973 and 1974 *SCI* files were reclustered using the so-called Jaccard normalization which is the integer co-citation count divided by the number of unique papers citing the document pair. More recently, *Salton*'s cosine formula[5] has been adopted for normalization because it deals with links between high and low cited papers more effectively. Using normalized co-citation it is possible to obtain many more clusters at a single co-citation level than with integer co-citation and the clusters obtained provide a more comprehensive representation of small and large research areas. Eventually, all years of the *SCI* from 1970 to 1980 were clustered using normalized co-citation at integer citation thresholds between 15 and 17 citations per document. Generally, the normalized co-citation thresholds were selected to create the largest possible number of clusters for the particular file. At these levels the largest cluster usually remained in the range of 100 highly cited documents.

Despite the improvements brought about by co-citation normalization, problems remained regarding the comprehensiveness of the clustering, and how broadly representative it was of the various scientific fields. For example, it proved difficult to obtain an adequate representation for fields such as mathematics and engineering within the broad mix of biological and physical sciences using the relatively high annual citation thresholds mentioned above. These thresholds favored fields with strong referencing patterns and high publication volume. Lowering the annual citation threshold and normalized co-citation thresholds for clustering gave more clusters in the lower cited fields, but the bias toward fields with stronger referencing patterns, such as biomedicine and biochemistry, remained.

Such a biomedical bias was borne out by the work of Martha *Dean* who documented an increasing biomedical representation in annual cluster analyses of the *SCI* over the

decade of the 1970s, as part of ISI's cluster string project[6]. *Dean* found the proportion of biomedical clusters had increased from about 60% in the early 1970s to about 70% in 1979. Furthermore, ISI's BIOMED file,[7] created as a subset of the *SCI* beginning in 1979 and designed to cover biomedicine comprehensively, was known to cover only about 50% of the source articles in the *SCI*. These facts pointed strongly to a biomedical over-representation in the annual *SCI* clusters.

It was well known that biomedical papers had longer reference lists on the average than papers in mathematics and engineering.[8] This higher reference intensity per paper could affect the co-citation clustering procedures in two ways: 1. by increasing the number and proportion of biomedical items which fall in the highly cited range and are hence accepted for clustering, and 2. by increasing the strength and density of co-citation links formed among biomedical items which are used directly in clustering. Both factors would favor the formation of clusters in biomedical areas, and make it more difficult for clusters to form in areas where shorter reference lists are the rule.

In an annual slice of science, certain small areas of research will remain inaccessible to our analysis, simply due to the paucity of articles published in a given year. But for fields with adequate publication volume, it should be possible to compensate for smaller numbers of references per paper by some form of reference normalization. A reasonable objective seems to be that the number of clusters for a field be proportional to its source article representation in the data base. If five percent of the articles in a year are in mathematics, then five percent of the clusters should be on mathematical topics. The problem becomes how to compensate for the differences in referencing patterns from field to field and indeed from article to article.

New techniques

Fractional citation counting

The first step in the clustering system is to set a threshold for the minimum number of citations a document needs to receive in order to participate in clustering. In previous experiments and production runs at ISI, this threshold has varied from four to twenty citations, depending on whether a subset or the entire database was the object of analysis, and the time period of the citation data cumulation. Each citation was considered equal to every other and given a count of one, and hence all citation thresholds were whole numbers. The idea of using fractional citation counting was originally suggested by the late Tyler *Thompson* of Rutgers University in 1976 and independently by Martha *Dean* of ISI as a way of overcoming reference length bias. In fractional citation counting, each citing item has a total voting

strength of one, but divides that single vote equally among all references it cites. If an item contains ten references, each citation has a fractional weight of 1/10. This procedure has the generally desirable effect of giving papers with short reference lists greater weight, and papers with long reference lists, such as review papers, less weight per reference. Earlier, *Derek Price* had advocated the use of fractional paper counting for the evaluation of productivity of scientists, to counteract the growing tendency toward multiple authorship.[9] In this procedure each author of an "n" author paper receives 1/n credit, and the total productivity of an author is a sum of fractional values. While not strictly analogous to the concept of fractional citation counting, the *Price* proposal did suggest thinking of a source document as having a single unit of credit which it dispenses among the references it cites or the authors who wrote it. Clearly, the important concept here is that all source items have but a single unit of credit to dispense.

