Reprinted from *Journal of the Indian Institute of Science* 57(2):61-74, 1975.

Also See Introduction: "Clusters and classification ", p.354, Essays, Vol:2

A System for Automatic Classification of Scientific Literature

by

Eugene Garfield, Morton V. Malin and Henry Small

Institute for Scientific Information
325 Chestnut Street
Philadelphia, Pa. 19106

ABSTRACT

A computer-based system for automatically classifying scientific articles is described. The unique feature of this system is that its structure is completely determined by citation patterns in the *Science Citation Index*. These citation patterns give rise to clusters or clumps of cited papers which correspond, in turn, to clusters of citing papers. Classification headings for each cluster are determined by examination of high frequency word pairs drawn from the titles of the citing papers. Classification of a new article is performed automatically by determining what clusters it cites and assigning appropriate (weighted) subject headings to it. The system will permit updating of the classification scheme on an annual basis, and the incorporation of new headings and deletion of old ones.

This paper describes an automatic classification system being developed at ISI®, with the unique feature that its structure is completely determined by citation patterns derived from the *Science Citation Index*® data base. I will also summarize some of the current research at ISI. ISI's research and development objectives include the development of new information products and services, and development of improved processing operations, methods, and systems. But we also conduct basic research in the area of information science; the last activity supports the first two. We believe that the more we learn about the characteristics of the scientific literature, and its relationship to sciene and research communication, the better we will be able to develop and provide services to the user. The automatic classification system discussed in this paper is an outgrowth of a basic research project now being conducted at ISI using the *Science Citation Index* data base.

This data base now consists of 13 years of back files containing nearly 3.4 million source articles and nearly 40 million citations. Thus, we have an unusual opportunity for conducting a broad program of research activities, using the file to study the characteristics of the literature and to conduct citation behavior studies in the history and sociology of science. These studies are very productive both for ISI and the scientific community.

Before describing some of the research and results a brief description of citation indexing would be useful, because it is necessary to understand this data base in order to understand the research work we are doing. Since the literature already contains excellent detailed descriptions of citation indexing,[1-3] I will not discourse in detail on the *SCI*®, but only describe the concept and the data one has available in the printed Index and on computer tape.

In brief, a citation index is a cumulation of journal article references arranged so that one can determine what later or more current articles have cited any earlier article or book. The *Science Citation Index* arrangement is alphabetical by the cited author of each cited item. Under each cited item is listed all the later articles which have cited it during a specified time period, *e.g.*, three months, one year, or five years. ISI now processes about 2400 journals for the *SCI*, and all references in all of the articles in these journals are keyed into the data base, and eventually appear in the printed *SCI* cumulations. At the same time, a number of other data elements are keyed from the source articles. These include: all authors of a given article, author addresses, full title of the article, and journal, volume, page, and year. The number of references keyed in 1973 was roughly 5 million coming from approximately 400,000 source items. It is virtually the only data base, available, which includes the bibliographic, that is cited, references.

It was long ago pointed (1955) that these cited references are a unique and important group of indexing terms.[4] Salton,[5] in particular, has confirmed the value of citation indexing for retrieval of information. Thus, bibliographic citations are important indicators of document content. Human indexers do not ordinarily think of citations as descriptors of the citing document, but they are, in fact alternate representations of the documents they identify.[4,6] Were this not true, the automatic classification system described below would not be possible.

I will digress briefly from the main topic to describe a project which illustrates how we benefit from research with our data. This project is called the Journal Citation Index. To create this compilation, every citation from the source items processed during the last quarter of 1969 was extracted from the total year's file. Through a series of sorts, a new type of citation file was created which, instead of obtaining citations to articles, obtained citations to the Journals in which the cited articles were published. Further programming then produced listings providing data showing for each cited journal which journals had been cited. Counts were made to show the frequency with which each journal was cited, and the year of the cited articles. The process of analysis continued until we were able to produce statistical indicators which would permit ranking of the journals based on factors other than just the frequency of citations. A description of this project can be obtained from my 1972 *Science* articles.[7] Indeed, this project illustrates well the classificatory power of citation analysis. What other means do we have available today for categorizing journal collections?

The purpose of the JCI project was to test the hypothesis that citation frequency is a measure of impact of a journal. We believed that such data could help us and others in developing a core list of scientific journals as well as aid in journal evaluation and selection procedures. In 1973, the listings were published by ISI as a service for libraries under the title *ISI's Journal Citation Reports*®. More recent data are now available covering the year 1972, and we plan to produce an updated *JCR*® on a regular basis.

