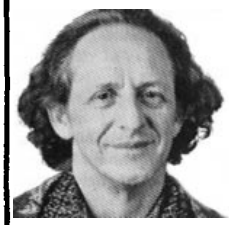
Essays of an Information Scientist, Vol:4, p.642-649, 1979-80 Current Contents, #41, p.5-12, October 13, 1980



Current Comments

ABCs of Cluster Mapping. Part 2. Most Active Fields in the Physical Sciences in 1978

Number 41 October 13, 1980

Last week, we listed the 100 top 1978 clusters in the life sciences.[1] This week, in Table 1, you'll find the list for the physical sciences. The clusters are ranked by the number of papers published in 1978. For those who want the details of how these lists were created—read on.

In the first part of this essay, I explained that ISI®'s clustering procedure differs from the classic techniques for a variety of reasons. Since our procedure is order independent, we begin our cluster analysis by choosing any document at random. No matter which document we start with, the resulting structure is the same in the end. What follows is a description of the computer procedure that is followed to create one particular cluster of the 2,000 generated from an annual *Science Citation Index® (SCI®)* file. After the clusters are generated, the computer creates two-dimensional "maps" showing the spatial co-citation relationships between the cited papers in each cluster.

Before we take a step-by-step look at the computerized generation of a cluster, let me reiterate two important terms. *Frequency threshold* refers to the number of citations a paper must receive in a given year before it is included in a cluster. We set frequency thresholds to reduce the number of cited papers or books included to a more manageable level. It would have been costly and time-consuming to examine all the different works cited by the 500,000 published papers we indexed in 1978. Instead, we selected only those cited 17 or more times—less than 1% of the 3.5 million cited in the 1978 *SCI*.[1]

Strength threshold refers to the degree of association between co-cited pairs of documents—that is, the proportion of their total citations that are co-citations. There is no way to tell in advance what the optimal strength threshold should be. Thus, we process the set of documents at several different strength thresholds to make sure we cover all significant co-citation relationships. Usually, the threshold is set so that no more than 100 cited documents appear in a single cluster.[1]

With these definitions in mind, we can now take a step-by-step look at how the computer actually generates a cluster in a specific field. The cluster we'll examine deals with the structure of red blood cell membranes. We've set the citation frequency at 20 or more for this cluster: that is, any cited document having less than 20 citations will be passed over. Also, the strength threshold is set at 24% or greater association: that is, any pair whose proportion of co-citations is less than 24% of their total citations will be passed over. A document must meet *both* frequency and strength requirements before it appears in the cluster.

(Figure 1) Remember that we have direct access to all documents stored in the computer memory. We arbitrarily decide to start the run with a paper whose primary author is Cabantchik, indicated by the empty square. Cabantchik's paper is linked to three other

