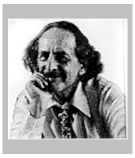
Essays of an Information Scientist, Vol:4, p.634, 1979-80 Current Contents, #40, p.5-12, October 6, 1980

Current Comments



ABCs of Cluster Mapping. Part 1. Most Active Fields in the Life Sciences in 1978

Number 40 October 6, 1980

When the 1978 Lasker award for basic science research was announced, honoring various investigators for their work in opiate receptors, a controversy developed over why several others were not named.[1] At that time, I presented four cluster "maps" of opiate receptor research covering the years 1974-77.[2] The maps not only graphically depicted the explosive growth of this scientific specialty, but they also identified the foundation, or "milestone," papers, when new developments arose, and who was responsible for advances in the field.

The opiate receptor exercise called special attention to ISI®'s clustering procedures, but we've been mapping scientific specialties for over five years now.[3] The opiate receptor cluster is only one of about 2,000 clusters generated in a computer run of a single annual *Science Citation Index® (SCI®)* file. Table 1 lists the 100 largest clusters in the life sciences in 1978. These were the most active fields in terms of papers published that year. Part 2 of this essay will cover the physical sciences.

I make no attempt here to "explain" these listings in cognitive terms. Any regular reader of *Current Contents® (CC®)* will recognize the fields closest to his or her own research. The clusters were named simply by identifying the terms most frequently used in the titles of the papers in each cluster. Thus, the subject matter of the cluster is described in the actual terms currently used by active researchers.

The number of "cited" papers for each cluster is also shown in Table 1. These papers could have been published at any time, but in most cases they are the *recent* milestone papers for the field. For example, of the 124 cited papers in the opiate receptor cluster, 108 were published between 1973-77, 11 between 1968-72, and five before 1968. The average year of publication was 1974. By contrast, 15 of the 24 cited papers in the acetylcholine receptors cluster were published between 1973-77, six between 1968-72, and three before 1968. The average year of publication here is 1972. This average year for cited publications presumably is one indication of how fast the field is moving. Another indicator is when some of the "primordial" papers disappear from the cluster. This is not to say that they are never cited. Rather, they are cited less because they become the common wisdom: formal citation is "obliterated."[4]

To evaluate the utility of Table 1, you could ask several colleagues to name the most active fields of research for 1978. On the other hand, an enterprising journal publisher might consider whether a field like opiate receptors (947 citing papers) is ready for its own journal. The same question can be asked about research on substance P (640), fibronectin (432), somatostatin (411), or sister chromatid exchanges (299). These cluster titles, and a few thousand more, will add to the unique approach to searching ISI's files in the future, both on-line and in print. The latter is to be tested in an almost completed *Atlas of Biochemistry*. The former will be tested soon at a dozen or more research institutions.

Table 1. Top 100 1978 clusters in the physical sciences, ranked by the number of citing articles in each cluster, that is, the number of papers published in that field.