More relevant to the subject of this paper is the second research project which I will describe: work being done by Dr. Henry Small of ISI's R&D staff, "Mapping of Scientific Specialties." The work is being supported by a grant from the National Science Foundation. Although primarily a project concerned with historical and sociological aspects of science, it has great relevance to information science, and to our automatic classification system. The objective of this research project was to test the hypothesis that citations to scientific articles could be used in identifying scientific specialties, in effect that citation data could be used for classificatory purposes. Thus, an understanding of the classification system described below requires initially an understanding of the research from which it stems.

It is appropriate also at this point to define automatic classification because clustering is an essential part of classification and of the specialty mapping research. *Webster's New Collegiate Dictionary* defines classification as "systematic arrangement in groups or categories according to established criteria." An automatic classification system can thus be defined as a method for systematically arranging documents in groups or categories by a process that requires no human intervention, save the keying of the text. In this context, text may be full text titles, abstracts, or citations in a bibliography. A system for automatic classification is, therefore, one which satisfies the requirement of clustering or bringing like things together or as a process which groups objects resembling one another in terms of their properties.

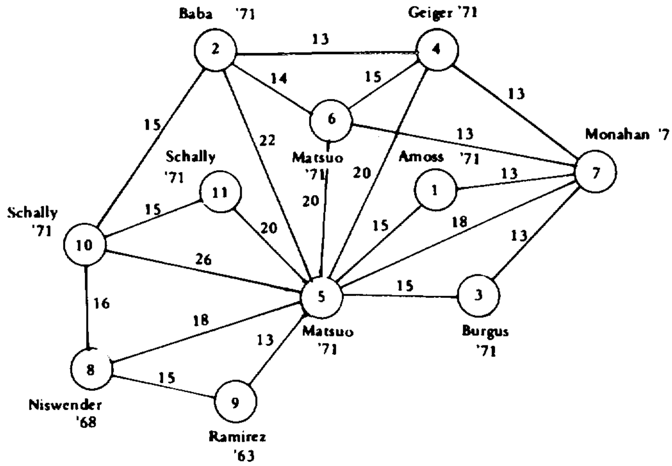
The specialty mapping project uses a computer based technique to identify clusters of highly cited and co-cited scientific articles. Co-citation is defined as the number of times two publications are cited together in the literature. The clusters formed are found to correspond to scientific specialties. The technique employed begins with identification of highly cited papers in an annual file of the *SCI*. To initiate the experiment, the 1973 file was used, and all papers cited at least fifteen times were extracted. This reduced the total file of approximately 4,000,000 citations to a more manageable one containing 430,000 citations. Out of more than 2,000,000 unique cited items in 1973, only 16,000 items were selected. For each cited paper we extracted the list of citing papers and this new file was then resorted so that we could identify pairs of papers cited together, *i.e.*, co-cited. The number of identical pairs of cited papers were then counted to establish the co-citation strength of each pair of papers and a total of 710,000 distinct co-cited pairs were generated through this method.

The next step was to apply a clustering algorithm in order to group together the most highly co-cited documents. The clustering algorithm used is a single-link procedure. Briefly, to describe this procedure, a minimum linkage level is specified and the computer begins by selecting an initial document and retrieving all of its linkages to other documents which equal

