Mapping of Science by Combined Co-Citation and Word Analysis. I. Structural Aspects

Robert R. Braam, Henk F. Moed, and Anthony F. J. van Raan
Center for Science and Technology Studies (CWTS), University of Leiden, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands

The claim that co-citation analysis is a useful tool to map subject-matter specialties of scientific research in a given period, is examined. A method has been developed using quantitative analysis of content-words related to publications in order to: (1) study *coherence of research topics* within sets of publications citing clusters, i.e., (part of) the "current work" of a specialty; (2) to study *differences in research topics* between sets of publications citing different clusters; and (3) to evaluate *recall* of "current work" publications concerning the specialties identified by co-citation analysis. Empirical support is found for the claim that co-citation analysis identifies indeed subject-matter specialties. However, different clusters may identify the same specialty, and results are far from complete concerning the identified "current work." These results are in accordance with the opinion of some experts in the fields. Low recall of co-citation analysis concerning the "current work" of specialties is shown to be related to the way in which researchers build their work on earlier publications: the "missed" publications equally build on very recent earlier work, but are less "consensual" and/or less "attentive" in their referencing practice. Evaluation of national research performance using co-citation analysis appears to be biased by this "incompleteness."

Introduction

Co-citation analysis is one of the major quantitative techniques in science studies to map the structure and dynamics of scientific research. This technique is claimed to be capable of identifying "research foci" and their relations, in particular at the level of research specialties (Small & Griffith, 1974; Griffith et al., 1974; Small, 1977; Garfield et al., 1977; Small & Crane, 1979). However, the validity and the practical usefulness of co-citation analysis for science-policy purposes is subject of recent studies and debates (ABRC, 1986; Healey et al., 1986; Hicks, 1987; Franklin, 1988; Hicks, 1988).

In this article we focus on the capability of co-citation analysis to map *structural aspects* of scientific research on the level of research specialties. For a discussion of *dynamical aspects* we refer to a related publication in this issue of *JASIS* (Braam et al., "Mapping II").

According to Price's theory of knowledge growth, scientific researchers constitute a "research front" by focusing their attention, as expressed by their references, to a small select part of the most recent literature (De Solla Price, 1965; Cozzens, 1985). Thus, a network of citation relations is created with relatively "high density areas" related to "research fronts." Price supposed that most papers are included in research front-subjects (De Solla Price, 1965).

Co-citation analysis is in fact an attempt to identify such "high density areas" in a citation network by clustering highly co-cited documents, thus indicating the existence of these research fronts. The citing literature of co-citation clusters, then, is considered to correspond to the group of publications that can be described as a subject-matter-specialty's published current work (Small & Griffith, 1974; Griffith et al., 1974). The cluster of co-cited documents is considered to represent the knowledge base of the specialty: the key concepts, methods, or experiments that researchers build on (Small, 1977, and 1978).

The question has been raised whether in this way the entire specialty, or only a subgroup of publications of the specialty is identified (e.g., Sullivan et al., 1977). Small, who introduced the co-citation technique (Small, 1973), claims that the citing authors of a cluster constitute a highly relevant subgroup of the current practioners of a specialty (Small, 1977). According to Rip (1988), only subgroups with "shared legitimatory tactics" are traced. In a recent review article on biblio-

This study is part of a project financed by the Ministry of Education and Sciences, through the Netherlands Advisory Council for Science Policy (RAWB).

Received February 24, 1989; revised October 4, 1989; accepted October 10, 1989.

metric indicators King (1987), sums up a number of objections against co-citation analysis: loss of relevant papers, inclusion of nonrelevant papers, overrepresentation of theoretical articles, time lag (between emergence of new specialties and capturing of them in a co-citation map), and subjectivity inherent in the setting of threshold levels, while these threshold levels strongly affect size and content of clusters. Furthermore, interpretation of the results is considered problematic: is indeed the cognitive structure of specialties of parts of these displayed, or the social structure of research (Callon et al., 1983; Rip & Courtial, 1984; Rip, 1988)? Others are much more sceptic, and maintain that clusters are mainly artifacts of the applied technique having no further identifiable significance (e.g., Oberski, 1988).