Cluster Number	Name	Cited Articles	Citing Articles
24	Weak Neutral-Current Reactions	49	368
436	Instantons	8	266
62	Spin Glasses	23	244
920	Macrocyclic Complexes	6	212
1,253	Heavy Quark Systems	4	201
1,416	Quantum Chromodynamics: Jet Processes	7	199
119	Large Transverse Momentum Hadron Production	21	189
125	Hadron Collisions & Lepton Pair Production	12	176
297	Clusters of Galaxies	17	168
523	MNDO Studies of Molecules	5	167
201	Charmonium Model	15	162
264	Rare-Gas—Halide Systems	21	156
13	Solitons	7	150
20	Charmed Hadrons	7	141
113	Angle-Resolved Photoemission	13	136
2,074	Molecular Orbital Calculation of Interactions	4	131
171	Electronic States in Amorphous Semiconductors	6	130
1,371	Cycloaddition Reactions	9	128
253	Instanton Solutions for Gauge Field Theory	10	127
209	Photoelectrochemistry: Solar Energy Conversion	20	125
919	Quantum Theory of Solitons	10	125
345	Electronic States of Semiconductor Surfaces	15	124
1,634	Gauge Theories of Gravitation	4	122
227	Heavy Ion Collisions	17	120
307	Organic Alloys	15	120
789	Tau Heavy Lepton Decay Mode	6	112
469	Phase Transitions in 2-Dimensional Systems	6	111
346	Bag Models of Hadrons	4	109
204	Magnetic Monopoles in Gauge Field Theories	6	105
371	Vibrational Relaxation Studies	11	104
735	Renormalization of Quantum Field Theory in Various Space-Times	9	101
127	Scaling Violations & Neutrino Scattering	10	100
379	K-Shell Ionization	6	100
2,024	Charge Density Wave States	4	98
333	Excited-State Electron Transfer in Transition-Metal Complexes	14	96
169	Properties of Amorphous Solids at Low Temperatures	7	96
1,632	Molecular Orbital Structure Studies	5	96
21	Baryonium Model	10	94
324	Planetary Nebulae	7	91
309	Critical Phenomena in Fluids	5	91
549	Multi-Photon-Induced Molecular Dissociation	5	91
240	X-Ray Absorption Fine Structure (EXAFS) Studies	12	90
382	Heavy Ion Fusion	8	90
1,703	High-Performance Liquid Chromatography	8	90
402	Renormalization Group Approach to Critical Dynamics	7	88
25	Asymmetric Reactions Catalyzed by Metal Complexes	9	87
66	Nuclear Proton Scattering	10	86
1,011	Kinetic Model for Classical Liquids	9	86
225	Geochemistry of Volcanics	6	85
718	Pleistocene Paleoclimates	5	85
355	Stellar Evolution & Mass Loss	5	84
226	Molecular & Atomic Low-Energy Scattering	4	83
1,089	Neutron Stars	5	82

626	Microtubule Assembly	16	183
1,449	Nucleosome Structure & Transcriptionally Active Chromatin	4	183
496	Genetics of Diabetes Mellitus	19	181
214	Epidermal Growth Factor	16	180
1,215	Calcium Transport System of Sarcoplasmic Reticulum	22	179
803	Plasmid Replication	18	179
926	Organization of Catecholamine Neurons in the Central Nervous System	10	179
1,115	Thyroid Hormone Receptors	24	176
489	Friend Cell Erythroid Differentiation	20	176
573	Phospholipid Exchange Proteins	16	176
540	Rotavirus in Gastroenteritis	26	175
59	Myelin Basic Proteins	19	175
754	Immunochemistry of Gonococcus	18	175
164	Estrogen Receptors	17	175
139	Computed Cranial Tomography	4	174
810	Regulation of Prolactin Secretion	17	173
53	Adenosine Deaminase Deficiency	18	172
109	Chromatin Fractionation & Functional Organization	14	170
635	Messenger RNA 5' Terminal Cap Structures	14	170
652	Reaction Center of Photosynthetic Bacteria & Plants	23	169
1,019	Inhibitory GABA & Amino Acid Inhibitors in the Central Nervous System	16	169
304	Tumor Promotors	18	168
348	DNA Polymerases	18	165
1,148	Beta-Adrenergic Receptors	12	165
534	Biosynthetic Precursors	11	165
82	Intercellular Junction Formation	13	163
58	Taurine in Central Nervous System	13	162
228	Teratocarcinoma Stem Cells	15	161
1,348	T-Cell Subpopulations	4	161
409	Retinal Photoreceptors	22	158
918	Guanylate Cyclase Activation	18	157
518	Lipoprotein Metabolism	12	157
1,425	Fluorescence Membrane Probes	8	157
705	Structure of Major Histocompatibility Antigens	17	156
716	Suppressor Cell Activity	6	156
4	DNA-Replication Proteins	17	155
753	Vasoactive Intestinal Polypeptide	7	154
47	Pathogenesis of Gallstones	15	153
1,463	Cyclic GMP	7	151
294	Non-Hodgkin Lymphoma	17	149
866	DNA Repair & Mutagenesis	12	148
11	Heparin & Antithrombin	17	147
574	Idiotype in Antibodies	14	147
366	Resistance Plasmids	11	146
317	Localization of Amino Acid Neurotransmitters	10	146
96	Computed Body Tomography	10	144
275	Erythroid Differentiation	10	144
102	Lectins	8	144
1,691	Clinical & Immunologic Aspects of Gonococcal Infection	11	142