First Citation Frequency			Bibliographical Data				Freq.	Strength
1.	19	Amoss M.	Biochem. Biophys. Res.	44	205	71		
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	15
		Monahan M.	Comptes Rendus, etc.	273	508	71	26	13
2.	28	Baba Y.	Biochem. Biophys. Res.	44	459	71		
		Geiger R.	Biochem. Biophys. Res.	45	767	71	24	43
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	22
		Matsuo H.	Biochem. Biophys. Res.	45	822	71	23	14
		Schally A V.	Biochem. Biophys. Res.	43	393	71	42	15
3.	18	Burgus R.	Comptes Rendus, etc.	273	1611	71		
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	15
		Monahan M.	Comptes Rendus, etc.	273	508	71	26	13
4.	24	Geiger R.	Biochem. Biophys. Res.	45	767	71		
		Baba Y.	Biochem. Biophys. Res.	44	459	71	28	13
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	20
		Matsuo H.	Biochem. Biophys. Res.	45	822	71	23	15
		Monahan M.	Comptes Rendus, etc.	273	508	71	26	13
5.	59	Matsuo H.	Biochem. Biophys. Res.	43	1334	71		
		Amoss M.	Biochem. Biophys. Res.	44	205	71	19	15
		Baba Y.	Biochem. Biophys. Res.	44	459	71		
		Burgus R.	Comptes Rendus, etc.	273	1611	71	18	15
		Geiger R.	Biochem. Biophys. Res.	45	767	71	24	20
		Matsuo H.	Biochem. Biophys. Res.	45	822	71	23	20
		Monahan M.	Comptes Rendus, etc.	273	508	71	26	18
		Niswender G D.	P. Soc. Exp. Biol. Med.	128	807	68	71	18
		Ramirez V D.	Endocrinology	73	193	63	23	13
		Schally A V.	Biochem. Biophys. Res.	43	393	71	42	26
		Schally A V.	Science	173	1036	71	44	20
6.	23	Matsuo H.	Biochem. Biophys. Res.	45	822	71		
		Baba Y.	Biochem. Biophys. Res.	44	459	71	28	14
		Geiger R.	Biochem. Biophys. Res.	45	767	71	24	15
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	20
		Monahan M.	Comptes Rendus, etc.	273	508	71	26	13
7.	26	Monahan M.	Comptes Rendus, etc.	273	508	71		
		Amoss M.	Biochem. Biophys. Res.	44	205	71	19	13
		Burgus R.	Comptes Rendus, etc.	273	1611	71	18	13
		Geiger R.	Biochem. Biophys. Res.	45	767	71	24	13
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	18
		Matsuo H.	Biochem. Biophys. Res.	45	822	71	23	13
8.	71	Niswender G D.	P. Soc. Exp. Biol. Med.	128	807	68		
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	18
		Ramirez V D.	Endocrinology	73	193	63	23	15
		Schally A V.	Biochem. Biophys. Res.	43	393	71	42	16
9.	23	Ramirez V D.	Endocrinology	73	193	63		
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	13
		Niswender G D.	P. Soc. Exp. Biol. Med.	128	807	68	71	15
10.	42	Schally A V.	Biochem. Biophys. Res.	43	393	71		
		Baba Y.	Biochem. Biophys. Res.	44	459	71	28	15
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	26
		Niswender G D.	P. Soc. Exp. Biol. Med.	128	807	68	71	16
		Schally A V.	Science	173	1036	71	44	15
11.	44	Schally A V.	Science	173	1036	71		
		Matsuo H.	Biochem. Biophys. Res.	43	1334	71	59	20
		Schally A V.	Biochem. Biophys. Res.	43	393	71	42	15

Figure 1. Cluster 33. First citations are item-numbered on the left. Bibliographical data consists of journal title abbreviation, volume, beginning page, and year.

1. Amoss M. Purification, amino-acid composition and N-terminus of hypothalamic luteinizing hormone releasing factor (LRF) of ovine origin. *Biochem. Biophys. Res.* 44:205, 1971.
2. Baba Y. Structure of porcine LH-releasing and FSH-releasing hormone. II. Confirmation of proposed structure by conventional sequential-analysis. *Biochem. Biophys. Res.* 44:459, 1971.
3. Burgus R. Molecular structure of hypothalamic factor of ovine origin controlling secretion of hypophyseal gonadotropic luteinizing-hormone (LH). *Comptes Rendus, etc.* 273:1611, 1971.
4. Geiger R. Synthesis and characterization of a decapeptide having LH-RH/FSH-RH activity, *Biochem. Biophys. Res.* 45:767, 1971.
5. Matsuo H. Structure of porcine LH-releasing and FH-releasing hormone. I. Proposed amino-acid sequence. *Biochem. Biophys. Res.* 43:1334, 1971.



6. Matsuo H. Synthesis of porcine LH-releasing and FSH-releasing hormone by solid-phase method. *Biochem. Biophys. Res.* 45:822, 1971.
7. Monahan M. Total synthesis by solid-phase of decapeptide stimulating secretion of hypophyseal gonadotropin LH and FSH. *Comptes Rendus, etc.* 273:508, 1971.
8. Niswender G D. Radioimmunoassay for rat luteinizing hormone with antiovine LH serum and ovine LH-1311. *P. Soc. Exp. Biol. Med.* 128:807, 1968.
9. Ramirez V D. A highly sensitive test for LH-releasing activity-ovarietomized, estrogen progesterone-blocked rat. *Endocrinology* 73:193, 1963.
10. Schally A V. Isolation and properties of FSH and LH-releasing hormone. *Biochem. Biophys. Res.* 43:393, 1971.
11. Schally A V. Gonadotropin-releasing hormone-one polypeptide regulates, secretion of luteinizing and follicle-stimulating hormones. *Science* 173:1036, 1971.

Figures 2 and 3. FSH- and LH-releasing hormones.

or exceed the minimum threshold. A cluster is complete when all documents have been identified which are linked together in a connected graph by linkages which satisfy the threshold criterion. At this point, the computer proceeds to the next unclustered document and generates another cluster. Clustering may be carried out at as many as four levels and the resulting clusters may be merged together to reveal the hierarchical or nested structure of the file.

Figures 1 through 3 show an example of a cluster obtained by this method. Figure 1 is a cluster as it emerges from the computer as a list of highly co-cited documents; Figure 2 is a cluster represented as a network with each document indicated by a circle and each line a co-citation linkage. Figure 3 is a list of the titles of the documents in the cluster which shows that the subject matter is quite narrowly focused on hormone releasing hormones.

