Information Processing & Management, Vol. 13, pp. 277–288. Pergamon Press 1977. Printed in Great Britain

JOURNAL CLUSTERING USING A BIBLIOGRAPHIC COUPLING METHOD

HENRY G. SMALL and MICHAEL E. D. KOENIG
Institute for Scientific Information, 325 Chestnut Street, Philadelphia, PA 19106, U.S.A.

(*Received 7 April 1977*)

Abstract—The classification of journal titles into fields or specialties is a problem of practical importance in library and information science. An algorithm is described which accomplishes such a classification using the single-link clustering technique and a novel application of the method of bibliographic coupling. The novelty consists in the use of two-step bibliographic coupling linkages, rather than the usual one-step linkages. This modification of the similarity measure leads to a marked improvement in the performance of single-link clustering in the formation of field or specialty clusters of journals. Results of an experiment using this algorithm are reported which grouped 890 journals into 168 clusters. This scope is an improvement of nearly an order of magnitude over previous journal clustering experiments. The results are evaluated by comparison with an independently derived manual classification of the same journal set. The generally good agreement indicates that this method of journal clustering will have significant practical utility for journal classification.

INTRODUCTION

The concept of algorithmically clustering or categorizing journals has aroused the interest of many members of the information science community. As CARPENTER and NARIN[1] point out, most work in the area seems to have been motivated by a combination of aesthetic and practical considerations. The aesthetic considerations include the challenge of doing algorithmically what has been a very non-trivial task intellectually—the classification of journals. The task is an almost pure problem in numerical taxonomy, that of partitioning a population on the basis of shared characteristics.

On the practical side the outcome of journal clustering can have various applications. The categories reveal the pattern, the mosaic of scholarly activity. An analysis over time would reveal shifts in that pattern, as journals entered or departed from clusters, and as clusters themselves emerged, merged, separated and disappeared. Such observations would have relevance for sociology, information science, and science policy. Clusters thus derived could also be used to analyze and promote the rationalization of journal coverage by secondary services. The DISISS (Design of Information Systems in the Social Sciences) project has proposed such an application[2]. Furthermore, journal cluster patterns would be useful for analyzing and validating thesauri, classification schemes, and indexing schemes.

A number of previous studies have described attempts to cluster journals. In their seminal work of 1967 XHIGNESSE and OSGOOD[3] examined the journal-to-journal citation patterns within a group of 21 psychology journals to obtain a similarity matrix. This was accomplished by means of Shepard's algorithm[4] which assigns distances between journals in n-dimensional space, keeping n as small as possible while preserving the rank orders of citation frequencies between journals. Nine of the 21 journals were assigned to three overlapping clusters, determined by the journals' proximity to each other in n-space. The multidimensionality of this approach limits it to relatively small numbers of journals.

PARKER, PAISLEY and GARRETT[5], later in 1967, undertook an analysis of 17 journals in the field of communication research. The measure of relatedness between journals was a form of co-citation—the frequency of co-occurrence of citations to journals within articles in the 17 source journals. (The term co-citation, more recently introduced[6], refers to a measure of relatedness between articles, defined as the frequency with which two articles are cited together by other articles.) Some 68 journals were cited frequently enough to be analyzed, of which approx. 30–35 were grouped into some 8–11 clusters (the exact number varies for each of the four time periods studied). A criticism as pointed out in the DISISS study described below is the lack of any attempt to normalize for the level of citations. Without normalization the procedure almost inevitably links highly cited journals. The technique is, however, capable of

providing "affiliates" as well as "members" for each cluster, but without normalization, the affiliates tend to be the most highly cited journals that are members of the most strongly linked clusters.

Large scale attempts at clustering journals using citation relationships were not possible until the advent of the *Science Citation Index®* (*SCI*) database (compiled by the Institute for Scientific Information). Particularly important was Garfield's reformating of the *SCI* to show journal to journal citation patterns [7] which revealed the existence of very strong direct citation linkages among journals. This work culminated in the publication of *Journal Citation Reports®* [8] which is an index of these journal to journal citation patterns. CARPENTER and NARIN [1] used these data to look at three disciplines: physics, chemistry and molecular biology. For each discipline the individual journals were manually pre-selected, and a separate journal-to-journal citation matrix was prepared. A "hill climbing" algorithm was used which for each attempt requires the number of clusters to be predetermined, as the algorithm creates no new clusters and rarely eliminates any. A measure of cluster quality is then used to determine which level of clusters has the "best" fit. In this study, nine different combinations of journal similarity measures and cluster quality measures were used and then combined to produce the final results. Each of the three disciplines, ranging in size from 81 to 106 journals, was clustered into 11 or 12 clusters, with 5–16 journals remaining unclustered, and the clusters produced had a high degree of face validity.