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

Cluster Number	Cluster Name	Cited Articles	Citing Articles
32	Opiate Receptors & Opioid Peptides	124	947
23	SV40 & Adenovirus Genome Structure	88	698
260	Substance-P	52	640
775	Prostaglandins & Thromboxanes	775	605
272	Chromatin Structure	50	534
99	Fibronectin	50	432
121	Hypothalamic Hormones	48	423
81	Somatostatin	38	411
625	Polycyclic Hydrocarbons & Cancer	40	372
207	Translation of RNA Tumor-Viruses	57	359
22	Myasthenia Gravis	59	353
405	Vitamin-D Metabolism	52	352
315	Bacteriorhodopsin	41	345
27	Red Blood Cell Membrane Structure	32	316
105	Sister Chromatid Exchanges	36	299
582	Cell Filaments	26	291
75	Cell-Mediated Cytotoxicity	34	288
506	Cell-Mediated Immunity & Major Histocompatibility Complex	8	271
274	Chromatin Reconstitution & Transcriptional Activity	25	262
258	Mixed-Function Oxidases	23	261
940	Deficient DNA Repair in Carcinogenicity & Xeroderma Pigmentosum	21	256
308	Monocular Deprivation	25	246
548	Dopamine Receptors	21	243
403	Type-C RNA Tumor Viruses	37	239
836	High Density Lipoprotein & Atherosclerosis	6	234
107	Human Neutrophils	27	230
7	Beta Endorphin	19	227
196	Platelets, Platelet Factors & Atherogenesis	14	226
662	Cell Colony Stimulating Activity	21	224
795	Neuroleptic Receptors	7	222
538	Adenylate Cyclase System	25	215
270	Vasodilator Therapy of Congestive Heart Failure	29	212
670	Intracellular Calcium Regulation	23	212
79	Localization of GABA-energic Projections in the Central Nervous System	16	207
1,645	Nasopharyngeal Carcinoma & Epstein-Barr Viral Markers	12	206
36	Acetylcholine Receptors	24	205
233	Histamine H2 Receptors & Antagonists	4	202
1,059	Interaction of Bacteria Toxins with Membrane Receptors	23	199
384	Outer Membrane Proteins of Bacteria	26	196
148	Calcium-Dependent Modulator Protein	25	195
267	Control of Myeloid Leukemia Cell Differentiation	15	194
1,106	Enterotoxigenic Escherichia coli	21	191
273	Messenger RNA Structure & Metabolism	18	191
14	Lymphocyte Membrane Immunoglobulin	21	190
205	Clinical Applications & Pharmacology of Theophylline	20	189
484	Thyroid Hormone Metabolism	18	189
1,291	Genetic Control of Immune Response	6	189
1,095	Macrophage Activation	9	186
557	Platelet Suppressant Therapy	20	185
98	Hippocampal Organization	25	184
696	Ischemic Myocardium	21	184

1,294	Hydrophobic Chromatography	8	81
876	Climate Modeling	7	78
2,196	Organoselenium Compounds	7	78
1,889	Solar Wind	6	78
339	Ion-Atom Collisions	4	78
798	Charge Transfer Complexes with TCNQ (Tetracyanoquinodimethane)	6	77
428	Beta Aluminas	6	77
1,001	Position Vacancy Interactions in Metals	5	77
841	Amorphous Semiconductors	5	76
572	Multi-Photon Dissociation of Molecules	6	75
1,147	Spin Labels & Membranes	4	75
296	Stellar Distribution Around Black Holes	8	74
435	Cochlear Mechanics	4	73
655	Mantle Xenoliths from Kimberlite	11	72
299	Heterogeneous Metal-Complex Catalysts	6	72
132	Conformations of Nucleosides & Nucleotides	5	72
2,041	Hydrides of Rare Earth Intermetallic Compounds	5	72
812	Mixed Valence Complexes	5	70
883	Dual Unitarization Approach to Hadron Reactions	8	69
157	Isotopic Structures in Solar-System Materials	6	69
709	Polymeric Sulfur Nitride	5	69
1,768	Acid Catalysis	4	69
712	Electron Hole Liquid in Semiconductors	4	69
846	Resonance Fluorescence	7	68
116	Resonances in Heavy-Ion Collisions: C_{12}-O_{16} System	7	68
1,272	Electron-Positron Annihilation in Quantum Chromodynamics	5	68
439	Chlorination of Organics in Water Treatment	4	68
1,722	Polyene Spectroscopy	7	67
1,119	H^1 NMR Studies	5	67
178	Fast-Rotating Heavy Nuclei	4	67
1,347	Supersymmetry & Superfields	4	67
395	Solid-State Polymerization	7	66
1,796	Josephson Junction	4	66
163	Quark-Parton Model	4	66
1,799	Renormalization Group Transformation	4	65
855	Upper Mantle Rheology	5	64
39	Seyfert Galaxies	5	64
368	Many-Body Perturbation Theory	4	63
1,600	Oxygen Radicals in Biological Reactions	4	63
808	Olefin Metathesis	7	62
421	Field Desorption Mass Spectrometry	6	62
525	2-Photon Laser Spectroscopy	6	62
972	A-15 Superconductors	5	62
1,954	Spiral Galaxies	4	62
762	Molecular Orbital Studies	4	61
1,056	Nitrous Oxide in Environment	4	61
949	Polymer Solutions	4	61

documents by co-citation. One is again authored by Cabantchik, and the other two by Ho and Lepke. The computer can now identify any documents co-cited with either Cabantchik (2), Ho, or Lepke since our clustering procedure is order independent. For convenience sake, we'll proceed in alphabetical order and see what is co-cited with Cabantchik (2). Keep in mind that if we chose to continue with Lepke instead, the resulting cluster would be the same.