At any single level, we may determine the linkages among clusters by counting the num-

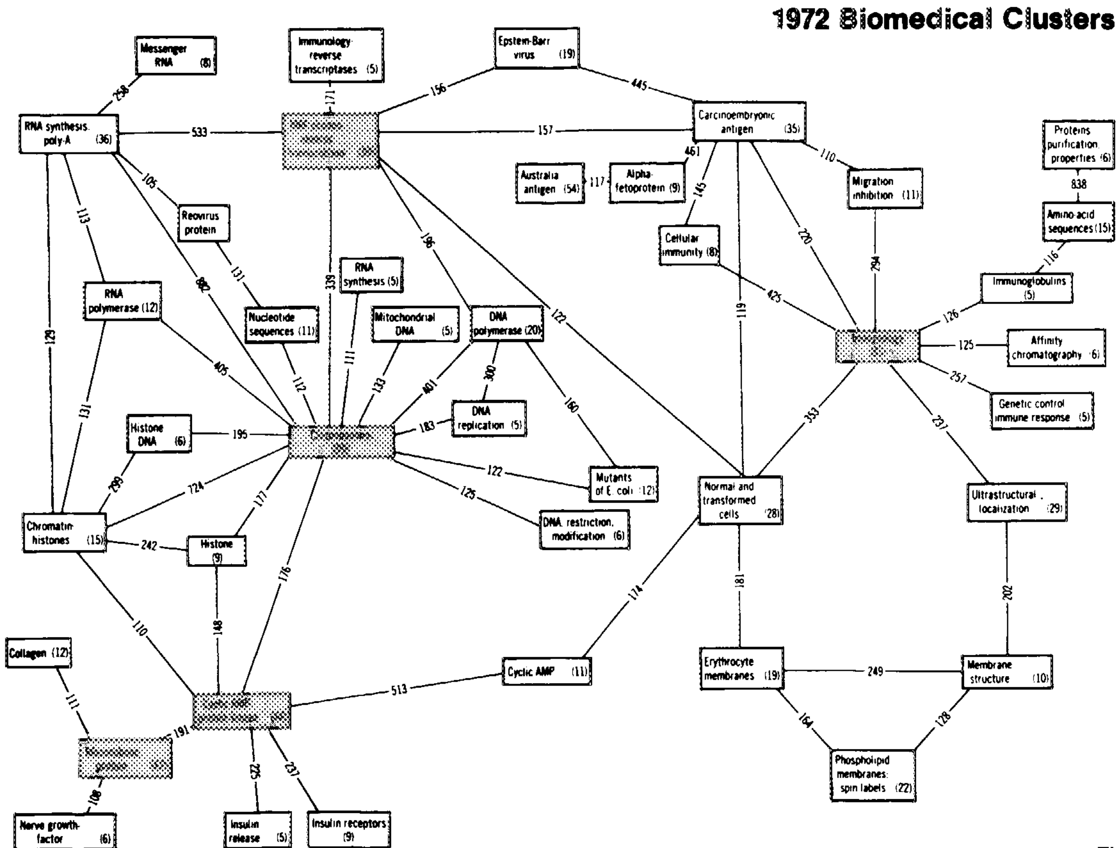

1972 Biomedical Clusters
Messenger RNA (8)
Immunology: reverse transcriptases (5)
Epstein-Barr virus (19)
RNA synthesis, poly-A (36)
Carcinoembryonic antigen (35)
Proteins purification properties (6)
Australia antigen (54)
Alpha-fetoprotein (9)
Migration inhibition (11)
Amino-acid sequences (15)
Reovirus protein
Cellular immunity (8)
RNA synthesis (5)
Immunoglobulins (5)
RNA polymerase (12)
Nucleotide sequences (11)
Mitochondrial DNA (5)
DNA polymerase (20)
Affinity chromatography (6)
Genetic control immune response (5)
Histone DNA (6)
DNA replication (5)
Mutants of E. coli (12)
Normal and transformed cells (28)
Ultrastructural localization (29)
DNA restriction, modification (6)
Chromatin-histones (15)
Histone (9)
Collagen (12)
Cyclic AMP (11)
Erythrocyte membranes (19)
Membrane structure (10)
Phospholipid membranes: spin labels (22)
Nerve growth-factor (6)
Insulin release (5)
Insulin receptors (9)

Figure 4.

ber of co-citations between documents in different clusters. This is called cluster co-citation. If, for example, we conduct a clustering run at level 11, the linkages among clusters are determined by co-citation links from one to ten. Using this information, we are able to draw as a graph, the network of most active specialties in science. By drawing such a map for each year, over a period of years, we can study how the links between specialties have changed, and where new specialties have emerged or old ones declined.