From the list in Table 1, it can be seen that ISI's clustering procedure is actually a classification system which groups documents into related fields and subspecialties. In this essay, I'll describe the first steps through which the computer automatically classifies the documents included in an annual *SCI* file. ISI's method is a unique variation on clustering techniques described in classic textbooks.[5-7] It is important to first discuss the basic concepts and definitions of clustering before you can see how ISI's particular methods differ from those in general use.

Cluster analysis has many different applications, but its objective is usually the same—to elucidate the structure underlying complex bodies of data by identifying resemblances between members of their populations.[8] Starting with a population of *cases,* cluster analysis reveals any organizational patterns that arise from similarities between their *variables.* For example, an anthropologist may use cluster analysis to define the social structure of a primitive tribe in terms of how its members (cases) are distributed according to similarities in sex, status, community roles, or marriage bonds (variables).

At ISI, we use cluster analysis to identify patterns in research. The papers covered in the *SCI* are the source documents (cases). These sources cite references (variables). To categorize or group the source (citing) documents, we cluster the works they cite.

Cluster analysis is usually applied to populations having a few hundred cases or variables at the most. But at ISI just one week's data includes over 10,000 articles or book chapters. The 1978 annual *SCI* included more than 500,000 published papers or chapters (source items) containing more than 7.5 million reference citations. In 1979, these figures increased to about 520,000 source items and almost 7.8 million reference citations. Even if we could afford the enormous amount of computer time required to *completely* cluster the *SCI* each year, it would overwhelm us with data. So we take a more selective and pragmatic approach which, in practice, proves to be quite useful and productive.

To be more selective, we set a citation *frequency threshold.* Instead of looking at *all* papers and books cited in a given year, we single out only those cited, say, 17 or more times. This leaves us with a set of about 23,500 highly-cited items, or less than 1% of all the items cited in a single year (Table 2).

Table 2: Citation frequency distribution data for 1978 *SCI*®.

Times cited	Number of items	% of file
1	2,675,936	70
2-4	876,993	23
5-9	199,210	5
10-16	49,741	1
17-25	14,694	.5
26-50	7,163	.3
51-100	1,415	.1
101-over	353	.1
total	3,825,505	100%

The set of most-cited papers, together with the lists of papers that cite them, are sorted to produce a series of indexes. Figure 1 shows the steps involved in preparing these data for clustering. The caption explains the procedure in detail.

A long processing step is required to determine how often each of the 23,500 highly-cited papers is co-cited with one another by the source publications. You can get an idea of how many possible pairs are generated by applying the standard formula 1/2n (n-l), where n = 23,500. Thus, there are almost 280,000,000 possible pairs!

Of course, none of the highly-cited papers actually co-occurs with *all* the other highly-cited papers. So, the actual number of pairs of co-cited papers is much lower than the 280,000,000 calculated above. In fact, more than 99.5% of the possible combinations are "zero-linked pairs"—highly-cited papers that are never actually co-cited with each other. This makes it feasible to program our computer to identify only the "non-zero-linked pairs"—papers that *are* co-cited. Typically, 800,000 co-cited pairs are identified.

You must remember the basic premise—the co-occurrence of two papers in reference lists uniquely identifies subject matter. Our ultimate objective is to group documents linked by such pairs so that they form clusters that identify *fields* of research. Thus, we are trying to go up the hierarchical scale from the more specific to the more generic. Using quantitative criteria we arrive at qualitative statements about current research activity.