(Figure 2) As it turns out, Cabantchik (2) is co-cited only with Cabantchik (1) at these frequency and strength levels. The computer automatically switches to Ho. Obviously, co-citation is a mutual relationship, and Ho's paper is linked

Figure 1: First step in computer-generated cluster development. Empty square indicates document which is being examined for co-citation links.

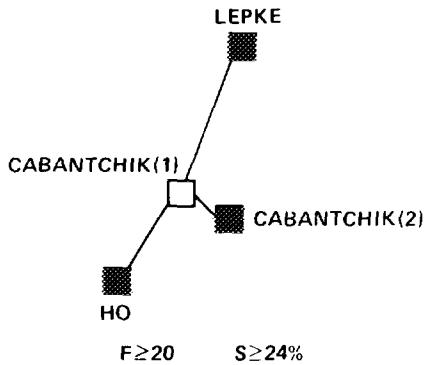
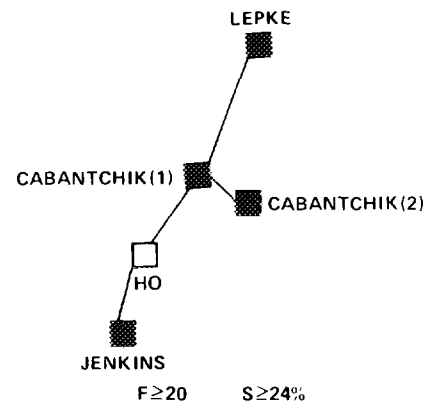


Figure 2: Second step in computer-generated cluster development.



with Cabantchik (1). The new addition to the developing cluster is Jenkins' document, co-cited with Ho's.

(Figure 3) The computer now centers on all documents co-cited with Jenkins, and four new items appear—Findlay, Steck, Tanner, and Yu.

(Figure 4) Jumping to Findlay, the computer identifies a link between it and Yu, who already appeared in the cluster, but no new items are added. At these frequency and strength thresholds, Findlay is linked only to Jenkins and Yu.

(Figure 5) Moving to Steck, the computer identifies two new co-citation links. One is to Yu (1), and the other is to a second paper authored by Yu (2), which appears in the cluster for the first time. The computer proceeds to Tanner, Yu (1), and Yu (2), but no new links or documents are generated. At a frequency of 20 or more citations and a strength of 24% or greater association, the cluster is complete. That is, no other documents in the entire file are co-cited at these thresholds with any of the items already in this particular cluster. At

Figure 3: Third step in computer-generated cluster development.

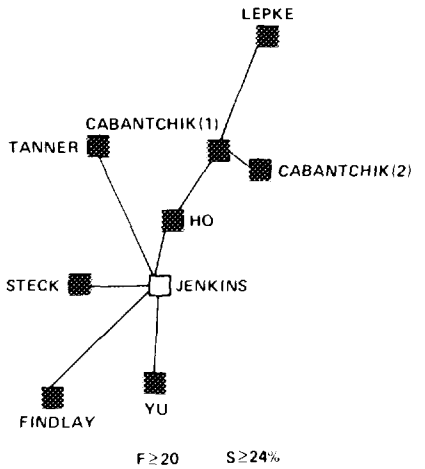


Figure 4: Fourth step in computer-generated cluster development.