Figure 4 is a map of biomedical clusters derived from the 1972 *SCI* file. Only clusters containing three or more documents have been included and only if they have been linked with another cluster on the diagram by a cluster co-citation threshhold of 100. The map shows four major areas of biomedicine. In the upper left hand corner are chromosomes and RNA viruses, and in the upper right is work on immunology. Attached to immunology in the lower right is research related to biological membranes. To the left of this, in the lower left hand corner is work related to cyclic AMP. The pattern of specialties and linkages changes from year to year, and we can observe the evolution of this network over time.

The purpose of the mapping of the science project is to increase our understanding of the processes of growth and change in science, and to apply this understanding in the area of science policy. The important finding of our mapping work is that the basic unit of science appears to be the scientific specialty, not the discipline or the isolated researcher. Further, we have found that growth and change in specialties can be extremely rapid. These findings have important implications for information retrieval. First, they indicate that we must gear information services and classification schemes to the specialty scale, because this scale is probably most relevant to the working scientist, and is the one at which he generates and utilizes information.

Second, a classification scheme, if it is to be effective, must be capable of changing very rapidly. Probably an annual update is needed to respond to new developments and growth in some new specialties, but this will vary for different specialties. Some have a lifetime of as little as one year—others ten years or more. The precise life expectancy of a specialty is a question of considerable interest,, and one we should be able to answer using the ISI data base.

The application of our clustering work to classification is, therefore a highly natural one. Only one important modification is necessary to adapt an essentially science policy oriented system, where the criterion is the level of activity, to an information science oriented system where the criterion is to generate as many classification categories and classify as many articles as possible. The change consists in adopting a normalized linkage measure, rather than an absolute measure. Earlier, I described the procedure for determining the frequency of co-citation between two highly cited documents. It is a simple step to convert this absolute frequency into a percentage overlap. In clustering terminology, this is known as a Jaccard-Sneath matching coefficient. For example, if paper A is cited fifteen times and paper B is cited twenty times, and together they are co-cited five times, the matching coefficient or percentage overlap is .16% [5/(15+20—5)]. This technique is illustrated in Figure 5 where we have calculated the Jaccard-Sneath coefficient for documents by Armstrong and Edmonds.

The results of our analysis of the 1973 *SCI*® illustrates the normalized clustering method. An initial citation frequency threshold of fifteen was selected, and a file consisting of 15,923 cited documents obtained. Co-citations among these documents were determined and the Jaccard-Sneath coefficients were calculated for each pair of cited documents. Clusters were formed at level .16 (16%), and a total of 16,001 clusters were formed, the largest cluster consisting of 117 cited documents.

These clusters were then used to retrieve 1973 source items processed by ISI for the *SCI* in 1973. About 25 per cent of the source items were retrieved. A higher fraction of the source items would be classifiable using a lower initial citation frequency threshhold.

Automatic indexing and classification is a goal which may never be completely attained, and creation of a system for classifying new documents may require some intervention of human judgement. The system described here is not entirely automatic because it requires human judgement to assign "headings" or labels to the groupings. This judgement is made on the basis of scanning titles with the aid of work pair frequency counts. This is, however, the only point at which human intervention is necessary.

Figure 6 shows a portion of the citing titles obtained for the cluster on hormone releasing hormones, and Figure 7 is a list of word pairs derived from these titles. The naming of the cluster can be readily done using both the titles and word pairs. The goal of an automatic classi-

Note: Solid connecting lines indicate co-citations that had appeared when this article was originally prepared. Broken lines indicate additional co-citations that appeared later in the year.