Figure 1 illustrates the concept of fractional citation counting and compares it with the traditional integer counting method. A hypothetical document "A" is cited by four later papers. The integer citation count weights each of these citations equally and gives each a value of one. The fractional method takes into account the length of the reference list of each of the citing papers, and apportions a single vote or unit of citing strength equally among each of the references it cites. Since citing paper No. 1 cites 5 previous papers in its reference list, item "A" receives 1/5 of a citing unit from paper No. 1. The fractional weights are totaled for the four citing papers

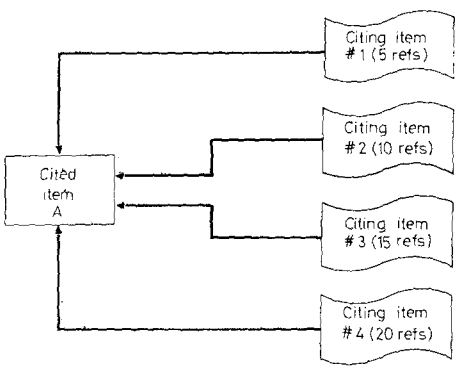

Fig. 1. Fractional citation counting. Integer citation count for A = 1 + 1 + 1 + 1 = 4. Fractional citation count for A = 1/5 + 1/10 + 1/15 + 1/20 = 0.417

$$\text{Fractional Citation Count} = \sum_{i=1}^{N} 1/R_i$$

N = Integer citation count

R_i = reference list length of citing item i

to give a fractional citation count of 0.417. The Figure gives the general formula for this counting method, which is a summation of the reciprocals of the reference list lengths of the citing papers.

Variable level clustering

Another methodology we examine in this paper is the use of an upper limit to cluster size coupled with variable co-citation levels, which we call variable level clustering. It is well-known that single-link clustering can be implemented most easily by setting a threshold for association and forming all connected groups whose links are at or above the threshold value.[10] Earlier work at ISI utilized this procedure. All co-citation links in the file are normalized using either the Jaccard coefficient or, more recently, the cosine function (Salton normalization, see Fig. 2), and all clusters are formed at a fixed value of this coefficient. Any number of cluster runs may be carried out at different values of this coefficient, but for any particular run, all clusters are formed at the specified threshold of normalized co-citation.

The problem with this approach is that the optimum co-citation level from the standpoint of recall and precision seems to vary from specialty to specialty, and it becomes difficult to select the best version of a cluster for each specialty from several different level runs. Thus, we developed a strategy whereby each cluster could form at a different, hopefully optimal, threshold. It was still necessary to decide what form of optimization should take place for each cluster within a single run of the program. The simplest parameter to limit is cluster size, the number of cited documents contained in the cluster. In other words, a cluster would form at the lowest possible co-citation level provided it did not exceed a certain specified cluster size. If the cluster exceeded this limit, the program would increment the level, and try to form the cluster again at a higher level. This upward level incrementing would continue until a cluster formed which did not exceed the size limit. This obviously breaks large clusters into smaller fragments, but since it allows the initial co-citation threshold to be set lower, it also allows smaller clusters to become larger. It also prevents the formation of amorphous macro-clusters by chaining, which is a problem with the single-link method when low co-citation levels are used.

While other stopping rules, such as limiting the density of linkages, were certainly possible, the maximum cluster size rule was relatively simple to implement, and found a general basis and rationale in Derek *Price*'s conjecture, first elaborated in 1963 in *Little Science, Big Science*[11], that there is an inherent maximum size limitation to an invisible college. In groups larger than about 100 members, *Price* argued, interpersonal communication between the members becomes difficult if not impossible, leading to

the breaking up of the group into smaller sub-groups. The existence of such a limit seems reasonable given the cognitive and social demands which would be placed on the individual scientist. While *Price* was not able to derive a rigid upper bound on invisible college size, empirical studies supported his conjecture. In particular the clustering of annual *SCI* files consistently gave mean cluster sizes in the range of 100 citing authors per specialty. (This roughly translates to about six highly cited works per cluster cited an average of 17 times each, and assumes that the number of citing papers assigned to a cluster is a rough measure of the size of the invisible college). We would argue, however, that the precise value of the maximum allowed cluster size is not critically important, as long as it is not so high to permit formation of heterogeneous subject groups and that the fragmentation which occurs at the size limit forms reasonable subdivisions of the larger specialties. In the second paper of this series, we will show how it is possible to preserve the relationships between such fragments by the iterative clustering of clusters.

The variable level cluster run requires the specification of three parameters: a maximum cluster size, a starting level for normalized co-citation, and a level increment indicating how large a step upward should be taken if the cluster formed is larger than the maximum. These three parameters, plus the initial citation frequency threshold (whether integer or fractional) are the only parameters that need to be set to define a unique clustering outcome for a specific citation file.

