Scientometrics, Vol. 15, Nos 5–6 (1989) 593–605

REPRESENTING A SCIENTIFIC FIELD: A BIBLIOMETRIC APPROACH

R. TODOROV

Centre for Scientific Information Bulgarian Academy of Sciences, BG-1040 Sofia, 7 Noemvri Str. No 1 (Bulgaria)

(Received June 16, 1988)

A new bibliometric method is proposed for representing links between subfields as defined by a classification scheme. The frequency of co-occurrence of articles from different subfields in selected journals is used for measuring the degree of relatedness between these subfields. The results of such quantitative analysis could be compared to the tree topology of the classification network established in a qualitative analysis. The method is applied to describe the internal links within the field of condensed matter physics using the 1984 Physics Abstracts database. A distinction is made between experimental and theoretical links on the basis of treatment codes assigned to journal articles. The links described by cluster analysis are matched against the cross-reference network of the International Classification for Physics.

Introduction

Co-citation analysis[1] has generated a number of discussions[2] since its first use for science mapping purposes.[3] At the same time it has been applied as a main or additional method in qualitative studies.[4] Independently of the convergence[5] or divergence[6] of results in such integrative approaches, many positive effects of the propagation of co-citation analysis have been registered. First of all, the method itself has been refined by introducing new measures such as normalized co-citation and fractional citation thresholds, as well as by using variable level clustering.[7] The statistical validity of co-citation graphs has also been investigated.[8] The shortcomings of single link clustering (e.g. highly chained clusters) have been noted.[9] Moreover, co-citation clustering has been replicated, in some sense, by another (co-word) analysis[10] which offers new opportunities for the validation of qualitative studies of science structure and development[11] as well as for a comparison of results from both quantitative methods.[12]

Co-citation clustering technique has also been extended and used for mapping science on a macro-level:[13] initial clusters (containing documents) are linked together into larger clusters to represent areas of current research.[14] It is evident that such macro-structures are rather difficult (if not impossible) to validate using a tra-

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

ditional qualitative approach: there are, for example, no experts on a higher level who could identify the sense of the aggregated clusters and their links.

In this paper we propose an alternative (bibliometric) method for representing a scientific field as defined by an established classification system. More specifically, we intend to describe the links between subfields within a given field and demonstrate a method involving a rather simple comparison of results against the cross-references (if available) of the classification scheme used.

A new bibliometric method

Bibliographic databases (or their printed versions) in science are created and developed on the basis of more or less elaborate classifications to categorize literature input. The classification schemes in use are in most cases discipline-oriented, and their (sub)divisions reflect in some way scientific (sub)fields or areas of current research. On the basis of such (even rough) classifications, off- and on-line bibliometric techniques could be (and are) applied to find the frequency distributions of articles over journal titles (JT) and (sub)fields F_i. Results are traditionally arranged in a matrix A, where the rows are the JT and the columns the F_i of the corresponding field F. An element of such a matrix expresses the number of articles of a given JT classified in the (sub)field F_i.

By applying a multivariate technique to this type of matrix[15] it is possible to describe the links between rows (objects) and columns (characteristics). Here we are interested only in representing the (internal) links between (pre)established subfields, i.e. to describe the tree topology of the field F on the basis of a classification. We assume that the frequency of co-occurrence of articles from different (sub) fields in a set of relevant journals may be used to measure the degree of association of these (sub) fields.

The method consists first in constructing the matrix A including all journal titles. It is evident that general journals which cover the field as a whole will introduce false links, since their articles are randomly or uniformly distributed. In order to select only 'specialized' journals, Pratt's absolute measure[16] of class concentration q for a given JT could be used:

q = SUM $i.a_i$ over all subdivisions n

with a_i proportion of articles from the JT in F_i.