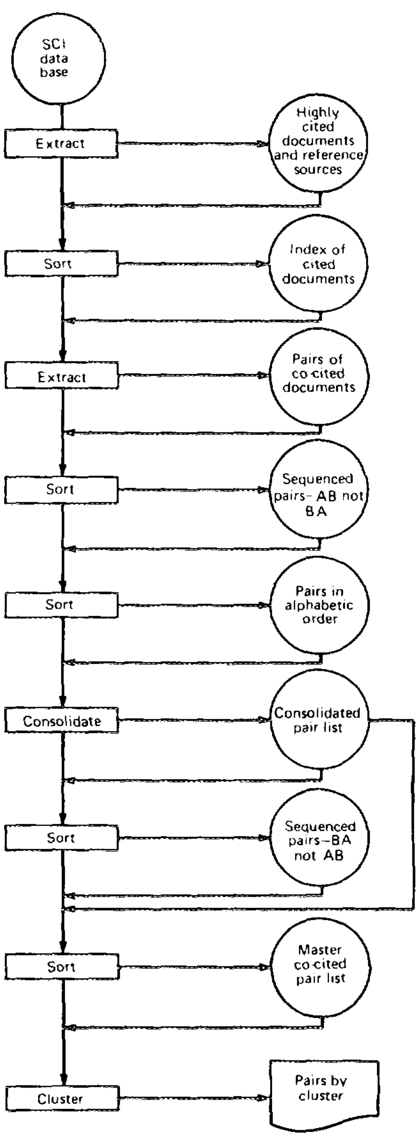
Figure 1: Diagram of ISI®'s clustering method. The initial input is a portion (quarter, semiannual, annual) of the *SCI*® data base. The product of the first processing step is a citation index of highly cited documents, containing the lists of papers that cited them. In the second step, this file is resorted to produce an index arranged by the source (citing) papers. In other words, for each paper that originally entered the file, only its references to highly-cited papers have been retained. In fact, any papers that contained no references to these highly-cited items have been eliminated—about half the file.

For each paper that remains, all possible pairs of highly-cited papers are formed; e.g., if three were cited, A,B,C, then AB, AC, and BC would be formed. If four were cited, AB, AC, AD, BC, BD, and CD would be formed. The result is a list of document pairs in which each pair appears in only one of the two sequences that are possible: as AB, but not as BA. This process is followed for every one of the hundreds of thousands of source papers in a typical annual file. Then this enormous file is sorted again so that all occurrences of each pair are now together.

The number of times each pair appears is counted, the total is recorded next to the first record of the pair, and all duplicate records are deleted. The effect of this is to reduce the number of records that must be processed. The next step is to duplicate the consolidated pair list, but with the document sequence of each pair reversed: AB becomes BA. This doubles the size of the file.

The two pair lists are then combined and put into alphabetical order. The effect of this sort is to produce a master list which batches together all the pairs for a particular cited document. In effect, this creates the master matrix. This matrix will be run through the clustering routine that forms aggregates of highly-cited documents. This is done by grouping together all cited papers that have a pair or linkage "strength" above a given value. The details of this procedure will be explained in Part 2 next week.

(Adapted from: **Garfield E.** *Citation indexing: its theory and application in science, technology, and humanities.* New York: Wiley, 1979. 274 p.)



Before clustering begins, using all of the co-cited pairs, we set a *strength threshold* that can range from 0-100%. "Strength" indicates how related two documents are, in terms of the proportion of their total citations that are co-citations. For example, document A is cited 20 times in one year, document B 50 times, and they are co-cited 10 times. The strength of association between them is calculated by:

$$\frac{\text{co-citations of A and B}}{(\text{total citations A + B}) - \left(\text{co-citations of A + B}\right)}$$