A pilot study to explore the feasibility of clustering social science journals was undertaken by the DISISS (Design of Information Systems in the Social Sciences) project at the University of Bath in the U.K. in the early 1970s [2, 9]. Citation data were obtained from 17 source journals. Again, a journal-to-journal citing matrix was used as the basic data form. The clustering algorithm called SCICON, operates on the basis of calculating the root mean square distance from members in n-dimensional space (n being the number of variables, in this case the 17 citing journals) to the center of gravity of each cluster and uses a "run-in" technique of starting with a large number of clusters and then reducing the number one at a time, examining at each step whether a better fit is accomplished by moving any journal to another cluster. The result of this technique used on 115 cited journals was three clusters: psychology (34 members), economics (21 members) and amorphous (60 members). Many of the smaller clusters produced during the run-in, when the number of clusters was higher, were meaningful however.

The work described above, although useful and frequently imaginative, has been limited in its scope. The largest number of journals clustered at one time is barely more than a hundred—a very small portion of the universe of journals. The constraint on size appears to originate not from the lack of data, but from the sheer impracticality of processing the matrices and multidimensional arrays inherent in the techniques used, when any significant number of journals is to be considered.

METHOD

The procedure used in this experiment is a novel combination of some standard methods known to bibliometricians and numerical taxonomists. First, we use the well known technique of bibliographic coupling to derive the basic journal-to-journal associations [10]. Co-citation could equally well have been used as bibliographic coupling, but for computational reasons, bibliographic coupling was the more convenient association measure. For our purposes, bibliographic coupling is defined as the citing of the same document by two journals. (Conventionally, bibliographic coupling is defined as the citing of the same document by two later documents.) The strength of bibliographic coupling (BC) is the number of identical, distinct documents cited by the two journals. This strength of coupling is normalized to compensate for the size effects of the two journals by dividing the bibliographic coupling strength by the sum of the number of references made by the two journals.

The second procedure used is single-link clustering. This mode of clustering has been described elsewhere [11]. We have used the fact that single-link clustering is equivalent to the application of a threshold on the item-to-item proximity measure. In our experiment, the method of single-link clustering was implemented in the following way: A file of journal-to-journal pairs with their appropriate coefficients of association is used as input. A threshold value of the journal-to-journal association is set and a journal is selected as a starting point. All

journals linked to this selected journal at or above the prescribed level of association are located in the file and assigned to the cluster. Each of these journals is then used as a starting point and all journals linked to them are assigned to the cluster. The cluster is complete when no "new" journals can be added to the cluster.

The program then proceeds to the next, unclustered journal, and attempts to create another cluster. After all journals have been examined and assigned to clusters, the program terminates. The smallest cluster created using our procedure is a two member (two journal) cluster since journals not linked to at least one other journal at the prescribed level of association are not searched. The clusters are created at a particular level of the journal-to-journal association and we have no way of knowing what level is "optimum" except by inspection of the results and comparison with results obtained using other procedures. In general, a level is sought in which no very large cluster exists (greater than 100 journals), realizing that such a level, while appropriate for some areas or disciplines, may not be appropriate for others.

A novel feature of our journal clustering system is the use of paths of "length two" between journals to determine the basic association measure used in clustering. Before we define what we mean by this, we can clarify our motivation by describing an earlier experiment which was not successful. We began with the file of journal pairs which were linked by normalized bibliographic coupling (NBC) described above. We then set a minimum threshold for NBC and extracted all journal pairs at or above this threshold.

This gave a file of "strong" journal-to-journal linkages. The problem which we encountered was that we could not obtain a satisfactory set of single-link clusters using the NBC measure. The journals tended to chain together forming very large and loosely linked clusters. It is well known that the single-link algorithm has a tendency to form clusters of this kind, and this tendency, combined with the strongly interdisciplinary character of journal relationships, created enormous chains of journals which resisted fragmentation when the level of NBC was raised. Eventually, when the journals finally did break up into reasonably small clusters at a very high level of NBC, too few of the journals remained in the clusters to consider the experiment a success.

As a result of this experience, we decided to modify our basic journal-to-journal measure. We had noticed that the chaining of journals to create gigantic clusters in the previous experiment was very often due to only a few links from a large or strongly interdisciplinary journal linking one journal "clump" to another. Our problem, then, was to enhance the "clumpiness" of the network so that inter-clump linkages could be "submerged" below some threshold value.

The method we chose was to determine the number of paths of "length two" between journals. For example, suppose we take some arbitrary starting journal. It is linked with a number of other journals with an NBC strength at or above some threshold. These journals are, in turn, linked to other journals at or above this threshold. Now we select a second arbitrary journal and find all the distinct paths which lead from it to the starting journal but which pass through other (third) journals as intermediate steps. These are the paths of "length two" between the two journals. For every pair of journals, then, there is some number of two step paths (including zero) which connect them. It is also clear that the number of such paths for any pair of journals is limited to the lesser number of paths of "length one" which originate from one or the other journal. For example, if journal A has five links to other journals and journal B has ten links, the number of two-step paths leading from A to B cannot exceed five. Hence, we can normalize the two-step paths as shown in Fig. 1. This normalization provides a new measure of journal-to-journal association (normalized two-step bibliographic coupling: NTSBC) which has the property of varying from zero to one.