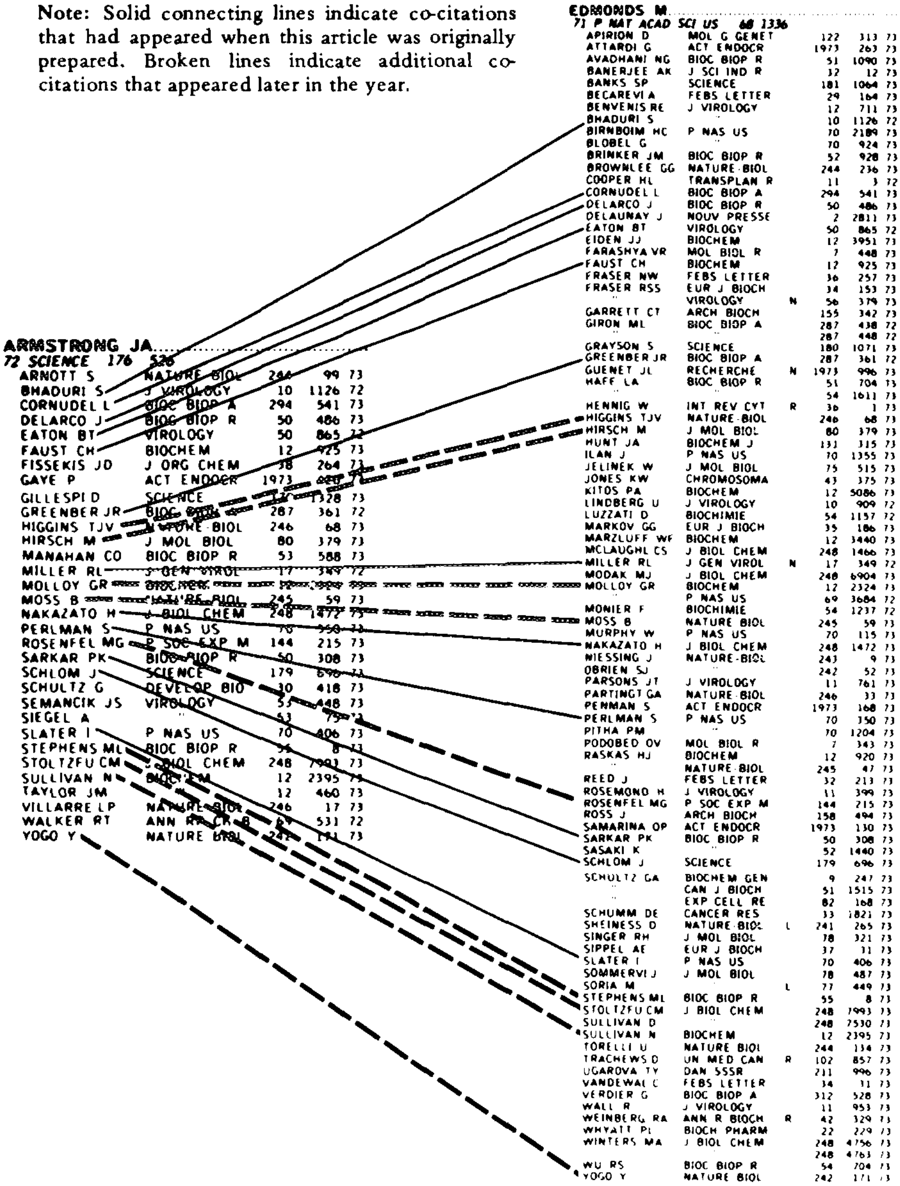


Figure 5. Co-citation of articles by J.A. Armstrong (*Science* 176:526, 1972) and M. Edmonds (*Proc. Nat. Acad. Sci. USA* 68:1336, 1971).

Frequency	Bibliographical Data
3	Abe K, Nagata N, Saito S, Tanaka K, Kaneko T, Shimizu N, & Yanaihar, N. Effects of synthetic luteinizing hormone-releasing hormone on plasma levels of luteinizing-hormone and follicle-stimulating hormone in man. *Endocrinol. Jap.* 19:77, 1972.
2	Akande EO, Carr PJ, Dutton A, Bonnar J, Corker CS, Mackinnon PC, & Robinson D. Effect of synthetic gonadotropin-releasing hormone in secondary amenorrhea. *Lancet* 2:112, 1972.
3	Akande EO, Carr PJ, Dutton A, Bonnar J, Corker CS, Mackinnon PC, & Robinson D. Effect of synthetic gonadotropin-releasing hormone in secondary amenorrhea. *Lancet* 11:112, 1972.
6	Amoss M, Blackwell R, & Guillemin R. Stimulation of ovulation in rabbit triggered by synthetic LRF. *J. Clin. Endocrin.* 34:434, 1972.
4	Amoss M, Rivier J, & Guillemin R. Release of gonadotropins by oral administration of synthetic LRF or a tripeptide fragment of LRF. *J. Clin. Endocrin.* 35:175, 1972.
5	Arimura A, Matsuo H, Baba Y, Debeljuk L, Sandow J, & Schally AV. Stimulation of release of LH by synthetic LH-RH in vivo. I. Comparative study of natural and synthetic hormones. *Endocrinology* 90:163, 1972.
7	Arimura A, Debeljuk L, & Schally AV. Stimulation of FSH release in vivo by prolonged infusion of synthetic LH-RH. *Endocrinology* 91:529, 1972.
3	Arimura A, Debeljuk L, Matsuo H, & Schally AV. Release of luteinizing-hormone by synthetic LH-releasing hormone in ewe and ram. *P. Soc. Exp. Biol. Med.* 139:851, 1972.
3	Besser GM, McNeilly AS, Anderson DC, Marshall JC, Harsoulis P, Hall R, Ormston BJ, Alexander L, & Collins WP. Hormonal responses to synthetic luteinizing-hormone and follicle stimulating hormone-releasing hormone in man. *Brit. Med. J.* 3:267, 1972.
8	Beyerman HC, Maat L, & Vanzon A. Synthesis of decapeptide sequence proposed for LH-releasing and FSH-releasing hormone. *Recueil. Trav. Chim.* 91:1239, 1972.
2	Bishop W, Fawcett CP, Krulich L, & McCann SM. Acute and chronic effects of hypothalamic lesions on release of FSH, LH and prolactin in intact and castrated rats. *Endocrinology* 91:643, 1972.
2	Bogdanove EM. Current knowledge of gonadotropin releasing factor(s). *Med. Coll. Va. Quart.* 8:5, 1972.
10	Borgeat P, Chavancy G, Dupont A, Labrie F, Arimura A, & Schally AV. Stimulation of adenosine 3'-5'-cyclic monophosphate accumulation in anterior-pituitary gland in vitro by synthetic luteinizing hormone-releasing hormone. *Proc. Nat. Acad. Sci. U.S.A.* 69:2677, 1972.
2	Borvendeg J, Hermann W, & Bajusz S. Ovulation induced by synthetic luteinizing-hormone releasing factor in androgen-sterilized female rats. *J. Endocrin.* 55:207, 1972.
6	Breton B, Weil C, Jalabert B, & Billard R. Reciprocal activity of hypothalamic factors of rams (ovis-aries) and teleostean fish on secretion in vitro of gonadotropin hormones C-HG and LH respectively by hypophysis of carps and rams. *Comptes Rendus Acad. Sci. D.* 274:2530, 1972.