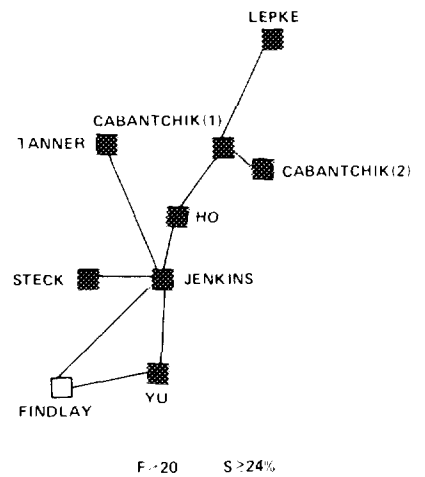
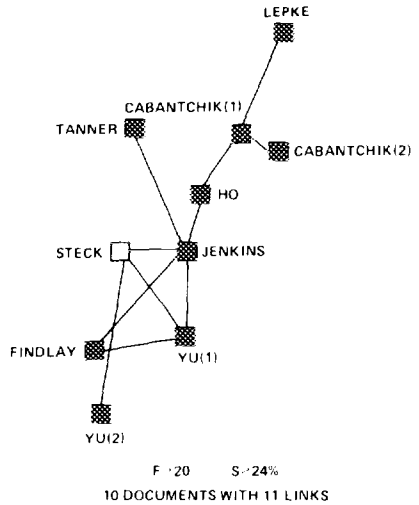


Figure 5: Fifth step in computer-generated cluster development. At these frequency and strength thresholds, no other documents in the file are linked by co-citation with those appearing in this figure.



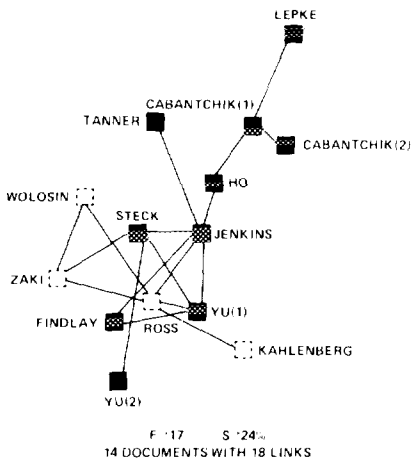
these values, ten cited documents with 11 links are identified that deal with red blood cell membrane structure.

(Figure 6) We can now vary the frequency threshold but hold the strength constant to see how this affects the cluster. Instead of searching for documents with 20 or more citations, the computer picks up any co-cited document with 17 or more citations that also meets the 24% strength threshold. Obviously, the ten documents in the original cluster remain, since they were cited 20 or more times that year. Lowering the frequency threshold to 17 or more citations adds four new papers, indicated by squares in broken lines —Kahlenberg, Ross, Wolosin, and Zaki. However, no new co-citation links are formed, except between the added documents, because the strength is held constant at 24% or greater. At these values, 14 documents with 18 links are now identified in the cluster on red blood cell membrane structure.

(Figure 7) We now return the frequency threshold to 20 but lower the co-citation strength limit. Instead of searching for documents whose proportion of co-citations is 24% or more of their total citations, the computer identifies those with a strength of at least 22%. Six new documents appear—Adair, Bender, Berg, Bretscher (1), Bretscher (2), and Wolosin. Also, ten new co-citation links are formed, indicated by the dotted lines. At these values, 16 documents with 21 links are identified in the cluster.

(Figure 8) We now lower both the frequency and strength thresholds. Instead of searching for documents having 20 or more total citations and an association strength of 24% or more, the computer isolates all documents having 17 or more total citations and a 22% or greater strength of association. Even this seemingly small recalibration of threshold values has very dramatic results. At these values, 32 documents with 47 links are identified.

Figure 6: Frequency is lowered from 20 citations to 17. Strength is held constant at 24%. Broken-line squares indicate new documents appearing as a result of lowering the frequency threshold.



