Information Processing & Management, Vol. 17, pp. 39–50, 1981
Printed in Great Britain

0306–4573/81/010039–12$02.00/0
Pergamon Press Ltd.

THE RELATIONSHIP OF INFORMATION SCIENCE TO THE SOCIAL SCIENCES: A CO-CITATION ANALYSIS†

HENRY SMALL
Institute for Scientific Information, 3501 Market Street, University City Science Center, Philadelphia, PA 19104, U.S.A.

(*Received* 3 *April* 1980)

Abstract—A co-citation cluster analysis of a three year (1975–77) cumulation of the *Social Sciences Citation Index* is described, and clusters of information science documents contained in this data-base are identified using a journal subset concentration measure. The internal structure of the information science clusters is analyzed in terms of co-citations among clusters, and external linkages to fields outside information science are explored. It is shown that clusters identified by the journal concentration method also cohere in a natural way through cluster co-citation. Conclusions are drawn regarding the relationship of information science to the social sciences, and suggestions are made on how these data might be used in planning an agenda for research in the field.

INTRODUCTION

One way of gaining insight into the state of a research field or discipline is to examine the publications produced by its practitioners. To the extent that practitioners in the field publish the results of their investigations, this mode for assessing the state of a field can reflect with great specificity the content and problem orientations of the group. Of the many ways that publications can be analyzed and counted, perhaps the most revealing kind of data are the references cited by the practitioner group in their publications. References to earlier literature tell us about the author making them as well as the items being cited. When references are cumulated over a significant volume of source literature such as in the *Science Citation Index* (*SCI*)® and *Social Sciences Citation Index* (*SSCI*)®, the collective patterns reveal the concerns of the field as symbolized by the documents and authors cited. This is how earlier cited literature can inform us on the current conceptual framework: the act of citing involves an association of a notion or idea expressed in the text with a cited document[1]. Hence, each reference is connected to a concept. The cumulative pattern of such contexts provides a representation of the cognitive structure of the research field.

The objective of a study currently underway at the Institute for Scientific Information is to examine the structure and development of the field of information science using the published literature of information science as data, and the techniques of citation analysis. We know very little about how the field of information science has developed over the past several years. On the one hand the field might be viewed from a technological standpoint, the primary accomplishments of which are creation of machine readable data bases and retrieval systems. From another perspective, however, the field of information science can be seen as an investigation into the nature of information, the theoretical basis for retrieval, the evaluation of retrieval, and the way that human beings use and transmit information. We expect that the literature of information science will reflect both the conceptual and technical concerns of the field. By using the statistical techniques of citation analysis we hope to get a picture of how the field has developed, the main lines of research in the field, its principal foci of interest, and where the field appears to be going.

Some studies have attempted to use published literature to arrive at insights into the structure of the field. Saracevic reviewed the first five volumes of the *Annual Review of Information Science and Technology* using bibliometric techniques and concluded that ARIST was biased toward the technology and practice of information science and against fundamental

†Research supported by NSF grant IST-78-16677.

research in the area[2]. In a breakdown of most cited authors in ARIST Saracevic found about 10% theoreticians, 40–50% experimenters, and the rest developers.

Donohue has provided the most ambitious bibliometric analysis of the "information science" literature, although his choice of journals to include in the database was somewhat idiosyncratic (e.g. inclusion of the *Journal of the Acoustical Society of America* as an information science journal)[3]. In general, the main corpus used in Donohue's study represented information science in its more technical, mathematical and engineering sense, rather than information science related to documentation, retrieval and library science. Therefore, the structure of "information science" in Donohue's study cannot be taken as a guide to what we expect to find in our study. (For example, the list of clusters Donohue obtained using bibliographic coupling among documents included groups designated as acoustics, computing, cybernetics, engineering, logic, numerical mathematics and statistics).