A $q = 1$ is the extreme case for total concentration while $q = (n + 1)/2$ is the limit for the even partition case. A random (Whitworth) distribution is characterized by $q = w = (n + 3)/4$. We are interested only in journals with no random distributions, i.e. whose q-values are comprised in the interval $(1, w)$ and differ significantly from w at a given level of confidence.[17]

Once the subset of specialized journals is determined and they are selected as source journals, the initial matrix is reduced to them. A multivariate technique (cluster analysis of variables, factor or correspondence analyses) could be applied to describe the links between the (sub)fields under consideration in the set of the selected journals. For example, cluster analysis does not provide spatial representation but seems to show a better fit with respect to small dissimilarities. The correlation between the subdivisions could be selected as a measure of similarity.[18] A simple (indirect) test of the structure obtained could be made by changing the measure used. To define the distance between a new cluster and other clusters, the average link procedure could be selected as a compromise between the criterion of withincluster similarity or single linkage and that of betweencluster separability. The average similarity is the arithmetic mean of the correlation using all possible pairings of the variables (subfields) between two clusters.

The solution obtained (especially from cluster and factor analyses) could be simply compared to the cross-references (where available) between the (sub)divisions. Such links (on the macro-level) are traditionally determined by teams of experts (assembled in working meetings) "usually incompetent to handle the lower levels, which are more properly the work of specialists."[19] The cross-references (of 'see also' type) could also indicate the direction of relatedness between (sub)divisions. To compare both links of 'qualitative' (cross-references) and 'quantitative' types, a matrix could be constructed whose rows and columns are the (sub)divisions under consideration. In this matrix the links from the 'quantitative' analysis could be introduced by their strengths as well (e.g. the values of correlation).

Comment

The proposed method is based on the assumption that the frequency of co-appearance of articles from different subfields (i.e. with different main classification codes) in selected (specialized) journals could be used to describe and measure the relatedness of these subfields. The term 'specialized' is used to denote journals whose article partitions over the given (sub)fields are neither uniform nor random. The existence of such a set of specialized journals is a precondition for the application of the method. In other words, a field which is not represented by a 'sufficient' number of such journals is rather 'small' or selfcontained for analysis by itself on the macro-level. This means that the field is covered predominantly (if not exclusively) by 'general' journals and does not include distinct subdivisions. Such a field could be considered as a constituent subdivision by studying a unit from a higher level (e.g. a discipline).

The method is developed as a means for comparison. Its results could be regarded as alternative to co-citation maps on higher level of aggregation. It is evident that the method only reproduces a kind of tree topology in terms of (sub)field-to-(sub)field relations, i.e. subfield boundaries are predetermined by a classification scheme. Conversely, co-citation analysis is describing 'newly formed' structure units and their links by aggregating initial clusters of documents in larger groups of clusters. It appears, however, that co-citation maps on higher level of aggregation are more questionable as compared to specialty pictures, since it has not been established:

– how many and which articles are lost in the iterative clustering procedure:[20]

– how much the structure is a result of the choice of threshold values and/or the applied multivariate technique;

– what macro-clusters represent since their names are approximate;[21]

– why "clusters often merge[d] where common sense indicate[d] they should remain distinct";[22]

– why centre-periphery patterns predominate and why peripheral areas appear also to be less basic.[23]

In view of these problems, the proposed method has some advantages as far as "old-fashioned" field representation is concerned:

– links between 'traditional' subfields are 'directly' described on macro- (or journal-) level, i.e., there is no aggregation of clusters from lower levels;

– all subdivisions covered by a 'sufficient' number of journals could be included in the analysis, i.e., the method could be applied on higher levels of aggregation (discipline, science);

– results could be simply compared to cross-references assigned to classification subdivisions by teams of experts.

It remains to be seen whether the method describes well the cognitive links between (sub)fields of a given field, and what kind of additional information is obtained (as compared to the results of qualitative analyses).

Application in the field of condensed matter physics

Data and analysis

Data for journals and their articles in the field of Condensed Matter Physics (CMP) are extracted from 1984 INSPEC database (printed version: *Physics Abstracts*). One INSPEC document record contains not only a bibliographic description and an abstract: the subject content of a given document is represented by classification codes (assigned on the basis of the International Classification for Physics) and controlled and free index terms, as well as by treatment codes or research characteristics. These latter indicate particular aspects of the subject treated in a document (for example, theoretical, experimental, applied, etc.).

For the purpose of the study, each CMP journal article from the database has been assigned to the corresponding source journal title (JT), and to one of the 17 CMP subdivisions according to its main (first) classification code (CC), and to its treatment code (TC). These data (JT, CC, TC) have been recorded on a separate tape. Appropriate software has been developed for combining and arranging the JT as rows and the 17 CMP subdivisions as columns, correspondingly, for all articles (matrix A), and for the experimental (matrix E), and the theoretical (matrix T) ones. The names of the CMP subdivisions are given in Table 1. They correspond to the second level headings of the International Classification for Physics (ICP).