The co-citation cluster structure is constructed as follows. From the reference lists of a set of publications published within a given period, for instance a year, documents are selected that are cited more than a specified number of times (the citation threshold). Out of these cited documents, pairs are selected that co-occur relatively frequently in the reference lists of publications in the dataset, i.e., these pairs measure up to some specified co-citation strength threshold. Next, a special clustering routine, "single-linkage clustering," aggregates clusters of cited documents by sequentially linking together all selected pairs of cited documents that have at least one cited document in common. Next, for each cluster all publications are identified that cite one or more of the clustered cited documents.

We developed a combination of co-citation and word analysis in order to evaluate the nature and magnitude of some of the problems mentioned above and possibly improve the co-citation mapping technique.

In this combined approach, words originating from publications citing to documents in co-citation clusters, are analyzed in a quantitative fashion. Analysis of content words of those publications enables one to describe research topics involved in sets of publications citing documents in co-citation clusters, and to study the coherence within and difference between these sets of publications. Moreover, this combined analysis offers a possibility to evaluate completeness of the results of co-citation analysis, i.e., to determine the "recall" of this clustering technique.*

The approach is, in principle, suited to analyze all types of words (or phrases) related to publications, e.g., title words, abstract words, author names, addresses, and also words attributed to these publications such as indexing terms and classification codes.†

Underlying this combined approach is the basic notion that a scientific specialty can be regarded as "a coherent set of subject-related research problems and concepts upon which attention is focussed by a number of scientific researchers," irrespective the social and intellectual background of the researchers involved.‡ If different researchers work on the same set of subject-related research problems and concepts, one would expect that they use, to a relatively large extent, the same words for important concepts and problems in their specialty. To the extent that these researchers also concentrate their references on a small, selected part of recent earlier literature, i.e., to the extent that they participate in a "research front" (De Solla Price, 1965), and as far as publications are representative carriers of both words and references, results of techniques based on references, in this case co-citation analysis, should converge in some way with results based on the analysis of words. Such congruence, if present, indicates that "sharing a focus on a set of subject-related problems and concepts" goes together with "sharing a focus on intellectual base literature." Presence of such congruence, however, is not a necessity, and it remains to be seen,

*Frequency analysis of words from the titles or, later on, words from citation passages in citing publications, have been used by Small and his co-workers to characterize subject matter, in particular concept consensus, related to individual cited documents (Griffith & Small, 1974; Small, 1986). Such word-profiles, however, have never been used to describe the content (research topics) involved in the whole set of publications citing a cluster, i.e., the research topics involved in the "current work" of a specialty related to a co-citation cluster. For this latter purpose a sentence (in English) is used, based on frequently recurring phrases in the titles of citing documents (e.g., Small & Crane, 1979). Such procedure, however, requires detailed insight into the scientific content of the field, and is not very suited to analyze in a quantitative fashion the cognitive coherence within and difference between clusters, or completeness of clustering results. For such purpose, it seems better to use word-profiles, in this case aggregated lists of all nonunique words that occur in the set of publications citing documents in co-citation clusters.

†Indexing terms, classification codes, and title and abstract words are structural elements of a scientific article that are similar in the following two senses (Mullins et al., 1988). First, all these words (or phrases) are related to the content of a scientific paper. They serve the purpose of summarizing, abstracting, or classifying a papers content concerning subject matter. Second, these words are seen as identification markers that refer to the paper itself, rather than to its author(s). Thus, all these words may be suited to describe research topics involved in sets of scientific papers, and to analyze cognitive coherence within and resemblance between sets of publications, e.g., sets of publications citing co-citation clusters.

‡This specialty concept is based on the notion of science as essentially a problem-solving activity (Laudan, 1977). The claim that co-citation analysis identifies "specialist communities" in the paradigmatic sense (Kuhn, 1970, postscript; Small & Greenlee, 1980) is a stronger claim than the claim we are investigating in this study: the claim that co-citation clusters identify coherent research topics (problems, related concepts, and methods) involved in the set of publications citing these clusters. Nor is it assumed in this study that definition of problems and delimitation of cognitive and methodological resources finds place entirely within a "specialist community" (Callon et al., 1983). The question we try to answer is, whether or not groups of publications citing co-citation clusters are coherent concerning research topics studied, whether different clusters identify different research topics, and whether all publications in a dataset, relevant to the research topics identified by co-citation clusters, also cite clustered documents.

for example, whether all researchers working on a particular set of subject-related research problems indeed do share such an intellectual focus on base literature.