It is also easy to see intuitively why this should enhance the linkages between journals in a "clump" and thus provide a better clustering than was obtained with the simple NBC. Suppose we have two clumps which are joined by only a few links. The number of two-step paths between journals within a clump will be high, while the number of two-step paths between journals in different clumps will be low. Hence, when a threshold on the two-step linkage measure is applied, the within-clump ties will remain and the between-clump ties will tend to be broken.

It should be noted that there was a direct connection (a one-step path) between J_1 and J_2 in

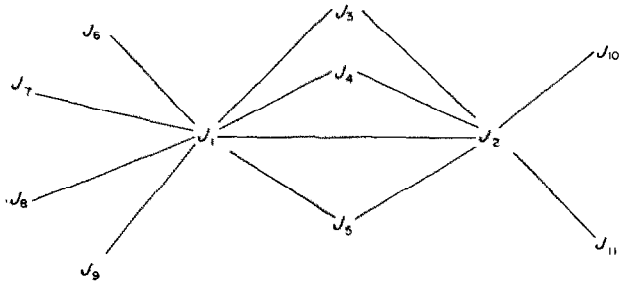

Fig. 1. Illustration of normalized two-step bibliographic coupling. Journals J_1 and J_2 are linked by three two-step paths. J_1 has a total of eight one-step paths leading from it and J_2 has a total of six. The normalized two-step bibliographic coupling (NTSBC) is calculated as follows:

$$\text{NTSBC} = \frac{\text{No. two-step 1–2}}{\text{No. one-step 1} + \text{No. one-step 2} - \text{No. two-step 1–2}}$$

$$= \frac{3}{8+6-3} = 0.273.$$

Fig. 1. The inclusion of this direct link actually weakens the normalized measure from what it would be if the link did not exist. It does so by making the denominator of the NTSBC formula larger. In other words, the strength of linkage between two journals connected by some number of two-step paths will be less if there is a one-step path between the two journals than if there is not. This seeming contradiction could be easily removed if we adopt the simple rule that every one-step path counts as one two-step path in our calculation of NTSBC. It is unlikely, however, that this refinement will have much impact on the results of the clustering since all directly linked journals experience the same disadvantage and few journal pairs having frequent two-step links fail to be directly linked as well. In any event, we do not want to give undue weight to the one-step paths since they are responsible for the chaining effects observed earlier.

Let us now review the method. We begin with an annual *Science Citation Index* and determine the bibliographic coupling strength for all pairs of source journals in this file. This BC strength is normalized by dividing by the sum of the number of references made by each journal during the year in question. A threshold is set on the normalized bibliographic coupling (NBC) and all journal pairs satisfying this threshold are selected. With this restricted file, the number of two-step paths between all pairs of journals is determined. This number is normalized by dividing by the sum of the number of one-step paths emanating from each journal minus the number of two-step paths (see Fig. 1). The normalized two-step bibliographic coupling (NTSBC) is used as input to the single-link clustering routine. Clusters of journals are obtained at a specified, but arbitrary level of the NTSBC.

CLUSTER FILE STATISTICS

Before discussing the specific journal clusters obtained, we will describe the statistical characteristics of the initial and intermediate files (see Table 1). As noted above, an annual *SCI* cumulation is used as the database, which in this experiment was the 1974 file (items 1 and 2 in Table 1). From this file we created a special file listing each document cited by two or more distinct source journals, and the journals citing it. If a document is cited more than once by a certain journal, it is nevertheless counted as though it were only a single citation. This reduces the number of records in the file by about 50% (item 3). (The documents cited by only one journal are dropped since they do not contribute to BC.)

The next step is to form all combinations of source journals which cite a given document, i.e. form all the bibliographic couplings in the file. There were almost seven million such couplings (item 4), which reduces to about 400,000 distinct pairs when identical journal pairs are gathered together and all pairs occurring only once are dropped. Each journal pair with its attached BC strength is then normalized by dividing by the sum of the number of references made by the pair of journals during 1974. A threshold of 0.01 was set on this NBC to eliminate weak linkages between journals (items 7 and 8). It is on this reduced file that the two-step paths are determined. This is done by forming pairs of journals which are linked to a common journal

(item 9). (This step is facilitated by the presence in the file of journal pairs in both the "forward" and "backward" versions, e.g. both AB and BA appear.) Again, identical pairs are gathered together and the frequency of two-step paths is attached to the pairs (items 10 and 11). The second normalization (according to Fig. 1) is carried out, and these data are input to the single-link clustering program. A threshold of 0.4 on the NTSBC resulted in 168 clusters containing a total of 890 journals, with an average cluster size of 5.3 journals per cluster. (The minimum cluster size is two journals and the largest cluster obtained at this level contains 96 journals.)