The next step has been to select the journals relevant for the analysis. This selection procedure consists in discarding journals with low productivity (for reasons of

Table 1
Subject subdivisions for Condensed Matter Physics (CMP)

Classification code	CMP subdivision
61	Structure of liquids and solids; Crystallography
62	Mechanical and acoustic properties of condensed matter
63	Lattice dynamics and crystal statistics
64	Equations of state, phase equilibria, and phase transitions
65	Thermal properties of condensed matter
66	Transport properties of condensed matter (nonelectronic)
67	Quantum fluids and solids: liquid and solid helium
68	Surfaces and interfaces; Thin films and whiskers
71	Electron states
72	Electronic transport in condensed matter
73	Electronic structure and electronic properties of surfaces; Interfaces and thin films
74	Superconductivity
75	Magnetic properties and materials
76	Magnetic-resonances and relaxation in condensed matter; Moessbauer effect
77	Dielectric properties and materials
78	Optical properties and condensed matter spectroscopy and other interactions of matter with particles and radiation
79	Electron and ion emission by liquids and solids; Impact phenomena

statistical validity), non-CMP journals (as the intention here is to describe only the internal links within the field of CMP), and general journals (which publish uniformly or randomly over the subdivisions and may introduce misleading relations). In all, 95 journals (from 718) have been selected as productive. They have been represented by more than 48 records in the CMP section of the 1984 INSPEC database. This threshold corresponds to a journal publishing 4 CMP articles per month. From this set, the non-CMP journals (for example, from other physics subfields, or cross-disciplinary areas) have been removed on the basis of their ratio of CMP articles to all articles classified in the INSPEC A database. This ratio is close to one (more than 0.9) for the CMP journals. Then, general journals have been discarded using *Pratt*'s absolute measure[24] of concentration q. The value of q for the random (Whitworth) distribution in this case is $w = (n + 3)/4 = 5$. For the analysis only specialized journals have been selected: their q-values are less than w and differ significantly from w at a confidence level of 0.95. In Table 2 the titles of 25 specialized journals retain-

Table 2
CMP source journals arranged alphabetically

Journal title	Country of origin	q value
Acta Metallurgica	USA	2.1
Crystal Research and Technology	GDR	2.7
Ferroelectrics Letters	GBR	2.2
Fizika i Khimiya Stekla	SUN	3.0
Fizika Metallov i Metallovedeniya	SUN	3.8
Fizika Nizkikh Temperatur	SUN	3.8
Fizika i Tekhnika Poluprovodnikov	SUN	2.7
Journal of Crystal Growth	NLD	1.6
Journal of the Less-Common Metals	SWI	2.9
Journal of Low Temperature Physics	USA	2.0
Journal of Magn. & Magnetic Materials	NLD	1.6
Journal of Non-Crystalline Solids	NLD	3.4
Journal of Polym. Science (Physics)	USA	2.8
Journal of Solid State Chemistry	USA	2.5
Kristallografiya	SUN	2.2
Molecular Crystals & Liquid Crystals	GBR	2.2
Philosophical Magazine A	GBR	2.7
Philosophical Magazine B	GBR	3.4
Radiation Effects	GBR	1.9
Scripta Metallurgica	USA	2.1
Solid-State Electronics	USA	2.0
Solid-State Ionics	NLD	2.1
Surface Science	NLD	1.8
Synthetical Metals	SWI	3.5
Thin Solid Films	SWI	3.4