The variable level clustering method is illustrated in Fig. 2. A hypothetical tree structure is presented with the number of cited items captured by a given cluster enclosed in a box. The scale of normalized co-citation is indicated to the left of the tree ranging from 0.1 to 0.4. The actual formula for co-citation normalization used in these experiments is given below the tree. Assuming a maximum cluster size of 50 and a level increment of 0.1, at the starting level 0.1 a single cluster "A" containing 100 documents has formed. Since the cluster exceeds the size limit, the program increments the co-citation level by the specified amount, and the next level attempted is 0.2. At this level cluster "B" is formed with 80 cited items, still too large, but a cluster "C" is also formed with 20 items, which is acceptable. Since cluster "B" is still too large, the level is incremented to 0.3 and clusters "D" and "E" are formed, the latter of which meets the size criterion, while the former does not. Continuing the process we finally arrive at a set of clusters (C, E, G, and F) which represent the disaggregation of cluster "A" at level 0.1.

The variable level clustering program "flags" the records of the documents which have been successfully clustered and proceeds to the next unclustered item. Hence a complete file of co-citation pairs consisting of hundreds of thousands of records can be clustered by the variable level method, and the method remains order-independent

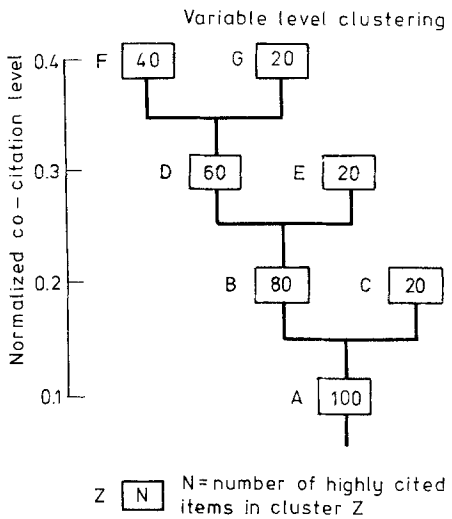

Fig. 2. Variable level clustering

$$\text{Normalized co-citation of items i \& j} = \frac{C_{ij}}{(C_i \times C_j)^{1/2}}$$

C_{ij} = number of co-citation of i and j
C_i = number of citations of item i (integer definition)
C_j = number of citations of item j (integer definition)

since all co-cited pairs remain accessible to the program even though the cited items they contain may already be clustered ("flagged") and can appear in only one cluster.

In combining fractional citation counting with variable level clustering, it is important to note that we did not use the fractional counts for the normalization of integer co-citation counts. The fractional counts were used only in the initial selection of cited items. We then determine the integer citation counts for the items which have been selected, obtain the integer co-citation values in the usual way, and normalize them using the integer citation counts. The fractional citation counts and the variable level clustering operate independently of one another, and hence their effects may be assessed separately. In the concluding section we will discuss how a fractional approach to the definition of co-citation could be implemented and suggest its likely effect.

Comparison of methods

The 1979 *Science Citation Index* was the data base used in these experiments. The machine readable version of this file consists of records each of which contains a citing item and a cited item. There are a total of about one-half million citing items in the 1979 *SCI*, and about 3.9 million unique cited items. The file contains 7.6 million citations (citing/cited pairs).

In order to compare the two new methods (fractional citation counting and variable level clustering) with the older methods (integer citation counting and constant level clustering), four experiments were carried out which allowed us to determine the effect of each of the old and new methods:

1. IC: Integer citation threshold with constant co-citation threshold,
2. IV: Integer citation threshold with variable co-citation threshold,
3. FC: Fractional citation threshold with constant co-citation threshold,
4. FV: Fractional citation threshold with variable co-citation threshold.

It was important, in the case of integer versus fractional citation counting, to select citation thresholds so that the methods could be reasonably compared. It was decided that equating on the number of citations received by cited items was more meaningful than on the number of cited items selected, since the latter would not take into account the lower mean citation rates of the fractionally selected cited items. Equating on total citations received by all items means that the retrieval potential of the cited items is the key factor.

Table 1 is a statistical summary of the four experiments carried out independently on the 1979 *SCI* file. Line 1 gives the integer and fractional citation thresholds used: 17 citations per document per year for the IC and IV runs, and 0.77 fractional citations per document for the FC and FV runs. As shown in line 2 these thresholds select more items as "highly cited" using the fractional threshold of 0.77 than are selected using the integer threshold of 17. In terms of total citations to all items (line 3), the thresholds are nearly equivalent, and it was on this basis that the fractional threshold of 0.77 was selected to match the integer cut-off of 17. Because the number of cited items selected is greater for the fractional file, the mean cites per cited item (line 4) is about one-half as large, indicating that more less-cited items are included in the fractional file.