We contrast this clustering outcome with one obtained in our previous unsuccessful attempt using NBC directly as input to single-link clustering. For a threshold of 0.025 NBC, which

Table 1. File statistics for journal clustering

1. 1974 *SCI* source journal with references	2376
2. 1974 *SCI* citations	5,168,119
3. Citations to documents cited two or more times by distinct source journals	2,478,207
4. Source journal pairs (bibliographic couplings)	6,839,380
5. Distinct source journal pairs	705,167
6. Journals in pairs at BC strength greater than 1	2359
7. Distinct journal pairs at NBC greater than 0.01	8044
8. Journals in pairs at NBC greater than 0.01	1679
9. Total two-step paths between journals	159,171
10. Distinct journal pairs connected by two-step paths	45,180
11. Journals linked by two-step paths	1586
12. Distinct journal pairs at 0.4 NTSBC	2071
13. Journals clustered at 0.4 NTSBC	890
14. Clusters formed at 0.4 NTSBC	168
15. Mean journals per cluster at 0.4 NTSBC	5.3
16. Journals in largest cluster at 0.4 NTSBC	96

represented the most successful NBC results obtained, 119 clusters resulted containing a total of 747 journals, with an average cluster size of 6.3 journals per cluster. This larger mean cluster size was due to the largest cluster which contained 297 journals, constituting nearly 40% of the journals clustered. By contrast, for the clusters obtained at 0.4 NTSBC, the largest cluster of 96 journals constituted only about 11% of the journals clustered. It is clear, then, that by using a two-step linkage measure the degree of chaining has been substantially reduced and the "clumpiness" of the journal network increased.

Other clustering levels of the NTSBC were also tried and it appears that the critical level at which a transition occurs from a highly chained and enormous cluster to a group of subject or discipline oriented clusters is between 0.2 and 0.3 NTSBC. At 0.2 there were only 40 clusters with the largest cluster containing 1276 journals, nearly the entire journal set. At 0.3 NTSBC a radical change occurred. We obtained 153 clusters with the largest cluster containing 360 journals. At level 0.4 we have increased the number of clusters by only 15 but the largest cluster declined in size nearly 75%.

The existence of a "critical point" in the clustering level where there is a sudden breaking up of the largest cluster is also found in experiments clustering highly cited documents rather than journals[12]. Whatever this may mean, it is clear that no one level of clustering is optimal for all scientific fields or specialties. Ideally one should adopt a variable level approach to seek out the best possible representation for a given area by varying the level up or down. This means that a way must be found of evaluating the quality of a cluster that is independent of the clustering methodology. This is a familiar situation in cluster analysis since it is generally recognized that adequate tests of cluster significance have not yet been developed and reliance on other means for evaluating results is necessary (e.g. their utility or agreement with classifications derived by other means). In the discussion of the clusters at level 0.4 NTSBC, which follows, we use two modes of "validating" the results. First, the classification obtained automatically is compared with one which was obtained manually and quite independently. Second, qualitative evaluations of some of the groupings of journals based on our understanding of the current state of the scientific subject matters involved are made.

EVALUATION OF JOURNAL CLUSTERS

We selected the level 0.4 NTSBC clusters for detailed examination because the largest cluster at this level contained 96 journals and was not so large as to suggest a completely meaningless journal grouping (the distribution of cluster size was the least skewed at this level). As we pointed out, as the level is raised, the size of the largest cluster decreases dramatically. This transition from macro-clusters to micro-clusters probably corresponds to the point at which an interdisciplinary chaining of journals breaks up and disciplinary or specialty groupings are formed. The cluster containing 96 journals seems to be such a disciplinary grouping in the biomedical field centered around cancer research. A closer look at the clusters obtained at the 0.4 level will provide some idea of what can be expected at other levels, but not too much significance should be placed on this particular level.

Of the 168 clusters obtained at this level, 79 contained three or more journals, and 89 clusters contained only two journals each. (Clusters of one journal do not emerge from our clustering procedure because the basic input record is the journal pair.) The 89 two-journal clusters are not considered in the following discussion, but we should attempt to explain their significance. Like other bibliometric data, the distribution of cluster sizes (that is, the number of clusters containing two journals, three journals, four journals, etc.) is very skewed and approximately hyperbolic. This is true at any clustering level selected except the very lowest where all journals are in a single gigantic cluster. Thus, there are many small clusters and few large ones (21 clusters have 10 or more journals and 58 have from three to nine journals at level 0.4). At lower levels of clustering, the small clusters may join up with one another or with a larger cluster. There are three possible interpretations of very small clusters: they may be genuinely isolated groupings; they may be tips of larger groupings which emerge at lower clustering levels; or they may be fragments of larger clusters which join up with the larger clusters at lower levels.