Thus, documents A and B have a

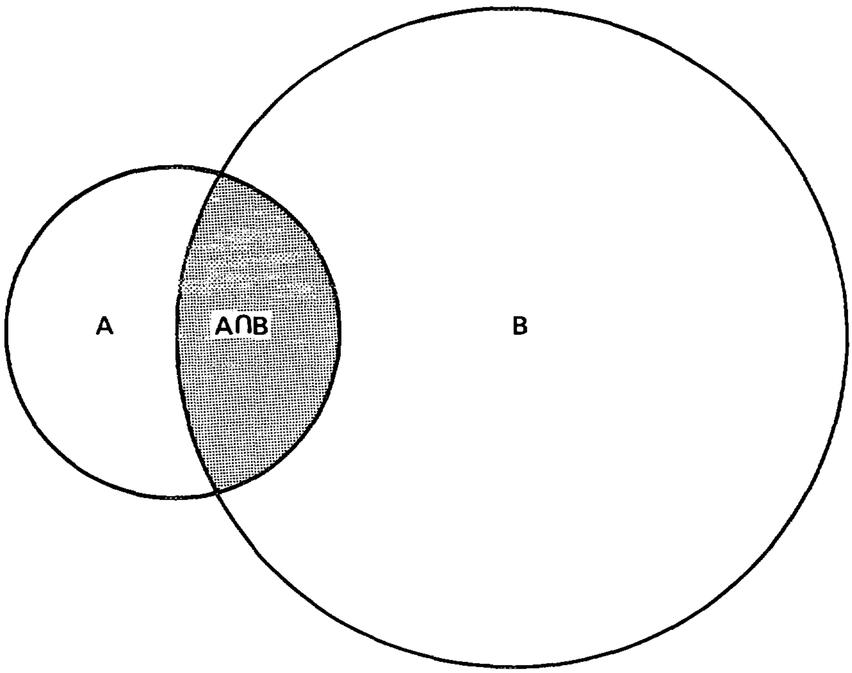
strength of association of .166 or 17%. This can be visualized as two overlapping circles, each with an area proportional to its total citations (Figure 2). The shaded area of overlap is 17% of the total (shaded and unshaded) area, i.e., ten citations out of the total 60 for A and B.

The strength threshold is important because it determines how sharply focused the cluster will be. For example, if we set the threshold at 0%, one very large cluster is generated composed of nearly *all* co-cited documents in the set. At the other extreme, a 100% threshold would generate almost as many separate "clusters" as there are individual cited papers, except for those which are invariably co-cited. Neither extreme would provide a meaningful or useful result.

If you are unfamiliar with clustering procedures, it may seem arbitrary to set strength thresholds "anywhere" between 0-100%. However, no clustering procedure has a predetermined or recommended threshold that produces a "valid" result. There is simply no substitute for exploring the data and evaluating the results. For example, to identify individual specialties, our experience tells us to set the threshold so that no more than 100 *cited* documents are included in a single cluster. Any larger cluster usually is a composite of more than one subject.

We normally process the set of co-cited documents at a number of different strength thresholds by a method called "single-link" clustering.[9] In single-link clustering, the computer selects a single document and searches for all the other items to which it is linked with a co-citation strength equaling or exceeding the specified threshold.

Figure 2: Strength of association between co-cited documents, A and B, represented as two overlapping spheres. The shaded area of intersection represents co-citations of A and B.



ISI's single-link method is different from the standard definitions included in clustering texts,[5-7] because the volume of files we cluster is unusually large. In particular, our approach differs from others in our use of a direct access disk to store the co-citations of linked documents. In other words, at a specific location on the disk are stored all the links that a particular document has to any other document. This greatly simplifies the implementation of single-link clustering.

Classifying research documents into related fields or clusters by co-citation analysis shouldn't be confused with another method of defining relationships between documents called "bibliographic coupling." Bibliographic coupling links *source* (citing) documents together. When two papers cite one or more references in common, they are bibliographically coupled. Co-citation is a relationship between *cited* documents—when two papers are cited together by a later paper, they are linked by co-citation.