The fractional citation counting method may be of interest independent of its effect on co-citation clustering. Let us digress for a moment to describe the cited items selected by the fractional threshold in terms of their integer citation values. Recall that following the selection of items by the fractional threshold the integer citation counts are determined for those same items. The frequency distribution of the number of items cited 1 to 49 times at the fractional cut-off of 0.77 is given in

Table 1

		IC	IV	FC	FV
1.	Integer or fractional citation threshold	17	17	0.77	0.77
2.	No. of cited documents selected at threshold	23 440	23 440	43 931	43 931
3.	Citations at threshold	703 513	703 513	697 416	697 416
4.	Mean cites per cited item	30.0	30.0	15.9	15.9
5.	No. of distinct co-cited pairs	1 834 390	1 834 390	1 103 607	1 103 607
6.	% Connected	0.66%	0.66%	0.11%	0.11%
7.	Normalized co-citation threshold	0.330	0.224+	0.280	0.180+
8.	No. of Clusters	2 317	2 460	3 695	3 932
9.	No. of Cited documents in clusters	11 557	15 480	14 744	21 149
10.	No. of Co-cited pairs Clustered	18 202	27 595	19 159	31 714
11.	Mean cited items per cluster	4.99	6.29	3.99	5.38
12.	Cited items in largest cluster	236	49	194	49
13.	% Clusters with two cited items	49.3%	42.4%	57.4%	48.3%
14.	% Cited items clustered	49.3%	66.0%	33.5%	48.1%

Fig. 3. The distribution for all items in the *SCI* cited 1 to 49 times is also given (the upper line in the Figure). The two distributions converge at about 30 times cited, or in other words, the fractional citation threshold of 0.77 selects nearly all items in the *SCI* cited 30 or more times in an integer sense. At lower citations rates, however, we see that the fractional threshold captures a smaller and smaller share of the cited items at a particular value of citation frequency. The number of cited items selected

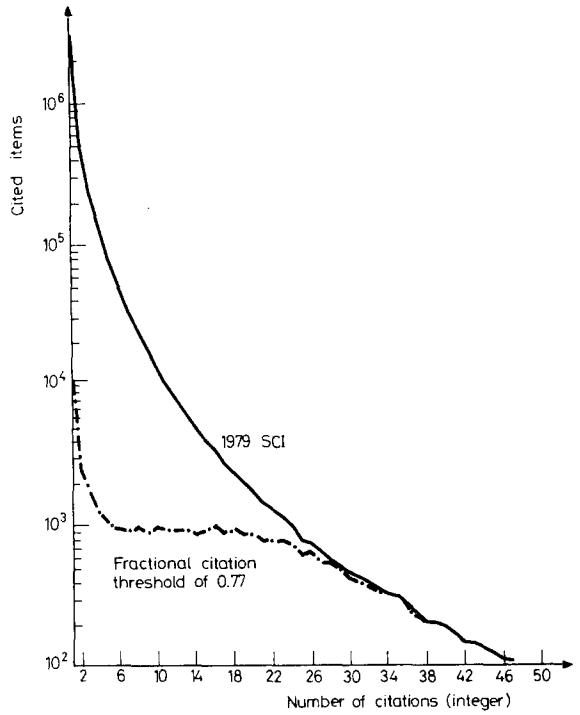

Fig. 3. Comparison of cited items selected by a fractional citation threshold with the overall 1979 *SCI* file

between 5 and about 15 citations per item is very nearly constant, even though the number of such items in the file as a whole increases by two orders of magnitude. The fractional threshold retrieves about one-half of the items cited 20 times. About 50% of the items retrieved fractionally are cited 10 times or less in the integer sense. Hence, the fractional cutoff has the effect of a "soft" integer cutoff.

Returning to our discussion of Table 1, lines 5 and 6 describe the co-citation characteristics of the selected cited items. It turns out that the integer threshold file with fewer cited items contains more co-cited document pairs than the fractional file with more cited items. Furthermore, the density of co-citation links, as measured by the "percent connected" of line 6, shows that the network of links is much more dense in the integer than the fractional file. Connectedness is calculated by dividing the number of unique pairs of co-cited items by the theoretically possible number of unique pairs based on all combinations of the cited documents selected. This means that the less cited items appearing in the fractional file do not interact as much as the more highly cited items in the integer file.

Several steps were taken to insure comparability across the four cluster runs. Line 7 gives the normalized co-citation threshold for clustering used in each file in terms of the Salton formula (see Fig. 2). In the two variable level files, IV and FV, a plus sign after the threshold value means that clusters were formed either at or above that level, or in other words, those values were the starting levels for variable level clustering. These starting levels were selected to draw upon an equal number of co-citation pairs in the IV and FV file, that is, about 47 000 co-cited item pairs (not shown in Table 1) were potentially available for the formation of clusters in both files. Hence, the same amount of data were made available to both variable level runs. In addition, a maximum cluster size of 49 items and a level increment of 0.01 were used in both runs. The principle of equalization was different for the two constant level runs, IC and FC. Both of these files were clustered at several normalized co-citation thresholds, and the threshold generating the largest number of clusters in each file was selected . These are the values given in line 7 for IC and FC. It proved difficult to find a basis for equating across the constant and variable level cluster runs, and thus the criterion of maximizing the number of clusters gave the constant level files their best chance to compare favorably.