Salton has undertaken a citation analysis of individual researchers in information science using two comprehensive bibliographies and two points in time a decade apart[4]. His main conclusion is that the field has changed significantly over the decade and that this perhaps indicates not only intellectual ferment, but perhaps also a lack of focus on key problems. From Salton's study we would anticipate a lack of lasting theoretical orientations and a predominance of experimental work without the benefit of a theoretical framework.

A recent survey of concerns in the field of information science by Pratt, based on a qualitative analysis of the ARIST series which does not use bibliometric techniques, finds the major topics of current research to be[5]:

(1) Library problems
(2) Economics of information
(3) The nature of information
(4) Techniques of measurements.

All of the studies mentioned provide some hints of what we could expect in our citation analysis, but certainly none stands out as a definitive benchmark against which to compare our results.

Our study takes two different approaches to the analysis of information science as a field. First, we are using a special citation file extracted from the *Social Sciences Citation Index* database, consisting of a core set of journals in information science over a nine year period (1969–77), to explore the internal structure of information science and its development. The second approach is to show how information science links to the other fields and disciplines in the social sciences. To accomplish this we use existing cluster data files at ISI. It is this latter work which I report on in the present paper.

CLUSTER ANALYSIS OF THE SSCI

The starting point for this study was a cluster analysis of a special three year cumulation (1975–77) of the *Social Sciences Citation Index* (*SSCI*). Earlier we had performed a similar analysis of the *SSCI* for the period 1972–74[6, 7], and the new analysis of the 1975–77 file will allow us, eventually, to examine rates of specialty change within the social sciences. The results I will report here are only for the 1975–77 file and do not attempt to assess change over time.

The clustering procedure was identical to that used in the original study. I will review only the essentials here. First, all documents in the file cited ten or more times during the three year period are selected. Table 1 presents some statistics on the cluster analysis, and indicates that of the over two million cited items in the three year file, about 25,000 were cited ten or more times. This is a fairly weak criterion for selection when compared with our usual threshold of 15 citations per document per year for cluster analyses of annual SCI's[8]. Following selection of these highly cited items, all co-citations among the 25,000 were determined, that is, the number of times any pair of them is cited together in the three year period. As indicated in the Table there were 1.8 million unique pairs of co-cited items thus formed. The raw co-citation counts for each pair were normalized by dividing by the sum of citation frequencies for the two items minus the number of co-citations. This is essentially the fraction of citations to the two items that are co-citations, which is equivalent to the so-called Jaccard coefficient used in numerical taxonomy[9]. These coefficients were the basis for the cluster analysis. The clustering algorithm used (called single link clustering)[10] requires only that we specify a threshold for

Table 1. Statistics on clusters from 1975–77 Cumulative *Social Sciences Citation Index* (*SSCI*)

1. total citations (1975-1977)	3,399,058
2. distinct cited items	2,196,127
3. highly cited items ($\geq$ 10 citations)	24,954 (1.14%)
4. distinct co-cited pairs of cited items	1,846,585
5. distinct co-cited pairs at level 22%	10,418
6. clusters at level 22% ($\geq$ 2 cited items)	2,095
7. mean cited items per cluster	4.1
8. mean citing items per cluster	39.9

the normalized co-citation strength to generate a set of disjoint clusters containing the cited items. Setting this threshold at 0.22 generated about 2000 clusters each containing two or more cited items. The average cluster size was 4.1 cited items. This set of 2000 clusters in the social and behavioral sciences formed the universe from which we selected clusters on information science topics. Of course, information science is expected to represent only a small fraction of the clusters in this file, which is dominated by fields such as psychology (experimental and social), sociology, economics, psychiatry, and so on.