Figure 6. Source titles associated with cluster 60. This is only a partial list.

Releasing	Factor	60	7	0	0	0	7	1.0
Synthetic	Hormone-releasing	60	0	5	0	1	7	2.8
Synthetic	Luteinizing-Hormone	60	5	0	0	1	7	2.1
LH	Hormone	60	0	0	3	2	9	5.4
Synthesis	Hormone	60	0	1	0	1	9	7.1
LH-Releasing	Hormone	60	9	1	0	0	10	1.1
Luteinizing-Hormone	Hormone	60	0	5	1	2	10	4.2
Synthetic	Releasing	60	1	6	2	1	11	2.5
Hormone	LH-RH	60	10	1	1	0	12	1.2
Hormone-Releasing	Hormone	60	10	0	0	0	12	1.7
LH	FSH	60	9	1	0	1	13	2.5
Luteinizing	Hormone-Releasing	60	12	0	0	0	13	1.3
Luteinizing-Hormone	Releasing	60	11	0	0	0	13	2.3
Luteinizing	Hormone	60	1	12	1	0	16	2.5
Releasing	Hormone	60	16	0	0	1	17	1.1
Synthetic	Hormone	60	0	7	7	2	19	3.3
Aromatic	Compounds	61	0	2	0	0	2	2.0
Electron-spin	Reactions	61	0	0	1	0	2	4.3
Electron-spin	Resonance	61	2	0	0	0	2	1.0
Electron-spin	Studies	61	0	1	0	1	2	3.0
Formation	Decay	61	1	1	0	0	2	1.5
Ions	Aqueous-solutions	61	2	0	0	0	2	1.0
Ketyl	Radicals	61	2	0	0	0	2	1.0
Pulse	Ions	61	0	0	1	1	2	3.5
Radiolysis	Ions	61	0	1	1	0	2	2.5
Resonance	Reactions	61	0	1	0	1	2	3.0
Resonance	Studies	61	1	0	1	0	2	2.0
Studies	Reactions	61	2	0	0	0	2	1.0
Pulse	Aqueous-solutions	61	0	1	0	1	3	3.6
Radiolysis	Aqueous-solutions	61	1	0	1	1	3	2.6
Pulse	Radiolysis	61	5	0	0	0	5	1.0
Amphetamine-induced	Behavior	62	0	1	1	0	2	2.5
Amphetamine-induced	Stereotyped	62	1	1	0	0	2	1.5
Apomorphine	L-Dopa	62	0	2	0	0	2	2.0
Apomorphine	Rats	62	0	2	0	0	2	2.0
Behavioral	Lesions	62	0	0	0	0	2	6.0
Behavioral	Rat	62	0	0	0	2	2	4.0
Central	Action	62	2	0	0	0	2	1.0
Central	Dopamine	62	0	1	1	0	2	2.5
Central	Dopaminergic	62	2	0	0	0	2	1.0
Central	Effect	62	2	0	0	0	2	1.0
Central	Mondamine	62	2	0	0	0	2	1.0
Central	Rats	62	0	0	1	0	2	4.5
Dopamine	Receptor	62	2	0	0	0	2	1.0
Dopamine	Receptors	62	2	0	0	0	2	1.0
Effect	Activity	62	0	0	1	1	2	3.5
Effect	Apomorphine	62	2	0	0	0	2	1.0
Effect	Locomotor	62	0	1	1	0	2	2.5
Effects	Central	62	0	1	0	1	2	3.0
Evidence	Dopamine	62	0	1	1	0	2	2.5
Evidence	Receptors	62	0	0	1	1	2	3.5
Induced	Rat	62	0	2	0	0	2	2.0
Locomotor	Activity	62	2	0	0	0	2	1.0
Model	Tardive	62	2	0	0	0	2	1.0
Mondamine	Neurons	62	2	0	0	0	2	1.0
Rat	Lesions	62	1	0	1	0	2	2.0
Amphetamine	Fat	62	1	1	0	1	3	2.3
Central	Neurons	62	0	2	0	1	3	2.6
Effect	Rats	62	0	0	2	0	3	3.6