The numbers of clusters generated in each file is given in line 8. The number of pairs actually utilized in cluster formation for all files is given on line 10, with the variable threshold experiments (IV and FV) consuming more co-citation pairs than the two constant threshold runs (IC and FC), even though thresholds for the latter were selected to maximize the number of clusters obtained. The same is true for the number of cited items included in clusters (line 9), although the differences here are not as pronounced as in the case of co-citation pairs. The FV run captured the largest number of highly cited items with 21 149, and moving from constant to variable levels increased the number of cited items clustered by about 30 percent for the integer files and 40 percent for the fractional files.

The mean cited items per cluster (line 11) indicates that the variable level clusters are larger on the average than the constant level clusters, but also that the integer clusters are larger than the fractional ones, whether they are constant or variable level. Using variable co-citation levels increases the average cluster size even with the imposed limit on the maximum size of 49 cited items, suggesting that the effect is due mainly to the small clusters growing larger. Line 12 indicates that the largest clusters obtained in the constant level files were 236 and 194 cited items for the IC and FC files, respectively.

Two additional measures of performance are: the percentage of clusters with two cited items (the smallest possible cluster size), and the percentage of the cited items clustered of those selected at the original citation threshold, whether integer or fractional (lines 13 and 14). For a less diverse or more cohesive file we expect to

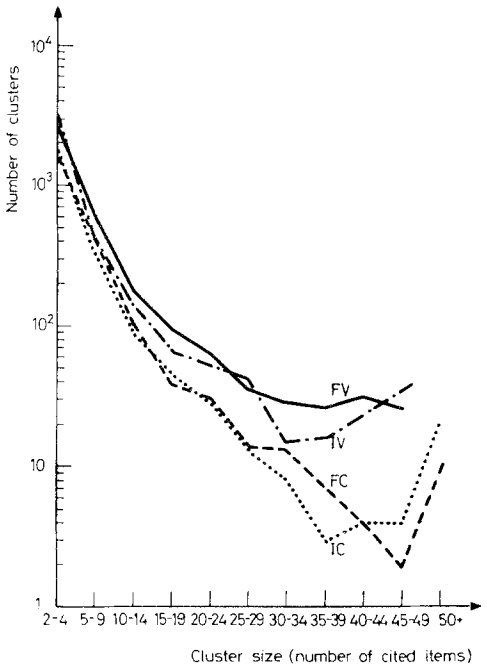

Fig. 4. Distribution of cluster sizes for the four files: Fractional-variable (FV), Integer-variable (IV), Fractional-constant (FC), Integer -constant (IC)

see proportionally fewer small clusters, and a higher percent of the cited items captured by clusters. On these measures, the IV showed the least diversity and highest cohesion, while the FC the most diversity and the least cohesion. The IC and FV files were comparable in this regard, perhaps indicating that what the fractional citation counting loses in cohesion of subject matter, the variable co-citation level regains by the ability to aggregate at lower levels. Going from constant to variable level clustering increases the percentage of cited documents clustered by about 15 percent.

Further information can be gained by examining the distribution of cluster sizes for the four runs. Figure 4 is a plot of the number of clusters (on a semi-log scale) obtained in specified ranges of cluster size from a low range of two to four cited items to a high range of 50 and greater. It is clear from the plot that the two constant level runs have a similar size distribution curve, and likewise the two variable level runs. The variable level runs have smaller ranges in the numbers of clusters of various sizes, while the constant level runs show a much wider range of variation. It is expected of course, that the constant level clusters vary more widely in size than the variable level ones which have a size limit. The up-swing at the high end of the distributions

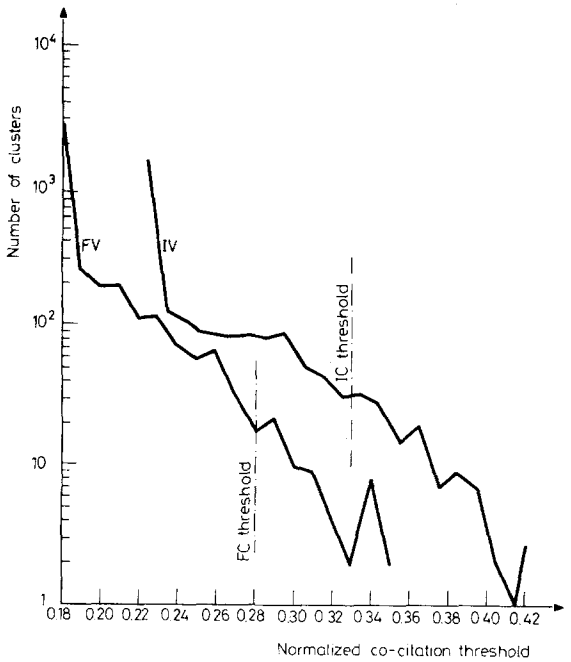

Fig. 5. Distribution of clusters over co-citation thresholds for Fractional-variable (FV) and Integer variable (IV) files