In the analysis presented in this article, the degree of convergence between these two aspects will be assessed empirically. Furthermore, in case of discrepancies, explanations will be suggested. Our working hypothesis holds that content words *do* reflect the research topics, i.e., studied problems and related concepts, involved in publications adequately, and the results of co-citation analysis will be discussed from this perspective. As such, we evaluate (claims of) co-citation analysis on the basis of an analysis of content words.

Although the *Science Citation Index* is the most important source of citation data, it is a limited source concerning "content" words, as only words from the titles of publications are available. This constraint is partially overcome by using information from different on-line databases in a combined fashion (Moed, 1988; Braam et al., 1988a). We say partially, as the success of such combination of databases is restricted by possible differences between these databases in coverage of the literature of the fields under study.

In the present study, indexing terms and classification codes have been extracted from the on-line versions of Chemical Abstracts (CA) and Biological Abstracts (BIOSIS) which were combined with data from the on-line version of ISI's *Science Citation Index* (SciSearch), using dedicated software (Moed, 1988). For related earlier work on this combined approach we refer to our previous publications (Braam et al, 1987, 1988a and 1988b).

Central issues of investigation discussed in this article are

(1) Is there sufficient amount of "cognitive coherence" within, and "cognitive difference" between, research topics involved in sets of publications citing co-citation clusters to justify the claim that indeed different research specialties are identified?

(2) If so, how complete then is the co-citation map? In particular, what is the maximum number of identifiable different research specialties in a specific dataset relative to the total number of specialties included in this dataset (recall of specialties)?

(3) How capable is co-citation analysis in covering at least a major share of the publications in the dataset that are relevant to the identified research topics (recall of sources relevant to these topics)?

(4) Does a higher-order clustering represent an "image" of higher-order structural hierarchies such as research (sub)fields, etcetera? The correlation of co-citation clusters with professional field classification codes is analysed to this end, because these field classification codes reflect a hierarchical level higher than the level of specialties.

(5) Does "incompleteness" concerning source publications relevant to identified research topics affect policy-relevant aspects of the results? In particular, we address the question whether co-citation analysis provides reliable results of output evaluations on a national level.

(6) Are the results of co-citation analysis recognizable and acceptable for researchers in the fields involved? In our opinion, legitimate practicle applications, particularly concerning science policy, of such quantitative techniques as co-citation analysis presuppose the recognized usefulness of results by researchers in the field(s) concerned. Results of this combined co-citation and word analysis were therefore also discussed with some experts in the field.

Empirical results of two case-studies will be presented and discussed in relation to the above mentioned problems concerning interpretation and application for science policy purposes. Further, conclusions will be drawn in this respect regarding the fruitfulness of combining co-citation and word analysis.

Data

We constructed datasets for agriculture-related biochemistry (*Chemical Abstracts,* 3400 source publications, 1985) and for chemoreception research (BIOSIS, 1384 publications, October 1985 to June 1986). The first dataset represents (a larger part of) a research field, the second a level between field and specialty.

In addition to bibliographic details, data on the cognitive content of the source publications involved have been collected, in particular indexing terms (controlled vocabulary as well as free terms) and field/subject classification codes.

Publications on agriculture-related biochemistry were selected on the basis of a set of about one hundred journals covering the field, and by using field classification codes, in this case eight relevant sections in *Chemical Abstracts* (CA). The distribution of publications in the dataset over the different sections is included at the bottom of Figure 3. Chemoreception publications were selected using a periodically published professional bibliography (Van der Starre, 1985 and 1986), excluding contributions to conferences and workshops. Selection from *Biological Abstracts* of publications for this bibliography is based on a set of keywords and classification codes. Citation data played no role in either selection of data.

For both datasets, reference lists of selected source publications were, as far as available, subtracted from SCISEARCH (the online version of the *Science Citation Index*). The different datasets were then combined using dedicated software (Moed, 1988). The new combined datasets cover 89% of the CA source publications and 75% of the original Chemoreception references, 3021 and 1033 publications respectively.