Figure 7. Listing of word pairs for clusters 1972 Level 11 Fred 2.

fication system is to classify new source documents and, therefore, the test of this system is whenever the clusters obtained from the 1973 file are capable of classifying articles published in 1974. Our research is still proceeding, and in the following, I will outline the procedure which is being developed.

The complete list of cluster names and identifying numbers is maintained on one disc file. A second disc contains the cluster number and all the cited references contained in that cluster. As a new document is being entered into the ISI data base, it is possible to match the cited references in the document against the file containing clustered documents and associated cluster numbers. If a new source document contains one or more references which match the cluster file, one or more classification headings can be assigned to the source document.

Suppose, for example, that a particular source document contains five references, three of which cite documents in one cluster, and two which cite documents in another cluster. The source item would then be assigned two classification headings, one with a weight of three and the other with a weight of two. A test of the effectiveness of this method must involve a comparison of the results of this automatic classification procedure with manual indexing procedure performed on a sample of source documents. The system must also be tested in user studies, since a great deal will depend on how well we have identified and named the subject of each of the clusters. As with any system, we cannot hope to please every user, but rather to develop a system which will satisfy the needs of a maximum number of users for the minimum cost. The advantage of the automatic procedure described in this paper is that all manipulations, save the naming of the clusters, are totally automatic and require no human judgement.

Since the theme of this conference is the ordering of global information networks, it is appropriate that we discuss the connections between our citation clustering experiment and the need for a global classification scheme. The application of citation data in the creation of a classification scheme has the advantage of being closely geared to the international activity of the scientific community which have established these citation patterns through their publications. Since scientific specialties do not have national boundaries, we believe the citation approach is a fair procedure for identifying subject areas which are of interest to many different countries. Secondly, bibliographic citations themselves are an international language. Clusters of citations may, therefore, be named in any language, but the content remains defined by the cited documents. Hence, it is possible to envision a truly international classification scheme based on the *Science Citation Index* with subject experts in every country naming clusters according to that country's scientific usage. This may not really be necessary if English becomes the international language of science, but even if this does not occur, citation indexing still remains an indexing language which is essentially free of semantic or linguistic problems. Our main problem is in dealing with the variety of alphabets and symbol systems in Japanese, Chinese, Russian, etc. Such a system would go a long way towards improving worldwide exchange of information to the benefit of all countries involved.

1. Garfield E. *Science Citation Index;* a new dimension in indexing. *Science* 144:649-54, 1964.
2. Malin M V. The *Science Citation Index;* a new concept in indexing. *Library Trends* 16:374-87, 1968.
3. Weinstock M. "Citation indexes." In: *Encyclopedia of Library and Information Science,* 5 vols. (New York: Marcel Dekker, 1971), vol. 5, p. 16-40. Reprinted in *Current Contents* No. 25, 23 June 1971, p. M21-8; No. 26, 30 June 1971, p. M29-36; No. 27, 7 July 1971, p. 59-70.
4. Garfield E. Citation indexes for science. *Science* 122:108-11, 1955. Reprinted in: *Current Contents* No. 46, 18 November 1970, p. 46-51.
5. Salton G. Associative document retrieval techniques using bibliographic information. *J. Assoc. Computing Machinery* 10:440, 1963.
6. Garfield E. "Can citation indexing be automated?" In: *Statistical association methods for mechanized documentation, symposium proceedings, Washington, 1964,* ed. by M.E. Stevens et al. (National Bureau of Standards Miscellaneous Publication 269, 15 December 1965), pp. 189-92. Reprinted in: *Current Contents* No. 9, 4 March 1970, pp. 5-14.
7. -----------. Citation analysis as a tool in journal evaluation. *Science* 178:471-79, 1972. Reprinted in: *Current Contents* No. 6, 7 February 1973, p. 7-24.
8. Sparck-Jones K. Some thoughts on classification for retrieval. *J. Documentation* 26:89, 1970.
9. Small H & Griffith B C. The structure of scientific literatures. I. Identifying and graphing specialties. *Science Studies,* 4:17, 1974.