for all files is due to the inclusion of all larger clusters in one category for the two constant threshold files (50 cited items and up), and the tendency for the variable level clusters to collect large fragments at the upperbound of cluster size.

For the two variable level files we can make an additional comparison: the number of clusters formed at specific level increments of the normalized co-citation. Figure 5 is a plot of these distributions, again with the number of clusters created at each level on a log scale. Note that the FV file has a higher number and percentage (70%) of clusters at the lowest possible level compared to IV with 62% at the lowest level. Recall that the starting levels for the two variable level files were determined by equating the number of potentially usable co-citation pairs in each file.

Both the lower starting level and the higher percentage of clusters formed at the lowest level again indicate the greater diversity of subject matter in the fractional citation file (FV) and the greater cohesiveness of the integer file (IV). This is also borne out by the wider range of levels required by the IV file to create clusters no larger than 49 cited items, and the more rapid decline in the number of clusters requiring higher co-citation levels in the FV file. It was necessary to go to higher

Table 2

	IC		IV		FC		FV	
	No.	%	No.	%	No.	%	No.	%
Biomedicine and biochemistry	1683	72.6	1743	70.8	2300	62.3	2398	61.0
Physics	383	16.5	387	15.7	825	22.3	818	20.8
Chemistry	123	5.3	152	6.2	267	7.2	343	8.7
Mathematics	13	0.6	29	1.2	92	2.5	175	4.5
Geosciences	34	1.5	40	1.6	52	1.4	56	1.4
Other	48	2.1	51	2.1	112	3.0	96	2.4
Unknown	33	1.4	58	2.4	47	1.3	46	1.2
Total clusters	2317	100%	2460	100%	3695	100%	3932	100%

thresholds in the IV file in order to break up the large, cohesive areas represented in this file. The fractional file, on the other hand, contained fewer of these larger areas, and presumably a greater diversity of subjects, so that more of these areas emerged at lower levels.

The most telling comparison between the files is the disciplinary distribution. The categorization is only into major fields based on the journal of publication of the cited items and is hence rough. Nevertheless, it is sufficient to see the effects of the different procedures. Table 2 shows the approximate disciplinary categories for all clusters in the four files. The categories are: biomedicine and biochemistry, physics, chemistry, mathematics, geosciences, other, and unknown. The last named category pertains to clusters whose journals of publication were exclusively multidisciplinary, and hence could not be classified on the basis of journal title alone. The "other" category contains clusters in disciplines such as psychology, agronomy, environmental science, meteorology, etc.

Comparing the IC and IV files reveals only small differences in the disciplinary distribution, and this also holds for the FC and FV files. Apparently the effect of using variable levels is to reduce the biomedical and biochemical representation by only about two percentage points. Physics is also slightly reduced, but chemistry is increased. On the other hand, the effect of going from an integer to a fractional citation threshold has a dramatic effect on the biomedical/biochemical representation, reducing it by about 10 percentage points. Physics also increases by about 5 percent, and chemistry by about two percent. Math increases going from integer to fractional about 2 to 3 percent, while geosciences and the "other" and "unknown" categories remain about the same. In general the highest non-biomedical representations are attained by the FV file.

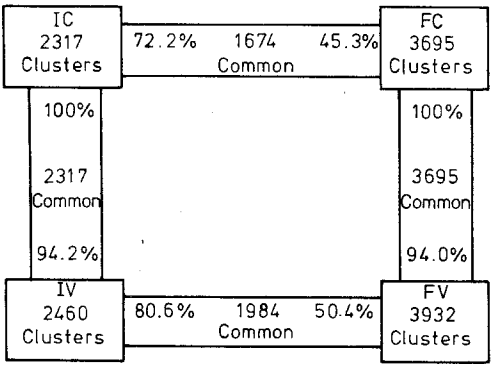

Fig. 6. Cluster overlap between the four files: Integer-constant (IC), Fractional-constant (FC), Fractional variable (FV), Integer variable (IV)