Methods

Co-Citation Cluster Analysis

In co-citation analysis, citation and co-citation strength thresholds are used to discriminate between

significant and insignificant (co-)cited documents. Optimal choice of these thresholds has been studied earlier (Braam et al., 1988a). Co-citation analysis has been performed for several combinations of thresholds and results were compared. As a first approach, the combination of thresholds generating the largest number of co-citation clusters has been used. It was assumed that in this way the number of different research topics related to clusters would be maximized.

Several enhancements recently implemented in the ISI clustering algorithm, such as fractional citation counting, variable level clustering, and iterative clustering of clusters (Small & Sweeney, 1985, Small et al., 1985), have not been applied by us. These alterations are mainly of importance when analyzing multidisciplinary data, in order to account for differences across fields in citation rates (Small & Garfield, 1986), while our datasets do not exceed the level of fields.

Presentation of the clustering results in a map is a problem in itself. We applied, for the moment, a graph theoretical approach to draw a map of clusters and their interrelations. The individual clusters are depicted by circles and the interrelations between clusters are indicated by lines between the circles. Hence, the exact position on the map and the distance between the clusters are not meaningful.

Description of Research Topics of Current Work by Word Analysis: Construction of Word-Profiles

An indication of the research topics involved in individual publications can be given by constructing a *publication "word-profile,"* i.e., a list of content words related to a publication.

If publications sharing citations to documents within the same co-citation cluster indeed represent the current work of a research specialty, then these (citing) publications are cognitively related and, as a consequence, are expected to contain, and to be indexed with, primarily the same content-related words. Thus, topics involved in a particular research specialty can be indicated by aggregating and listing these words, together with their frequency of occurrence, for the set of citing publications of each cluster. In this way a *cluster "word-profile"* can be constructed that represents the research topics involved in the current work of a specialty indicated by the cluster. For the present study we applied a frequency analysis of *indexing terms* and *classification codes* occuring in publications citing the various clusters (Fig. 1).

It should be noted that these words are associated with the *citing* publications and *not* with the *cited* documents, i.e., these words are associated with the "current research" in a specialty rather than with its "intellectual base."

In order to exclude *isolated aspects* of research, only words occurring in more than one citing publication per cluster have been listed for each cluster. Further, we distinguish between "central" and "peripheral" citing publications (Fig. 1). Source publications citing exclusively to one cluster are called "central," since they probably better represent the specific character of topics involved in the clusters. Source publications citing to several different clusters, are called "peripheral sources." These peripheral publications emphasize similarity between clusters, as their terms are included in the word-profiles of several different clusters. In order to maximize differences between clusters, only indexing terms from central publications are analyzed.

In order to improve the interpretability of co-citation maps, cluster word-profiles can be printed in the co-citation map near the clusters they belong to (see for example the map presented in Figure 4).

Evaluation of Coherence Within and Difference between Sets of Publications Citing Clusters

Analysis of Word-Profile Similarities. Using word-profiles to describe research topics involved in (sets of) publications, a quantitative analysis of coherence within and difference between publication groups becomes possible. For this purpose we make use of the concept of similarity, as it is developed in information retrieval.

Similarity measures yield an indication of the relevance of an object (a document, in this case a publication word-profile) to a given standard (the query, in this case a cluster word-profile). When both object and query are represented as collections of "weighted" terms (in our case word-profiles), several conventional information retrieval similarity measures can be used to establish this relevance (Jones & Furnas, 1987). One of these similarity measures, the cosine formula, also used in co-citation analysis, has been used in this study. The similarity between an object (O) and a query (Q) is then defined as