Table 3

Tree printed over correlation matrix. Clustering by average distance method

```
Sub-
field
      -----------------------------------------------------/
 61    80/64/45 55 44 47 33 33/48/42 42 44 48 44 60 49/
        / /                  / /                     /
       / /                  / /                     /
 64   /68/59 74 58 54 40 41/47/41 40 39 40 36 44 47/
       /                  / /                     /
      /                  / /                     /
 62   /55 60 56 43 43 45/50/42 41 37 40 38 45 49/
                       / /                     /
      ----------------/ /                     /
 63    76/59 61/65 68/58/54 55 63 61 53 58 41/
        /     /     / /                     /
       /     /     / /                     /
 65   /64 70/58 62/62/38 39 44 43 35 43 40/
          /     / /                     /
      ----/    / /                     /
 75    90/50 53/45/45 45 51 48 45 43 42/
        /     / /                     /
       /     / /                     /
 76   /53 56/43/41 43 69 66 57 60 39/
          / /                     /
      ---/ /                     /
 67    99/42/45 46 54 47 45 40 44/
        / /                     /
       / /                     /
 74   /42/45 45 55 48 46 41 43/
       /                     /
      /                     /
 66   /44 43 44 47 47 47 47/
                           /
      --------------------/
 68    98/46 45 58 57/46/
        /           / /
       /           / /
 79   /48 47 56 57/46/
                  / /
      -----------/ /
 71    98/92/89/46/
        / / / /
       / / / /
 72   /93/92/47/
       / / /
      / / /
 73   /91/46/
       / /
      / /
 78   /52/
       /
      /
 77   /
```

The tree diagram shows the clusters formed during the amalgamation. The horizontal and diagonal lines determine the clusters. For example, 67 and 74 form one cluster; all the subfields listed from 71 to 78 form another cluster. The first number in each line is the measure of similiarity of the subfield to the left of the line with the one immediately below it.

The values in the tree have been scaled 0 to 100 according to the following table:

Value above	Correlation	Value above	Correlation
30	–0.4	65	0.3
35	–0.3	70	0.4
40	–0.2	75	0.5
45	–0.1	80	0.6
50	0.0	85	0.7
55	0.1	90	0.8
60	0.2	95	0.9

Table 4
Tree printed over correlation matrix.
(Experimental articles only)

```
Sub-
field
     ---------------------------------------------/
61     81/79/63/47/32 30 38 38 45 46 45 59 42 42/
        /  /  /  /                             /
       /  /  /  /                             /
77    /62/54/56/39 37 46 45 43 44 45 52 39 40/
        /  /  /                             /
       /  /  /                             /
64    /65/43/38 40 38 38 42 41 38 44 53 51/
        /  /                             /
       /  /                             /
62    /46/38 43 42 42 38 40 42 51 56 48/
        /                             /
       /                             /
66    /41 41 44 43 45 48 49 50 43 42/
                                   /
     -----------------------------/
67     97/44 45 52 45 44 37 45 44/
        /                       /
       /                       /
74    /44 43 52 45 44 35 48 47/
                             /
     -----------------------/
68     97/45 43 57 57 43 38/
        /                 /
       /                 /
79    /46 44 54 56 43 40/
                       /
     -----------------/
71     98/94/87/48 62/
        /  /  /     /
       /  /  /     /
72    /95/90/47 62/
        /  /     /
       /  /     /
73    /94/44 54/
        /     /
       /     /
78    /40 52/
           /
     -----/
75     93/
        /
       /
76    /
```

ed in the selection procedure are given together with their q-values. These 25 titles have been used (as rows) for the formation of a reduced matrix A (with the 17 CMP subdivisions as columns). The reduced matrices E and T contain 20 rows/15 columns and 10 rows/13 columns respectively. Some of the journals and the subdivisions have been excluded because of an insufficient number of experimental or theoretical articles.

Table 5
Tree printed over correlation matrix.
(Theoretical articles only)

```
Sub-
field
      -------------------------------------/
61     62 70 73 60/16 18 40 40 52 49 56 42/
                  /                      /
      -----------/                      /
62     93/89/67/43 41 34 31 32 32 24 26/
        /  /  /                       /
       /  /  /                       /
66    /91/71/34 35 37 35 46 43 33 39/
        /  /                      /
       /  /                      /
64    /78/32 32 41 39 39 35 28 33/
        /                     /
       /                     /
75    /51 55 39 41 67 49 39 42/
                            /
      ---------------------/
67     99/40 40 42 38 32 42/
        /               /
       /               /
74    /38 39 48 41 33 43/
                       /
      ----------------/
68     99/38 40 48 60/
        /          /
       /          /
79    /43 43 50 61/
                /
      ---------/
71     90 82 80/
              /
      ------/
72     94/94/
        /  /
       /  /
78    /91/
        /
       /
73    /
```

Further cluster analysis of variables (characteristics) has been applied to the three reduced matrices to describe the internal macro-links within the field of CMP. The program used[25] forms clusters on the basis of the correlation between the variables (the CMP subdivisions). The criterion applied here is based on the average linkage, which is the arithmetic mean of the variables between two clusters.