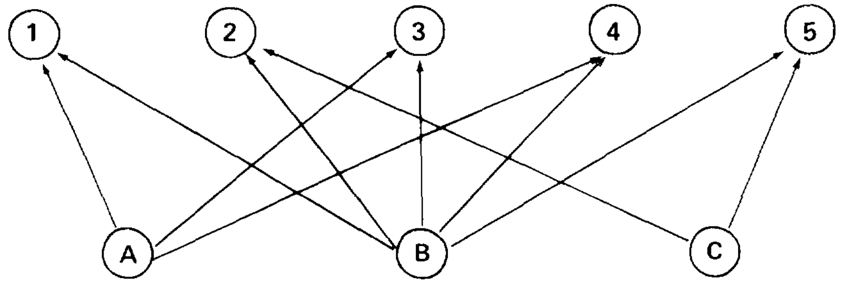
Figure 3 illustrates the difference between the two. Paper A cites references 1, 3, and 4. Thus, papers (1 + 3), (1 + 4), and (3 + 4) are co-cited documents. Paper B cites all five references, yielding ten different pairs of co-cited documents—(1 + 2), (1 + 3), (1 + 4), (1 + 5), (2 + 3), (2 + 4), (2 + 5), (3 + 4), (3 + 5), and (4 + 5). Paper C cites only 2 and 5, and that is the only co-cited pair of documents it yields. Thus, *all* five references are co-cited with each other at least once by papers A, B, and C.

Also, papers A and B are bibliographically coupled because they both cite references 1, 3, and 4. Papers B and C are coupled by common references to document 5 in their bibliographies. However, A and C are *not* bibliographically coupled since they share no common references. But these two papers could be co-cited in the future by one or more papers—and the more often they are co-cited, the greater is the chance that they have become a "marker" or identifier for a new subject or even a field.

The disadvantage of classifying documents by bibliographic coupling is that it can't take into account the constant evolution of research—two source items either are or are not bibliographically coupled for all time. However, co-citation relationships are dynamic and reflect the evolution, decline, and merger of research fields—two documents that are not presently co-cited may be linked together in later publications.

Co-citation analysis is an *automatic* classification procedure[10] that minimizes reliance on arbitrary human judgments. In conventional systems not only may different indexers index the same paper inconsistently, but, more important,

Figure 3: Co-citation links and bibliographic coupling. Lettered circles represent source documents and numbered circles are cited references. All five references are co-cited at least once by sources A,B, and C. A-B and B-C are bibliographically coupled, but A and C are not.



their hierarchical classification systems frequently get out of touch with the realities of current science. In the procedure I am describing, human operators input the bibliographic information for each paper and the computer takes the citations used by the author to assign the paper to its appropriate category.

In the second part of this essay, we'll take a step-by-step look at how the computer actually generates a cluster in a specific field—the structure of red blood cell membranes. We'll also show how the size of the cluster and the links between co-cited documents change when the frequency and strength thresholds are varied. These exercises will demonstrate how ISI's clustering procedure is used both as a tool for analyzing the structure of science and as an information retrieval system with high precision and recall.

* * * *

My thanks to Alfred Welljams-Dorof for his help in the preparation of this essay and to Ronald Levine, Jim Shea, and Beta Starchild for the naming of the clusters.

REFERENCES

1. **Marx J L.** Lasker award stirs controversy. *Science* 203:341, 1979.
2. **Garfield E.** Controversies over opiate receptor research typify problems facing awards committees. *Current Contents* (20):5-19, 14 May 1979.
3. ------------. ISI is studying the structure of science through co-citation analysis. *Current Contents* (7):5-10, 13 February 1974.*
4. ------------. The 'obliteration phenomenon' in science—and the advantage of being obliterated! *Current Contents* (51/52):5-7, 22 December 1975.*
5. **Anderberg M R.** *Cluster analysis for applications.* New York: Academic Press, 1973. 359 p.
6. **Hartigan J A.** *Clustering algorithms.* New York: Wiley, 1975. 351 p.
7. **Sneath P H A & Sokal R R.** *Numerical taxonomy.* San Francisco: W.H. Freeman, 1973. 573 p.
8. **Garfield E.** Mapping the structure of science. *Citation indexing: its theory and application in science, technology, and humanities.* New York: Wiley, 1979. p. 98-147.
9. **Small H G & Griffith B C.** The structure of scientific literature, I: identifying and graphing specialties. *Sci. Stud.* 4:17-40, 1974.
10. **Garfield E, Malin M V & Small H.** A system for automatic classification of scientific literature. *J. Indian Inst. Sci.* 57:61-74, 1975.

*Reprinted in: **Garfield E.** *Essays of an information scientist.* Philadelphia: ISI Press, 1980. 3 vols.