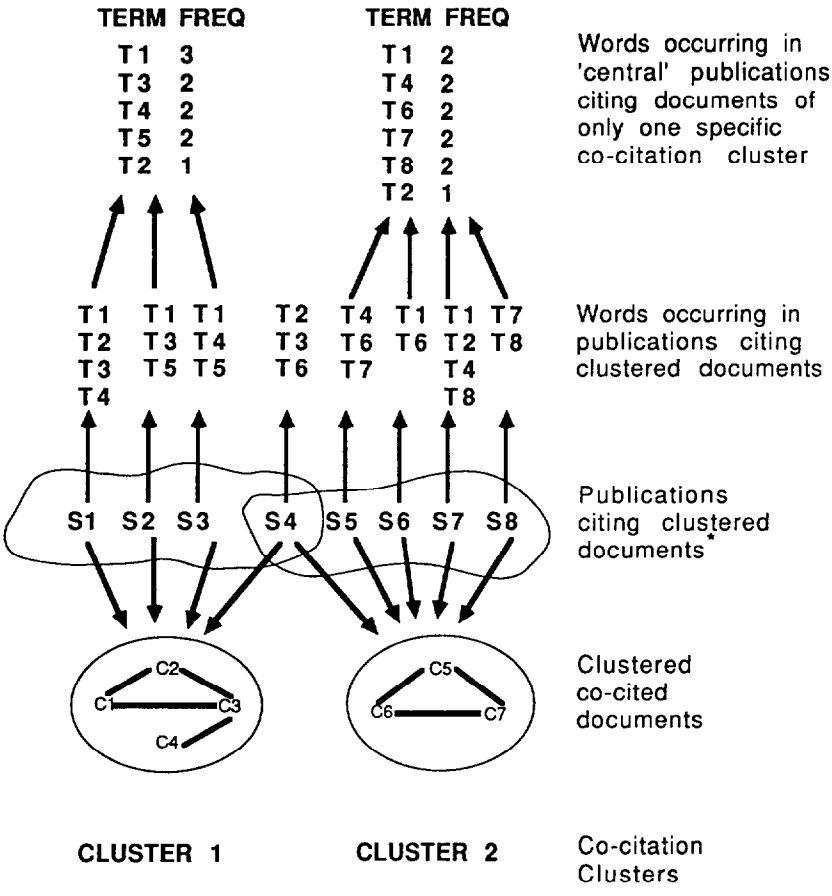
$$\mathrm{Sim}(O,Q) = \frac{\sum_{i=1}^{n} W(Oi) * W(Qi)}{\sqrt{\sum_{k=1}^{n} (W(Ok))^2} * \sqrt{\sum_{j=1}^{n} (W(Qj))^2}} \qquad (1)$$

$W(Oi)$ = weight of object term i, in the boolean case ($W(\mathrm{Oi}) = 0$ or 1;

$W(Qi)$ = weight of query term i, in the boolean case $W(Qi) = 0$ or 1;

n = total number of terms.

In the Boolean case, the cosine formula expresses the relative number of terms in the intersection of the set of terms associated with the query, and the set of terms associated with the object. The cosine formula normalizes for the length of the word-profiles of both object and query. Thus, objects with long word-profiles



* Publications citing exclusively to documents of one cluster are called 'central', publications citing documents of more than one cluster are called 'peripheral'.

FIG. 1. Method to describe research topics involved in publications citing documents in co-citation clusters.

can be "penalized" for their "representational richness" if this does not correspond to a richness in the query's representation (Jones & Furnas, 1987).

The similarity of each individual publication word-profile to the aggregated whole (cluster word-profile) is used to measure the "coherence" within a cluster, i.e., the extent to which the publications citing to a cluster share research topics. This coherence is computed for each cluster as the average value of the similarity to the aggregated whole, applying the cosine formula for the boolean case, for all source publications citing the cluster

$$Coh(Ck) = \frac{1}{m} \sum_{i=1}^{m} \mathrm{Sim}(O,Q) \qquad (2)$$

Ck = coherence within a co-citation cluster with index number k;
m = number of publications (objects) citing to cluster k;

This coherence attains a maximum value of 1 when all source publications have exactly the same terms, and reaches a minimal value close to 0 when all source publications have exclusively different terms. In order to obtain a minimal value of zero, terms with a frequency of one should first be excluded from the query word-profile (in this way isolated aspects of research are excluded). This procedure is only of importance if very few citing publications are involved in clusters. For large numbers of publications, it will not make much difference whether or not terms with a frequency of 1 are included in the query word-profile.