For the purpose of comparison with the 79 clusters containing three or more journals at level 0.4, an independently derived journal classification was used. This classification appears as Table A-3 in Narin's *Evaluative Bibliometrics* [13]. The Table lists Narin and co-workers' manual classification of the 1973 source journal list of the *Science Citation Index*. Roughly 2000 journals are classified into nine major headings (fields) and 106 subheadings (subfields). Two points should be noted about the comparison of our journal clusters with Narin's manual classification. First, Narin has classified nearly all source journals in the 1973 source list (over 2000 journals), while our clusters at level 0.4 with three or more journals comprise only 701 journals. Hence, we would not expect to find all journals Narin includes under a subheading in our clusters. Second, since our clustering experiment was done using the 1974 *Science Citation Index*, a few journals which were dropped or added to the *Science Citation Index* coverage since 1973 do not match up when Narin's classification is compared with our clusters. The number of such cases is, however, small in relation to the number of journals in either file.

One way of evaluating the match between these two classifications is to count the number of journals shared by one of our clusters and the Narin subheading to which it is most strongly related. This overlap is expressed as a fraction of the number of journals in *our* cluster and not as a fraction of the number of journals in the related Narin subheading because of the greater comprehensivity of the latter. This measure reflects, in effect, the dispersion of our cluster over Narin's subheadings. For example, one of our clusters may contain journals which Narin has placed into several different subheadings, although usually there will be a single subheading which has the greatest overlap with our cluster. The fraction of journals in this cluster which falls in the single most closely related subheading will measure the degree to which the cluster's journals are dispersed over Narin's subheadings: the larger the fraction, the less the dispersion.

Figure 2 shows the distribution of fractions of shared or overlapping journals for each of the 79 clusters with the Narin subheading with which it has the largest overlap. Six small clusters (each having 3 journals) have fractions equal to zero because they did not match with any of Narin's subheadings. Most of the journals in these clusters were not classified by Narin because they were newly added to the *Science Citation Index* coverage in 1974. The figure shows that 22 of the 79 clusters had fractional overlaps with Narin's subheadings of from 0.91 to 1.0. Of the 22 clusters, 20 were perfect matches, i.e. all journals in the cluster appear under a single

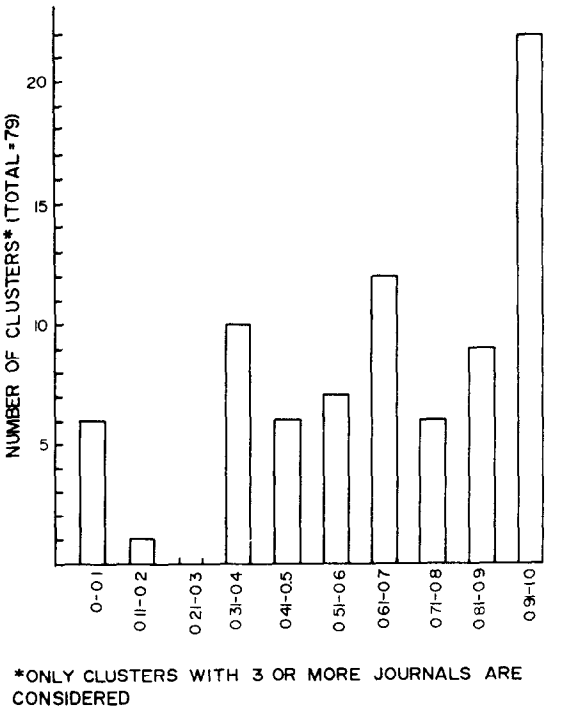

Fig. 2. Fractional overlaps between journal clusters and subheadings in Narin's classification scheme.

subheading. Thirty-seven (47%) clusters had fractional overlaps of 0.75 or better, and 56 (71%) had overlaps of 0.5 or better.

An example of a "good" match between a cluster and a subheading in the manual classification is cluster No. 1, a group of 10 journals (see Table 2). The fact that this group of journals comes out as cluster No. 1 has no significance except that the first journal in the cluster, A GRAEFES A, is early in the alphabet and its pairs were the first to be selected by the computer. (ISI standard 11 character journal abbreviations are used throughout. For full titles see Ref. 14.) The subheading in Narin's scheme which matches this cluster is titled "ophthalmology" and contains 18 journals. All 10 of the journals in cluster No. 1 appear under Narin's "ophthalmology" subheading. Narin's scheme lists eight additional journals which do not appear in cluster No. 1 or among any of the other clusters at level 0.4 NTSBC. It remains to be

Table 2. Match between cluster No. 1 and Narin's "ophthalmology" subheading

Cluster No. 1	"ophthalmology"
A GRAEFES A	A GRAEFES A
ACT OPHTH K	ACT OPHTH K
AM J OPHTH	ADV OPHTHAL
ARCH OPHTH	AM J OPHTH
BR J OPHTH	AM J OPTOM
CAN J OPHTH	ARCH OPHTAL
EXP EYE RES	ARCH OPHTH
INV OPHTH	ARCH S A OF
KLIN MONATS	BR J OPHTH
OPHTHALMOLA	BR J PHYS O
	CAN J OPHTH
	DOC OPHTHAL
	EXP EYE RES
	EYE EAR NOS
	INV OPHTH
	KLIN MONATS
	OPHTHAL RES
	OPHTHALMOLA

determined whether lowering the various levels used in creating the clustered file would result in adding these journals to the cluster.