THE SELECTION OF INFORMATION SCIENCE CLUSTERS

The procedure used to select information science clusters from the 2000 1975–77 *SSCI* clusters was to define a set of information science journals which appear as source journals in the *SSCI*. Fifty journals were selected and are listed in Table 2. This list should not be regarded

Table 2. Information science journal subset

1.	American Archivist
2.	Annual Review of Information Science and Technology
3.	Aslib Proceedings
4.	Bulletin of the Copyright Society of the U.S.A.
5.	Bulletin of the Medical Library Association
6.	Canadian Journal of Information Science
7.	Canadian Library Journal
8.	College and Research Libraries
9.	Drexel Library Quarterly
10.	Government Publications Review
11.	IEEE Transactions on Information Theory
12.	IEEE Transactions on Engineering Management
13.	IEEE Transactions on Professional Communication
14.	Information and Control
15.	Information Processing & Management
16.	Information Sciences
17.	Information Scientist
18.	International Classification
19.	International Forum on Information and Documentation
20.	International Journal of Computer & Information Sciences
21.	Journal of the American Society for Information Science
22.	Journal of Chemical Information & Computer Sciences
23.	Journal of Documentation
24.	Journal of Education for Librarianship

Table 2 (*Contd*).

26.	Journal of Library Automation
27.	Journal of Library History Phiosophy & Comparative Librarianship
28.	Journal of the Patent Office Society
29.	Law Library Journal
30.	Library & Information Science
31.	Library Resources & Technical Services
32.	Library Trends
33.	Library Quaterly
34.	Libri
35.	Methods of Information in Medicine
36.	Nachrichten Für Dokumentation
37.	Nauchno-Teknischeskaya Informatsiya. Seriya 1. Organizatsiya I Metodika Informatsionnoi Raboty
38.	Nauchno-Teknicheskaya Informatsiya. Seriya 2. Informatsionnye Protessy I Sistemy
39.	On-line Review
40.	Proceedings of the American Society for Information Science
41.	Pattern Recognition
42.	Program-New of Computers in Libraries
43.	Review of Public Data Use
44.	Social Science Information
45.	Social Studies of Science
46.	Special Libraries
47.	Unesco Bulletin for Libraries
48.	Wilson Library Bulletin
49.	Zeitschrift für Bibliothekswesen Und Bibliographie
50.	Zentralblatt für Bibliothkswesen

as a definitive list of journals in the field, but rather as one way of defining the field, which is subject to an empirical test later in our analysis. It should be noted, for example, that computer science journals were intentionally not included in the list, and that we were slanting the list toward the library/information science direction. Some journals included were concerned with the mathematical study of communication (information), but the majority deal with the more traditional view of information science as an off-shoot of documentation.

The procedure to select clusters, similar to that used in an earlier study [7], was to calculate the fraction of source (citing) papers for each cluster which fall in the specified journal set. A distribution of these fractions is obtained (see Fig. 1) which ranges from clusters having 100% of their citing papers in journals which are members of the set, to clusters which have none of their citing papers in these journals. For purposes of comparison a similar distribution for the field of chemistry is included on the same graph (from *SCI* not *SSCI* cluster data). The chemistry distribution shows clearly a group of disciplinary clusters centered on about 80% concentration in the journal sub-set for chemistry, and a smaller interdisciplinary group of clusters centered at 55%. The up-swing of the distribution to the left shows all the clusters which are not, or only marginally, in the field. The information science distribution shows similar disciplinary and interdisciplinary peaks, though on a much smaller scale: there are only eleven clusters of the 2000 which have 30% or more of their citing papers in information science journals, and only 22 with 10% or more. The latter group is listed in Table 3.

CHARACTERISTICS OF THE INFORMATION SCIENCE CLUSTERS

Information science by our definition therefore comprises at most one percent of the