The amount of similarity *between* clusters can be established, using formula (2), in the following three ways. First, by comparing terms of sources citing cluster 1 (objects) with terms representing cluster 2 (query). This procedure is not symmetric, i.e., similarity of cluster 1 as defined by its objects to cluster 2 as defined by the query may be different from the similarity of cluster 2 as defined by its objects to cluster 1 as defined by its query. Such procedure is suited to analyze for example whether publications citing a particular cluster are on the average more similar, and thus most relevant, to the research topics related to this cluster than to research topics related to other clusters not cited by these publi-

cations. Second, to establish a measure of similarity between two clusters, is to regard both clusters as queries and compute a similarity value for cluster 1 as a query with cluster 2 as a query. This procedure results in a symmetric similarity measure. Third, to establish a similarity measure between clusters would be to compute similarity values, using formula (1), for all objects (citing publications) of cluster 1 to all objects of cluster 2, and to calculate an average value of these object-object similarities over all possible cases (Formula 2, with m equal to the number of all possible object-object combinations). This procedure also results in a symmetric similarity value.

To evaluate whether different clusters represent also different topics, it seems most suited use the second way, i.e., to represent both clusters as queries, and to compute the similarity between them. This way, we obtain an indication of the similarity between topics involved in the clusters, apart from their respective coherence. The difference, then, between two clusters is the complement of their similarity:

$$\text{Dif}(C1, C2) = 1 - \text{Sim}(Q1, Q2) \qquad (3)$$

with $\text{Sim}(Q1, Q2)$ being the similarity between queries for cluster 1 and cluster 2.

In order to examine whether the amount of coherence within clusters is sufficient to justify, in a quantitative sense, the claim that clusters indeed represent research specialties (i.e., a set of cognitively related publications), this value should be related to a value for publications outside clusters. We proceed as follows. If selection of publications would be random, publications involved in a cluster are not expected to be more similar to the "topics" involved in the cluster than any other publication outside the cluster (assuming exclusion of isolated aspects, i.e., terms with a frequency of 1, from these cluster "topics"). Thus, a comparison of the average similarity to cluster "topics" for publications citing a cluster with publications not citing that cluster might be applied to evaluate the importance of the computed coherence. The magnitude and significance of the difference between the two above averages indicate whether the amount of coherence is sufficient to justify the claim that clusters represent research specialties.

The average strength of pairwise co-citation relations within clusters is another, and completely independent from the above measure, indication of the coherence of sources in a cluster. Mean values of the strength within clusters have been computed excluding pairs with zero strength. Pairs of co-cited publications not reaching the specified co-citation threshold, interrelate clusters in as far as their elements occur (in other pairs) in clusters. The strength of these interrelations is computed as the mean strength of such pairs for each combination of clusters, and indicates resemblance between them in cognitive content. Comparison of these two different, and independent, measures of coherence within and resemblance between clusters is useful, in order to see whether citing practices and use of words are indeed related characteristics of research specialties.

Discussion of Results with an Expert in the Field. For both areas studied, results were discussed with an expert with a broad overview of the particular (sub)-field. Maps of interrelated clusters were sent to these experts, together with information describing research topics involved in publications citing clusters, (word-profiles of indexing terms, classification codes), and some additional information on sources and addresses (journal titles and country names or full addresses) of the publications involved. For each cluster, coherence of cognitive information was discussed as well as recognizability of research topics known by the expert, and interpretation of interrelations between clusters. Furthermore, the expert was asked to name, if possible, the research topics involved in the clusters and to give comments on each topic. The main questions asked were:

(1) Are the word-profiles used to describe the cognitive content of clusters coherent?
(2) Do clusters represent specific research topics?
(3) Do these topics differ reasonably well among each other?
(4) Do interrelated clusters represent specific research areas (i.e., cognitively interrelated research topics)?
(5) Is the information used to describe the cognitive content of the clusters in this study adequate for this purpose?

Results of these discussions are presented in a later section.

Evaluation of Completeness of Co-Citation Analysis, by Means of Word Analysis

Completeness in Terms of the Number of Identified Topics. The question, to what extent co-citation clusters "cover" the (sub)field under study can not be answered without reference to some external "correct" overview of this field. As co-citation analysis is meant to create such overview, as an alternative view, the best one can do is to maximize the number of distinct identifiable topics. This second problem can be investigated by comparing results (cluster word-profiles) of co-citation clustering generated for different citation and co-citation threshold sets. For example, when low co-citation threshold values are used, topics may be identified in clusters not present at higher threshold values. In this study we do not present any empirical results on this subject, but only indicate a possible way to explore such questions quantitatively.

