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_i and J_j in

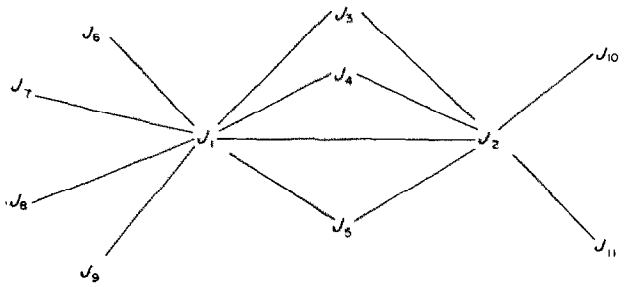

Fig. 1. Illustration of normalized two-step bibliographic coupling. Journals J_1 and J_2 are linked by three two-step paths. J_1 has a total of eight one-step paths leading from it and J_2 has a total of six. The normalized two-step bibliographic coupling (NTSBC) is calculated as follows:

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

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

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.

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.