Results and validation

The results of the application of cluster analysis on the reduced matrices *A, E,* and *T* are shown as tree diagrams in Table 3, 4, and 5. Three of the clusters (67/74, 68/79, and 71/72/73/78) appear on all diagrams, while the cluster 75/76 is 'experimental', and 62/64/66 rather 'theoretical'. Similar results have been obtained

Table 6
Comparison of cross-references (upper-left) with quantitative links (down-right)

CMP subfields	61	62	64	66	63	65	75	76	67	74	68	79	71	72	73	78
78					Y								E2	E3	N	//
73													Y	Y	//	a
72													N	//	a	a
71	Y				Y								//	a	a	a
79											Y	//				
68											//	a				
74									N	//						
67									//	a						
76							E1	//								
75							//	e								
65					Y	//										
63					//	a*							?			?
66		E4	N	//												
64		N	//	t												
62		//	t	t												
61	//												?			
	61	62	64	66	63	65	75	76	67	74	68	79	71	72	73	78

Notations:

a,e,t: high correlation values (more than 0.8) calculated on the basis of the reduced matrices A,E,T
a*: correlation value is 0.52
? : no significant correlation found

Y : 'see also' reference from the ICP
N : no corresponding cross-reference in the ICP
Ei: possible explanation of qualitative links between CMP subfields(in brackets) on micro-level:

E1: (75) properties - effects (76)
E2: (71) models - phenomena (78)
E3: (78) processes - effects (72)
E4: (62) properties - processes (66)

by using another measure of distance: the angle between two variables. In this case the average similarity has been computed in a different way, namely as

$$\text{SUM SUM } s_{im} \;/\; \text{SUM SUM } s_{jk} \,.\, \text{SUM SUM } s_{np}, \text{ where}$$

s_{im} = cos (angle between variables i and m), and i, j, k are in the first cluster and m, n, p in the second.

This provides an indirect way to verify (only positively) the structure described by the selected multivariate technique.

The links between the CMP subfields as represented by the cluster analysis have been compared to the cross-references attached by experts to the corresponding headings of the ICP and indicating the existence of some cognitive relatedness between the subfields. In Table 6 are shown the correspondences between the 'see also' references of the ICP (upper-left part of the table) and the 'quantitative' links (down-right). An ideal coincidence would be translated by a symmetric configuration (against the diagonal). In our case this 'mirror effect' is distorted: there are links described by the cluster analysis which have no cross-reference counterparts, e.g. 73/68, 71/72, 67/74. They are designated by the letter 'N' in the upper-left part of Table 6. The reverse is true for the 'qualitative' links 61/71, 63/71, and 63/78: their correlation values on the cluster diagrams are not significant. This absence of relatedness is expressed by a question mark in Table 6.

Discussion

The method proposed in the first part is applicable since a set of specialized journals could be selected from all journals covering the field of CMP. This means that the structure of other fields of similar size (e.g. analytical chemistry, organic chemistry, physical chemistry, immunology, microbiology, neurology, etc.) as well as of disciplines (such as biology, chemistry, physics, etc.) could be analyzed on the macro-level.

Here, only the internal links within the field of CMP have been described and compared to the tree topology established on the basis of qualitative analysis (decisions are taken by a team of experts on the corresponding level). Both representations of links (see Table 6) appear to be complementary and show different advantages. Qualitative links indicate the direction of relatedness, i.e. the source and the receiver subfields (of theoretical conceptions, models, properties of substances, etc.). The latter are not shown in Table 6. Quantitative links (in this case) are more recent and indicate the strength of association between two or more subfields.

Based on the existence of a (not necessarily elaborated) classification scheme, this bibliometric method describes links between size-comparable subdivisions of a traditional field. In this respect its results appear to be different from co-citation maps on higher level of aggregation and need further investigation and interpretation.

Summary and conclusions

A new bibliometric macro-method is proposed as an alternative to co-citation analysis of fields and disciplines. The method describes only an inherent tree topology between (sub)fields defined by a classification scheme. The idea underlying this quantitative analysis (on journal level) is that the co-appearance of articles from different subfields (i.e. with different main classification codes) in a set of specialized journals of the corresponding field may be used to describe and measure the degree of relatedness between these subfields.