One could also compare results of alternative clustering procedures (e.g., "complete-linkage clustering" versus "single-linkage clustering"), or results based on clustering of different aspects of publications (e.g., citations versus words). Word similarity analysis, then, could be used to investigate whether these different

procedures identify cluster topics not covered by co-citation clusters. Results of a comparison of clustering procedures based on different aspects, in a particular co-citations versus co-words, will be published elsewhere (Braam et al., 1989).

In the present study we investigate clustering results of co-citation analysis using a fixed co-citation strength threshold level and single-linkage clustering. The criterion for the threshold level was maximization of the number of clusters.

Retrieval of Sources Relevant to Topics Involved in Clusters (recall). Another type of completeness concerns the number of publications relevant to identified cluster topics, i.e., relevancy of publications to research topics involved in the set of publications citing documents in co-citation clusters. Possibly not all relevant publications cite the clusters involved. In the same manner as described above in an earlier section, word similarity between publications and cluster topics are computed for all publications in the database. It is then established how many source publications in the dataset have high cognitive resemblance to identified research topics, but are *not* citing the involved clusters. The critical similarity value is based on the distribution of similarity values for publications citing clusters. When similarity values occur for source publications not citing a specific cluster, that are higher than, say, the median of the similarity distribution for source publications citing this cluster, then it seems reasonable to consider these former source publications as highly resembling the publications citing that cluster, and thus as relevant to these research topics.

Another, and additional way to establish a measure of relevance to research topics for publications involved in the set of publications citing clusters, is the computation of a type of Inclusion-index that expresses, in this particular case, the number of terms a source publication has in common with the word-profile representing a "cluster topic", relative to the total number of terms of the source publication*

$$\mathrm{Inc}(O,Q) = \frac{\sum_{i=1}^{n} W(Oi) * W(Qi)}{\sum_{k=1}^{n} W(Ok)} \qquad (4)$$

$W(Oi)$ = weight of object term i, in the boolean case $W(Oi) = 0$ or 1;
$W(Qi)$ = weight of query term i, in the boolean case $W(Qi) = 0$ or 1;
n = total number of terms.

A source publication can have all its terms in common with a "cluster topic" (Inclusion-index = 1.0), though, due to a difference in number of terms, similarity according to the cosine formula (Formula 1) is low. Such source publications can be seen as dealing with *aspects* of research topics involved in publications citing this particular cluster, and, as a consequence, such publications should be considered relevant to these research topics.

The retrieval effectiveness of co-citation analysis, can now be expressed by the ratio of the number of source publications citing clusters, and the total number of source publications relevant to research topics involved in these clusters ("recall of publications").

Analysis of Relative Contributions of Countries to Research Topics

If co-citation analysis indeed enables us to display (important aspects of) the cognitive structure of scientific research, then it might be also useful in exploring the research activity of a country or institute in that particular field, subfield, or specialty. For instance, we could think of an analysis of "strengths and weaknesses" of national performance by computing a country's or institute's share in the publications involved in clusters (Mombers et al., 1985). We compared the number of contributions of countries to the *citing literature* of clusters, with those for all literature related to clusters, either by citations or by word similarity. In this way the effect of "incompleteness" of co-citation analysis on, for instance, ranking of countries can be revealed.

Results and Discussion

The Case of Agriculture-related Biochemical Research

First, some comments will be made on the type of indexing terms used to describe research topics involved in sets of publications citing clusters. In an earlier study (Braam et al., 1987) we found that *controlled* terms discriminate better between interrelated clusters (using Salton's formula to compute the co-citation strength) than between individual clusters, implying that these terms probably relate to a higher hierarchical level than the level of individual clusters. Other terms, such as keywords, are perhaps more useful to describe the specific content of individual clusters. For this reason *uncontrolled* terms of *Chemical Abstracts* (CA-keywords) were also taken into account. These uncontrolled terms originate from titles and abstracts, while the controlled terms originate from a subject thesaurus. Thus, CA-keywords are more text-specific then controlled terms. But, as both types of terms are selected by indexers of *Chemical Abstracts*, neither directly reflects the opinion and preferences of the authors themselves. CA-keywords are often arranged in phrases of four or more words. We did not use the full keyword-phrases, but counted word-pairs occurring

*Such inclusion index could also be used to inspect inclusion of research topics related to one cluster into topics related to other clusters. In this way, subspecialties can possibly be identified.

within these phrases in order to avoid very low frequencies. In addition to the above two types of indexing terms, we analyzed the distribution of CA-sections over the clusters.