With this particular cluster as an example, you can visualize the same procedure being followed for each of the as yet unclustered documents in the file. The actual map in Figure 8 was generated by a separate computerized pro-

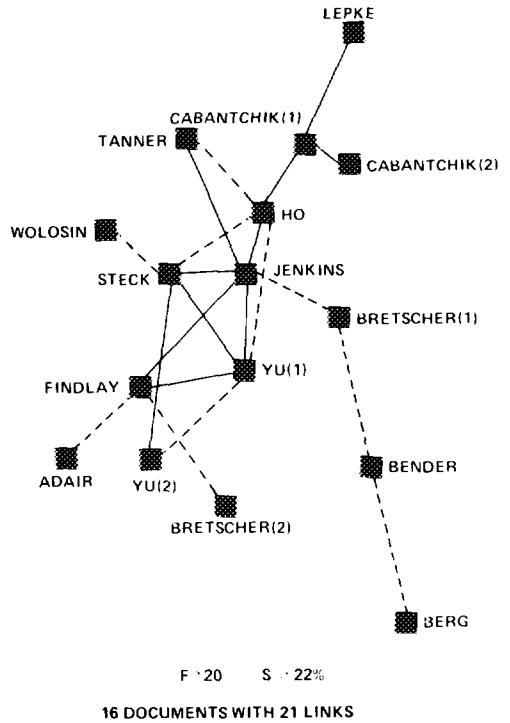


Figure 7: Strength is lowered from 24% to 22%. Frequency is held constant at 20 citations. Broken lines indicate new co-citation links appearing as a result of lowering the strength threshold.

cedure called "multidimensional scaling."[2] The maps in Figures 1-7 were derived from that master map. Any objects that have a specific relationship to one another can be arranged in a *spatial* configuration by multidimensional scaling.

For example, if the objects are cities and their specific relationships are the distances between them, multidimensional scaling would generate a two-dimensional "road map" showing the configuration of the cities in space. In this cluster, the objects are the cited items and their specific relationships are the strengths of association between them. Thus, ISI's cluster analysis generates a "road map" or "atlas" of science. The actual scaling procedure utilizes the quantitative linkages between points to assign them positions relative to one another. The distances between the points reflect the magnitude of the linkage measure.[2]

Cluster analysis is a very powerful and useful tool for science analysts. It can bring into focus the macroscopic structure of science and show how chemistry, physics, biology, and medicine are related by setting the strength threshold at a low value. Or it can resolve the microscopic structure of opiate receptor research, membrane structure research, or any other small specialty by setting a high strength threshold. Moreover, cluster analysis can show how science evolves by generating maps covering a sequence of years.[3] The progress and stagnation of specialties, the mergers and divisions of fields, the identity of "gatekeeper" researchers, the contributions of individual institutions—all this can be made accessible to the non-

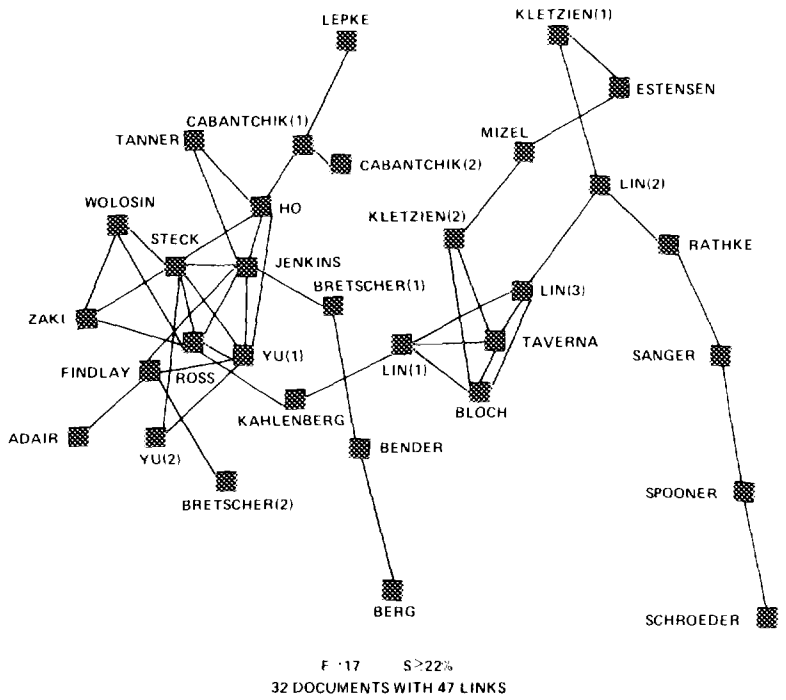


Figure 8: Frequency is lowered from 20 citations to 17. Also, strength is lowered from 24% to 22%.