Cluster No. 19 which contains 10 journals is an example of a cluster which matches with more than one of Narin's subheadings. This cluster corresponds to two of Narin's subheadings, one titled "obstetrics and gynecology" containing 12 journals and another called "fertility" containing five journals (see Table 3). Eight of cluster No. 19's 10 journals overlap with the "obstetrics and gynecology" subheading, and two of the cluster's journals overlap with the "fertility" subheading. This is an instance where the manual classification and the citation-based clustering disagree on how journals should be grouped. Despite the fact that this cluster is not a "good" match to a particular subheading, it does have a high face validity. The clustering suggests that due to the commonality of the literature cited, the "obstetrics and gynecology" journals should perhaps be merged with the "fertility" group.

To see how the cluster grouped these journals, actual linkages among the ten journals were drawn (Fig. 3). We see that of the two journals Narin placed in the "fertility" subheading one, FERT STERIL, was linked to the group only through CONTRACEPT, which was the other journal placed in the "fertility" subheading. CONTRACEPT, on the other hand, was strongly linked to the remainder of the cluster which Narin had classified under "obstetrics and gynecology."

An example of a "not-so-good" match with Narin's scheme is cluster No. 5 which contains 19 journals. As shown in Table 4, cluster No. 5 contains journals which appear in six of Narin's

Table 3. Match between cluster No. 19 and Narin's "obstetrics and gynecology" and "fertility" subheadings

Cluster No. 19	"obstetrics and gynecology"
ACT OBST SC	ACT OBST SC
AM J OBST G	ADV OBSTET
AUST NZ J O	AM J OBST G
CONTRACEPT	ARCH GYNAK
FERT STERIL	AUST NZ J O
GYNAKOLOGE	FORTSC GEB
J OBSTET GY	GYNAKOLOGE
J REPRO MED	GYNECOL INV
OBSTET GYN	J OBSTET GY
REV F GY OB	J REPRO MED
	OBSTET GYN
	REV F GY OB
	"fertility"
	BIOL REPROD
	CONTRACEPT
	FERT STERIL
	INT J FERT
	J REPR FERT

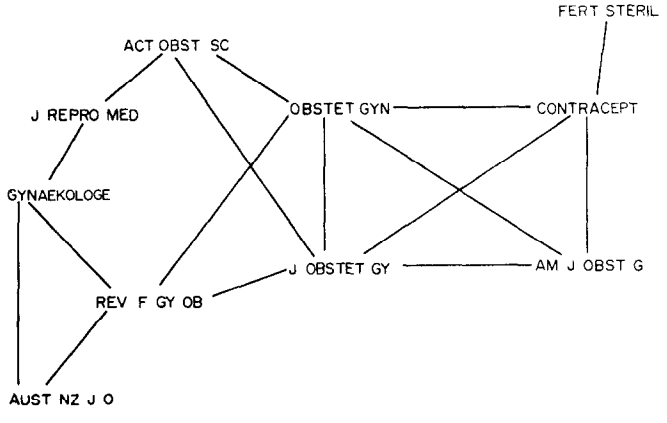

Fig. 3. Cluster No. 19 at level 0.4 NTSBC.

Table 4. Match between cluster No. 5 and Narin's subheadings

Subheadings	Cluster No. 5 (19)
Cell Biology, Cytology and Histology (28)	CELL TIS RE
	CYTOBIOLOG
	CYTOBIOS
	HISTOCHEM J
	HISTOCHEMIS
	J CELL SCI
	J HIST CYTO
	J ULTRA RES
	TISSUE CELL
	Z ZELL MIKR
Anatomy and Morphology (9)	ACT ANATOM
	AM J ANAT
	ANAT REC
	J ANAT
	J MORPH
Microscopy (8)	J MICROSCOP
Neurology and Neurosurgery (45)	J NEUROCYT
General Biomedical Research (82)	PHI T ROY B
Embryology (8)	Z ANAT ENTW

subheadings. The cluster has the largest overlap with the subheading titled "cell biology, cytology and histology", which includes 10 of the cluster's 19 journals. Another subheading, "anatomy and morphology" contains five of the cluster's journals, and four additional subheadings each contain one of the cluster's journals. The most notable feature of the cluster is that it joins together the classical fields of anatomy and morphology with the more modern fields of cell biology, cytology and histology. The cluster suggests that the distinction between these subject areas may be breaking down, and that the modern study of anatomy and morphology is concerned with structure at the microscopic and cellular level.