If the classification scheme includes already a tree topology on macro-level (determined by teams of experts), then a comparison (or simple validation) could be carried out. This matching of results from quantitative and qualitative analyses shows that both approaches are rather complementary, and reveal different relations and aspects.

Results of the new method on the disciplinary level could be compared with already existing macro co-citation maps (on C4 level) in order to test structure patterns in traditional fields and co-citation clusters.

References

1. H. SMALL, Co-citation in the scientific literature: A new measure of relationship between two documents, *Journal of the American Society for Information Science,* 24 (1973) 265–269.
2. D. SULLIVAN, D. H. WHITE, E. BARBONI, Co-citation analyses of science: An evaluation, *Social Studies of Science,* 7 (1977) 223–240.
3. H. G. SMALL, B. C. GRIFFITH, The structure of scientific literatures 1: Identifying and graphing specialties, *Science Studies,* 4 (1974) 17–40.
4. E. NOMA, Co-citation analysis and the invisible college, *Journal of the American Society for Information Science,* 35 (1984) 29–33.
5. A. PICKERING, E. NADEL, Charm revisited: A quantitative analysis of the HEP literature, *Social Studies of Science,* 17 (1987) 87–113.
6. D. HICKS, Limitations of co-citation analysis as a tool for science policy, *Social Studies of Science,* 17 (1987) 295–316.
7. H. SMALL, E. SWEENEY, Clustering the Science Citation Index using co-citations I. Comparison of methods, *Scientometrics,* 7 (1985) 391–409.
8. W. M. SHAW, Jr., Critical thresholds in co-citation graphs. *Journal of the American Society for Information Science,* 36 (1985) 38–43.
9. H. SMALL, E. GARFIELD, The geography of science: Disciplinary and national mappings, *Journal of Information Science,* 11 (1985) 147–159.
10. A. RIP, J.-P. COURTIAL, Co-word maps of biotechnology: An example of cognitive scientometrics, *Scientometrics,* 6 (1984) 381–400.
11. M. CALLON, J. LAW, A. RIP (Eds), *Mapping the Dynamics of Science and Technology,* London. Macmillan, 1986.
12. L. LEYDESDORFF, Various methods for the mapping of science, *Scientometrics,* 11 (1987) 295–324.
13. H. SMALL, E. SWEENEY, E. GREENLEE, Clustering the Science Citation Index using co-citations II. Mapping science, *Scientometrics,* 8 (1985) 321–340.
14. H. SMALL, E. GARFIELD, op. cit., Ref. 9.
15. R. J. W. TIJSSEN, J. De LEEUW, A. F. J. Van RAAN, Exploring Quantitative Relations within a Set of Scientific Entitites. First International Conference on Bibliometrics, August 24–28, 1987, Diepenbeek, Belgium.
16. A. D. PRATT, A measure of class concentration in bibliometrics, *Journal of the American Society for Information Science,* 28 (1977) 285–292.

17. A. SCHUBERT, W. GLAENZEL, Statistical reliability of comparisons based on the citation impact of scientific publications, *Scientometrics*, 5 (1983) 59–74.
18. J. HARTIGAN, Cluster analysis of variables, in: W. J. DIXON, M. B. BROWN (Eds), *Biomedical Computer Programs*, P-Series 1979, Univ. California Press, Los Angeles, CA, 1979. 623–632.
19. A. BERTHELOT, P. CLAGUE, S. SCHIMINOVICH, W. ZWIRNER, The ICSU AB International Classification System for physics: Its history and future, *Journal of the American Society for Information Science*, 30 (1979) 343–352
20. H. SMALL, E. GARFIELD, Op. cit., Ref. 9, p. 150.
21. H. SMALL, E. SWEENEY, E. GREENLEE, Op. cit., Ref. 13, p. 335.
22. D. HICKS, Op. cit., Ref. 6, p. 296.
23. H. SMALL, E. GARFIELD, Op. cit., Ref. 9, p. 159.
24. A. D. PRATT, Op. cit., Ref. 16.
25. J. HARTIGAN, Op. cit., Ref. 18.