A final comparison of the four files was the overlap of clusters, or the number of common clusters generated. This comparison is possible because the same data base, the 1979 *SCI,* was used for all experiments. This overlap was determined by counting the common cited items in clusters in the four files. Even though the cited items in the fractional and integer files were selected on different criteria, a sufficient number of items were selected by both methods to enable us to match clusters on this basis. If we count the number of clusters in one file which share one or more cited items with a cluster or clusters in another file, we obtain a measure of the degree of overlap between clusters in the two files. We call this the number of "common clusters". In Fig. 6 we report four of the six possible comparisons: IC to IV; FC to FV; IC to FC; and IV to FV. The number of common clusters for each comparison is indicated within the connecting bar, and to either side, this number is expressed as a percentage of the total number of clusters. Examining first the "vertical" comparisons between IC and IV, and FC and FV, we see that 100% of the clusters generated by the constant threshold method are identified by the variable threshold method, whether using integer or fractional citation thresholds for selecting highly cited items. The variable level files identify 6% more clusters than the constant level method. This is true for both the integer and fractional citation files. Examining the "horizontal" comparisons between IC and FC, and IV and FV files, a more diverse pattern emerges due to the different populations of cited items selected by the fractional and integer methods. A greater commonality exists between the IV and FV files with 80% of the clusters identified by the integer citation method also identified by the fractional method, but only 50% of the fractional clusters identified by the integer method. These percentages

are somewhat lower for the IC-FC comparison, perhaps because the 6% fewer clusters obtained by the constant level files include many which are common to the integer and fractional files. In summary, the fractional method with variable levels (FV) is the most inclusive procedure of the four.

Discussion

The experiments indicate that fractional citation counting is a promising strategy for removing the biomedical bias which results from the use of integer citation thresholds as a basis for co-citation clustering. A future paper will show how well the FV file has succeeded in capturing a balanced representation of the various fields of science. In that paper we will show how it is possible to use variable level clustering in a multistage "clustering of clusters" which leads, after four iterations, to an overall map of science. One of the striking results is that only 47% of the higher level "clusters of clusters" on this overall map are biomedical in nature, compared to 74% biomedical for a similarly derived overall map, based on an integer citation cutoff. It seems, therefore, that fractional citation counting is a step toward the ultimate goal of mapping science in a comprehensive way, as envisioned by *Derek Price.*[1,2]

One potential problem of the fractional approach, however, is that it gives undue weight to citing papers which have very short reference lists, and cited items which have very low citation rates in the integer sense. In the case of a reference list of one, this single citation may suffice to place the item over the selection threshold, such as the 0.77 threshold used in this experiment. In fact, of the 43931 cited items selected at the 0.77 threshold, 11108 (25%) were cited one time by papers containing only one reference (a fractional citation count of one). Since these items were not co-cited, they had no effect on clustering.

To avoid including such infrequently cited items, one strategy is to have an initial low cut-off for integer citation frequency. In other words, drop all items cited fewer than N times, where N is not so great that it removes desired subject matters. If this is done prior to the calculation of the fractional citation counts, then the effective reference list length of a citing paper becomes shorter since all references to infrequently cited items have been removed. The fractional counts are then determined on the basis of truncated reference list lengths. This procedure, in addition to eliminating noise due to infrequently cited items, also introduces computer run-time savings in the initial stages by reducing the size of the citation files. In a recent run of the 1983 citation file at ISI, items cited fewer than 5 times (in an integer sense) were dropped prior to determining the fractional citation counts. Since reference lists were effectively shortened, the range of fractional counts was shifted upwards, so that a fractional

threshold of 1.5 yielded about 70 000 cited items (compared with 43 931 items for a 0.77 threshold in the 1979 file). It is not known as yet what effect this drop strategy has on the subject distribution of clusters, but initial indications are that it may even strengthen low referencing fields since their reference lists are shortened more.

It is also interesting to speculate what the effect would be of extending the concept of fractional citation counting to fractional co-citation counting, that is, converting the present integer co-citation approach to a fractional one. By analogy, fractional co-citation would assign a single unit of co-citing strength to each citing paper, and apportion that unit equally among all the pairs of references cited by that paper. If, for example, a paper cites "n" highly cited items, each pair of cited items would be assigned a weighted co-citation equal to $1/[1/2n(n-1)]$. The summation of all such fractional co-citation contributions from all citing papers for a given pair of cited items would constitute the fractional co-citation count for that cited pair. While experiments have not been carried out as yet using fractional co-citation, we might expect a thinning out of links among highly co-citing fields of science, and an enhancement of links in more weakly co-citing areas. This should parallel the redistribution of clusters by field observed with the fractional citation method, and extend the more balanced representation to structural features of the fields as well, e.g., the density of links.