Based on our comparison of clusters and subheadings, we can conjecture on the significance of journal groups formed on commonality of cited literature and how this "classification" differs from one derived "intellectually". It appears that journal clusters reflect current research interests and research front activity since, in general, the literature cited by articles in a journal is a function of the state of development of the various specialties represented by those articles. At the research front citation patterns will in some cases link together journals from historically separate fields. This inter-relatedness at the research front might not be obvious to someone who was charged with classifying journals on the basis of their subject scope, unless that person possessed a detailed knowledge of each of the fields. Of course, as funding patterns change and new fields of knowledge open up, these emphases could change and the journal cluster could be dissolved or transformed. At present, we have no evidence for such shifts, since we have examined only one year of data (1974). However, it seems reasonable to expect changes in journal cluster membership over time. We expect that this analysis of cluster shift will be a fruitful area for further research.

The clusters at level 0.4 NTSBC also reflect both disciplinary aggregations and very small, highly specific groupings. An example of the former is cluster No. 25 (see Table 5) which consists of 35 journals in various fields of physics and corresponds to five of Narin's physics subheadings (solid state, applied physics, general physics, nuclear physics and miscellaneous physics). In this case, we are probably not dealing with a research front aggregate but rather a disciplinary aggregate. In order to break up this disciplinary group into its subdisciplinary units it would be necessary to raise the NTSBC clustering threshold.

At the other extreme in size, we have cluster No. 164 which contains four journals on the technology of paper making (Table 6). There is no question that these journals belong together, but to see where they belong in a larger disciplinary framework, requires that we recluster at a lower level of NTSBC until these journals link with others, perhaps on the science of materials. The point here is that clusters at this or any level must be seen as a composite of hierarchical levels, with some areas of science aggregated at the field or disciplinary level and others at the specialty or subdisciplinary level. This means simply that to derive maximum benefit from this

Table 5. Match between cluster No. 25 and Narin's subheadings

Subheadings	Cluster No. 25 (35)
General Physics (114)	ACT PHYS AU
	ADV PHYSICS
	ANN PHYSICS
	ANN R NUCL
	CAN J PHYS
	FORTSCHR PH
	J PHYS A
	J PHYS JAP
	LETT NOUV C
	NOUV CIM A
	NOUV CIM B
	PHYS LETT B
	PHYS REV A
	PHYS REV L
	PHYSICA
	PROG T PHYS
	REV M PHYS
	SOV J NUC R
	Z PHYS
Solid State Physics (9)	J PHYS C
	J PHYS CH S
	PHYS LETT A
	PHYS REV B
	PHYS ST S-B
	SOL ST COMM
Nuclear and Particle Physics (6)	NUCL PHYS A
	NUCL PHYS B
	PHYS REV C
	PHYS REV D
	USP FIZ NAU
Applied Physics (36)	J L TEMP PH
	J PHYS F
Miscellaneous Physics (6)	J MATH PHYS
	FIZ TVERD T*
	ZH EKSP TEO*

*Journal not listed in Narin's Classification.

Table 6. Cluster No. 164: paper technology

PAP PUU
PULP PAPER
SVENS PAP T
TAPPI

clustering technique we need to adopt a variable threshold approach to the creation of clusters and select appropriate thresholds for each field.

CONCLUSIONS

The results of the present experiment hold out some hope that clustering will prove to be not only an exercise of theoretical interest, but a practical method for organizing journal sets. In fact, the results have already been used to assist in updating ISI's classification scheme. Our experiment was modest (890 journals and 168 clusters containing two or more journals) compared to the number of source journals covered by the *SCI* (2400) which are potentially available for clustering. There are no indications that a scaling-up of the experiment to include all of the source journals would not be feasible. Experience with the same clustering algorithm (single-link) applied to highly cited documents rather than journals has shown that the algorithm is capable of handling several thousand objects efficiently[12].

By comparison to earlier clustering experiments involving journals, however, our experiment was on a rather large scale. By far the largest previous experiment was that reported by Carpenter and Narin which involved three sets of approx. 100 journals each. The principal drawback of their approach is that the outcome of clustering is order-dependent. This means that the precise make-up of a cluster depends on the order in which the journals are added to it. One advantage of our method is that it is order-independent: the clusters contain the same journals regardless of the order in which they are added.

The principal disadvantage of our method is the tendency for single-link clusters to chain. The important progress we have made in this regard is to perform, in effect, a squaring of the "adjacency matrix", which as is well known, converts one-step into two-step paths. This results

in a preferential enhancement of intracluster links where one-step linkage densities are high. This strategy could be described as adapting the proximity measure to fit the data, rather than adapting the clustering algorithm itself. The latter approach would be far more difficult, if not impossible in our case, due to the enormous number of journals to be clustered. Single-link clustering appears to be the only algorithm capable of operating efficiently on data sets with thousands of members.

It is interesting to speculate on what would happen if we went from two-step links to four-step links and so on. The question would be whether this is a convergent or a divergent process, i.e. whether the clusters formed at each successive step-order would be the same as or different from those formed using the previous step-order.