specialist in a graphic and easy-to-read map through cluster analysis.

Cluster analysis can also be part of an accurate and comprehensive on-line information retrieval system in the future. The *SCI* annual file usually yields about 2,000 clusters after the computer identifies all co-citation links. These clusters are then used as "pigeonholes" to classify all newly published papers. In order to locate a specific cluster, the researcher simply will request the computer to display the classified index to all the clusters for a specific year. What appears on the CRT screen is a "menu" of cluster titles, much like those shown in Table 1, in the natural language currently used by researchers in the field. Once the specific cluster is located, its number is entered and a complete bibliography of all papers we classified in the cluster will appear on the screen or be off-printed in hard copy.

For example, if you are interested in the effects of low temperatures on amorphous solids, you would be directed to cluster number 169, "properties of amorphous solids at low temperatures," which includes seven cited documents published in various years. You would retrieve 96 papers published in 1978 which cited various pairs of these milestone papers in that field (Table 1). If you also want to search the 1977, 1976, or other earlier years' literature, the computer will guide you to the appropriate cluster numbers, even if the cluster's name has changed over the years. The linkage is maintained through the cited papers rather than through title words.

Also, it is theoretically possible to specify varying frequency and strength thresholds when retrieving documents citing a specific cluster. This would be an important feature of a cluster-based

search for two reasons. First, as a cluster evolves over the years, the total number of citations for a particular article and the strength of association between co-cited documents obviously will change. This is particularly true when a cluster first emerges—the thresholds should be set low enough to "focus" on the cluster when it is still relatively small.

Second, the number of papers that cross an arbitrary citation threshold will depend on the number of references cited in an average paper, which varies to some extent according to the size of the field. Thus, clustering papers in mathematics would require different thresholds than those for biochemistry. Therefore, it may be necessary to segment files by discipline to ensure that smaller specialties are not neglected.

ISI is now testing a new search strategy based on clustering in a new on-line biomedical data base. These are essentially the articles listed in the life sciences edition of *Current Contents*® (*CC*®). As it turns out, these *CC* journals in biomedical research account for about 40% of the annual *SCI* file. In addition to being searchable by specialty cluster headings, articles in the file are also accessible by title word, source, and citation searching. The biomedical file is designed for use by the smaller research library or the individual specialist researcher who needs a comprehensive and timely bibliography on a very specific topic. I'll have more to say about ISI's on-line biomedical file in another essay.

After these initial tests, similar procedures will be developed for the physical sciences, mathematics, and social sciences.[4] It is not absolutely necessary for you to fully understand the details of our clustering procedures in order to appreciate or be critical of our results. Any searching procedure — whether it is based on co-citation, citation, or title words—can only be tested in the field if it is to conform to the needs of users. We have by no means exhausted all the possibilities in exploiting these methods, even for citation-based approaches. But when we consider the possibility of combining the co-citation method with word co-occurrence clustering, we may approach an ideal system that uses the best of both worlds.

* * * *

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

REFERENCES

1. **Garfield E.** ABCs of cluster mapping. Part 1. Most active fields in the life sciences in 1978. *Current Contents* (40):5-12, 6 October 1980.
2. **Kruskal J B.** Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1-27, 1964.
3. **Small H G.** Structural dynamics of scientific literature. *Int. Class.* 3:67-74, 1976.
4. **Garfield E.** *Social Sciences Citation Index* clusters. *Current Contents* (27):5-11, 5 July 1976.*

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