In the experiments reported here the variable level clustering procedure was applied using only one combination of the three specifiable parameters for each of the files. For the FV (fractional-variable level) file these were: a maximum cluster size of 49, a starting co-citation level of 0.18, and a level increment of 0.01. Experience with the 1983 file and other files at ISI is, however, beginning to show how the three parameters affect the outcome of clustering. First, the higher the maximum cluster size parameter, the higher the percentage of initially selected cited items captured by clusters (recall). The recall in the FV file for the present experiment was only 48%, but for a similar 1983 file a recall of 71% was obtained using a maximum cluster size of 60 rather than 49. On the other hand, using a higher maximum cluster size *reduces* the total number of clusters since there is an added agglomerative effect. Secondly, lowering the starting co-citation level for clustering increases both the number of cited items clustered and the number of clusters created. Furthermore, there is some evidence that the lower this threshold, the more diverse are the subjects of clusters formed. Finally, increasing the level increment generates more clusters, which however contain fewer cited items, since there is less fine tuning of clusters below the maximum cluster size limit. The additional clusters generated are not new subject areas, but rather fragments of large clusters.

General guidelines for setting the variable level parameters are: set the maximum cluster size as large as possible without creating macro-clusters which incorporate multiple subject matters; set the starting co-citation level as low as possible to bring in as many diverse subject areas as possible; and set the level increment as small as possible to fine-tune clusters to fall within the size limitation.

The principal conclusion of the present study is that both the fractional citation counting method, and variable level clustering improve the results of clustering an interdisciplinary data base such as the *SCI* by increasing the percentage of non-biomedical clusters which emerge, and at least partially restoring subject area balance. Of the two prcedures, the fractional citation method has the greater effect in reducing the biomedical bias. The next step may be to extend the fractional concept to co-citation as well. The main advantage of variable level clustering is its ability to increase the recall of the clusters as measured by the percentage of highly cited items initially selected which are clustered. Possible increases in recall seem to be in the range of 20 percentage points. In the fractional file, the increase amounted to about 6 400 additional cited items included in clusters even though the number of clusters formed increased only by about 200. The ability to cluster at variable levels is also reflected in a decrease in the proportion of small clusters. Variable level clustering may have a larger impact on reducing disciplinary bias when much lower starting co-citation levels are used, since it will then be able to dig more deeply into the low lying co-cited pairs which have a greater subject diversity, particularly if a fractional citation cutoff is used.

References

1. E. GARFIELD, *ISI Atlas of Science: Biochemistry and Molecular Biology* Philadelphia, Institute for Scientific Information, 1981.
2. H. G. SMALL, B. C. GRIFFITH, The Structure of Scientific Literatures I: Identifying and Graphing Specialties, *Science Studies* 4 (1974) 17–40; B. C. GRIFFITH, H. G. SMALL, J. A. STONEHILL, S. DEY, The Structure of Scientific Literatures II: Toward a Macro- and Microstructure for Science, *Science Studies* 4 (1974) 339–65.
3. E. GARFIELD, M. V. MALIN, H. G. SMALL, Citation Data as Science Indicators, in: Y. ELKANA (Ed.) *Toward a Metric of Science: The Advent of Science Indicators* John Wiley, New York, 1978, 179–207.
4. H. G. SMALL, Structural Dynamics of Scientific Literature, *International Classification* 3, (1976) No. 2, 67–74.
5. G. SALTON, D. BERGMARK, A Citation Study of Computer Science Literature, *IEEE Transactions on Professional Communication* PC–22 (1979), No. 3, 146–58.
6. H. G. SMALL, H. R. COWARD, M. C. DEAN, E. J. GREENLEE, S. E. COZZENS, *Co-citation Clusters as Indicators of Specialty Development,* Final Report on NSF grant SRS-7912157 Philadelphia, Institute for Scientific Information, 1983, 58.
7. E. GARFIELD, ISI's On-Line System Makes Searching So Easy Even a Scientist Can Do It: Introducing Metadex Automatic Indexing and ISI/BIOMED Search (Editorial) *Current Contents*

(26 January 1981), reprinted in: E. GARFIELD, *Essays of an Information Scientist,* Vol. 5 ISI Press, Philadelphia, 1983, 11–14.
8. F. NARIN, *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity,* Final report on NSF contract C–627, Computer Horizons Inc., Cherry Hill, NJ, 1976, 171.
9. D. J. de S. PRICE, D. BEAVER, Collaboration in an Invisible College, *American Psychologist* 21 (1966) No. 11, 1011–18.
10. J. A. HARTIGAN, *Clustering Algorithms* John Wiley, New York, 1975, 119.
11. D. J. de S. PRICE, *Little Science, Big Science* Columbia University Press, New York, 1965, 63–91.
12. D. J. de S. PRICE, The Science of Scientists, *Medical Opinion and Review* 1, (1966), No. 10, 88–97.