Another means of defeating chaining is to adopt a flexible approach to the clustering level. It appears that fields of science and subfields within them vary widely in linkage density and strength as measured by bibliographic coupling. The causes of such variations are not clear, but changes in the research front of science, the size and growth of fields or sub-fields, the tradition of citation in a field, may all contribute to the observed variations. Whatever the causes, it is clear that clusters at a single level of our proximity measure can range from small, highly specific groupings (paper technology) to large disciplinary groupings (physics). In other words, the level which yields paper technology as a distinct grouping is not the level which yields solid state physics. To obtain the latter, we must raise the clustering level and in doing this, the former may be submerged. The solution is to cluster at several widely spaced levels, and select from among the various levels the version of the cluster which seems most reasonable or useful.

The problem of cluster validity or quality is at this point crucial. As yet we possess no tests of statistical significance for clusters or generally accepted measures of cluster quality. Hence, we have no recourse other than to rely on external criteria such as: (1) Do the clusters make sense?, (2) Do they agree with alternatively derived classifications?, (3) Are they useful for our purposes? In our validation, we have relied mainly on method (2) using an independently derived manual classification of journals. The results were encouraging, suggesting wide areas of agreement between the two classifications, and revealing some interesting differences for which we have offered tentative explanations.

Nevertheless, a difficult problem remains in establishing internal criteria for clusters. For example, at what level should a cluster be considered optimum? There are indications that an operational or heuristic solution to this problem is possible. We want to include as many journals as possible in a cluster without adding groups of journals which are clearly on a different subject. Therefore, the level should be lowered as long as journals are added singly to the clusters. At the point that a cluster of N journals is added (where N is empirically determined) we know we have lowered the level too much, and the optimum level lies just above this. While this is not a formally satisfying criterion of cluster quality, it is a simple rule of thumb which would prove useful, lacking more rigorous tests of cluster quality.

Other problems remain for future research, including the stability of clusters over time, and the application of the techniques developed in this paper to the full source journal coverage of the *SCI*. The ultimate test of the system will be whether it provides a practical alternative or adjunct to current manual journal classification methods. The advantages of an automated system are obvious when we consider the difficulties raised by adding and dropping journals to the database and the sometimes rapid shifts in research fields. If we can establish the viability of automatic *retrospective* journal classification, perhaps we can then take the next step toward a "real time" dynamic classification.

Acknowledgements—We would like to thank Mr. Lou Holmes of ISI for systems and programming of the clustering algorithm, and Mr. Jim Gibson of ISI for assistance in analysis of clusters.

REFERENCES

[1] M. P. CARPENTER and F. NARIN, Clustering of scientific journals. *JASIS* 1973, **24**, 425.

[2] Bath University. Design of Information Systems in the Social Sciences. *The structure of social science literature as shown by citations.* (Research Report A). Bath University (1976).

[3] L. V. XHIGNESSE and C. E. OSGOOD, Bibliographic citation characteristics of the psychological journal network in 1950 and 1960. *Am. Psychol.* 1967, **22**, 778.

[4] R. N. SHEPARD, The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika* 1962, **27**, 125, 219.
[5] E. B. PARKER, W. J. PAISLEY and R. GARRETT, *Bibliographic citation as unobtrusive measures of scientific communication.* Stanford University, Institute for Communication Research, Stanford (1967).
[6] H. SMALL, Co-citation in the scientific literature: a new measure of the relationship between two documents. *JASIS* 1973, **24**, 265.
[7] E. GARFIELD, Citation analysis as a tool in journal evaluation. *Science* 1972, **178**, 471.
[8] E. GARFIELD, *Journal citation reports: a bibliometric analysis of references processed for the* 1974 *Science Citation Index.* (Science Citation Index 1975, Vol. 9). Institute for Scientific Information, Philadelphia (1976).
[9] Bath University. Design of Information Systems in the Social Sciences, *Clustering of journal titles according to citation data: report on preparatory work, design, data collection and preliminary analyses.* (Working paper No. 11). Bath University Library, Bath (1973).
[10] M. M. KESSLER, Bibliographic coupling between scientific papers. *Am. Docum.* 1963, **14**, 10.
[11] P. H. A. SNEATH and R. R. SOKAL, *Numerical taxonomy*, p. 216. W. H. Freeman, San Francisco (1973).
[12] H. G. SMALL and B. C. GRIFFITH, Automatic classification of scientific literature using co-citation clustering. *Proc. 12th Ann. Allerton Conf. on Circuit and System Theory*, p. 512 (1974).
[13] F. NARIN, *Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity.* Computer Horizons, Cherry Hill, NJ (1976). The journal classification referred to in the text as "Narin's Classification" was the work of Dr. Gabriel Pinski of Computer Horizons.
[14] *Science Citation Index* 1974 *guide and journal lists.* Institute for Scientific Information, Philadelphia (1975).