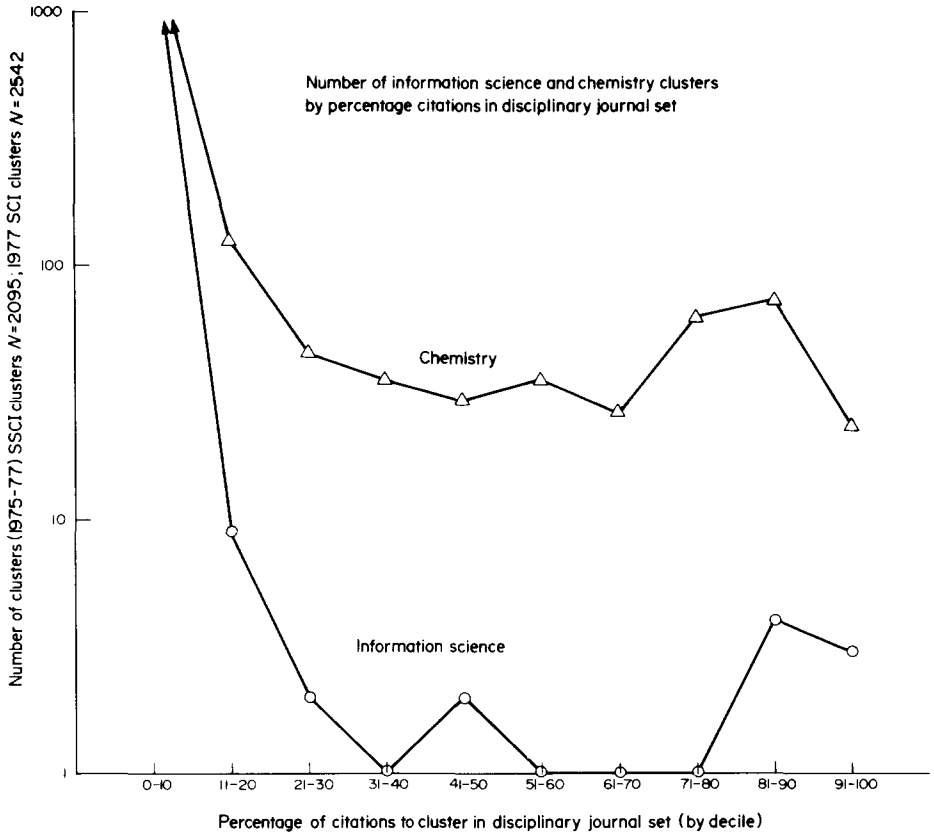

Fig. 1. Number of information science and chemistry clusters by percentage citations in disciplinary journal set.

clusters). Nevertheless, we can examine the clusters identified to see what they can tell us about the field. Referring to Table 3, each of the clusters with 10% or more of their citing papers in information science journals has been listed in descending order by percentage concentration. The order of clusters in Table 3 could be interpreted as the degree of disciplinary purity. This ranges from 100% for the "Precis" cluster to 14.3% for "copyright law" and "medical use of computers".

The cluster number in the left column is an arbitrary identification number assigned by the computer to each cluster as it is generated. An approximate name was given to each cluster based on an examination of the titles of the cited and citing documents. In the last two columns the size of the cluster is given both in terms of the number of cited items and the number of items citing them.

Table 4 lists all documents cited ten or more times which comprise the clusters having at least 40% of their citations from information science journals. The number of times each document was cited in the *SSCI* from 1975 to 1977 is indicated in the right-hand column. Of course, citations to an item can come from any source journal in the *SSCI* coverage, not just those journals listed in Table 2.

Most of the clusters are very small (the smallest number of cited items a cluster can have is two). In addition, a number of clusters have been given the same name (e.g. there are "on-line retrieval and data bases" clusters "a" and "b"). This redundancy occurs because it was not possible to distinguish the content of these clusters based on the titles of the papers, and they were therefore given the same name. This suggests that these areas are being fragmented at this level of association (22% normalized co-citation) and we will show in a moment that this is indeed the case. Both the small size of the clusters and their fragmentation tell us that information science has a weaker structure than areas such as psychology, sociology or economics which emerge as larger and more coherent specialties at this level. If the co-citation threshold had been lowered, the fragmented information science areas would have congealed, but for other subject areas composite macro clusters would have formed, indicating too low a

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.