In the clustering routine the set of thresholds was chosen to yield a maximum number of 39 clusters within a sensible range of threshold values. For detailed information we refer to the above mentioned earlier publication. The numbers of publications related to these clusters are in the range of 2–12 cited documents, and 6–46 citing publications of which 2–29 publications are "central." In Table 1 results are presented of cognitive-coherence analysis, i.e., average similarity of individual publications to word-profiles representing research topics related to the clusters these publications are citing, both for controlled terms and keywords. Table 1 also displays the average value of the similarity to these same word-profiles for publications not citing any cluster. Average similarity for publicatons citing clusters (coherence values summed over all clusters) appears to be quite high compared to average similarity to these word-profiles (representing the involved research topics) for publications that do not cite to any cluster. For example, keyword word-profiles of individual source publications citing clusters have an average similarity of 0.34 to the cluster's word-profiles, against an average similarity of 0.10 for all other source publications. Results presented in Table 1 offer empirical support to the claim that clusters represent research specialties. Furthermore, it is shown that central sources, on the average, are somewhat more similar to cluster topics than peripheral sources (Table 1). This indicates that central sources are indeed more specific for the topics involved in clusters than peripheral sources. Also, we find a positive and significant correlation (though weak for controlled terms) between coherence of clusters in terms of words and coherence in terms of citations, i.e., average word-similarity versus average co-citation strength within clusters ($r = 0.7$ at level 0.0001 for uncontrolled terms, and $r = 0.3$ at level 0.09 for controlled terms). This indicates that a relation exists between citing practices and the use of indexing terms (words) *within* research specialties.

As far as relations *between* clusters concerns, one could ask whether different clusters with high word-similarity between their respective word-profiles represent parts of one larger specialty. For example, as shown in Figure 2, clusters with index number 11, 13, 30, and 36 are strongly related by word-similarity for controlled terms, and the same holds for cluster 8 and cluster 15. For a number of clusters there seems to be no convergence of citing practices and use of indexing terms with respect to relations *between* them, e.g., cluster number 33 has other, and much more, word-similarity relations to other clusters than one would expect from its co-citation relations (Fig. 2). For each cluster, we compared the average similarity to the cluster's word-profile of all source publications citing the cluster to the average similarity for the same publications to the word-profiles of all other clusters. The conclusion is that, in most cases, source publications are, on the average, most similar, and consequently most relevant, to the topics involved in the clusters they cite. Differences appeared to be larger for keywords than for controlled terms, indicating a more specific character of keywords.

A further important finding is a correlation, in particular at the level of "super-clusters," between co-citation clusters and field-classification codes. Almost all clusters are dominated by one single classification code (i.e., this code covers 75% or more of the source publications citing the cluster), while clusters interrelated by co-citations share the same dominating code. For example, Figure 3 shows clusters with index number 32, 33, 35, and 38 are interrelated by co-citations,

TABLE 1. Average word similarity of source publications to topics involved in clusters for datasets on agriculture-related biochemistry and chemoreception research.

		Source Publications Citing to Co-citation Clusters						Sources Not Citing to Any Co-citation Cluster		
		Central Sources			Peripheral Sources					
Dataset	Words Used	N^a	Mean	Std	N^a	Mean	Std	N^a	Mean	Std
Agriculture-related biochemistry	Controlled terms from *Chem. Abs.*	414	0.36	0.14	64	0.25	0.10	2920	0.15	0.07
	Keywords from *Chem. Abs.*	414	0.34	0.16	64	0.16	0.13	2920	0.10	0.07
Chemo-reception research	Supplementary terms from BIOSIS	167	0.44	0.17	54	0.43	0.15	1163	0.16	0.09

[a]N = number of source publications involved in analysis of word similarity. Mean = Average word similarity as measured by Salton's cosine formula (boolean case) of source publications to topics involved in co-citation clusters. Std = standard deviation from the mean for word similarity.

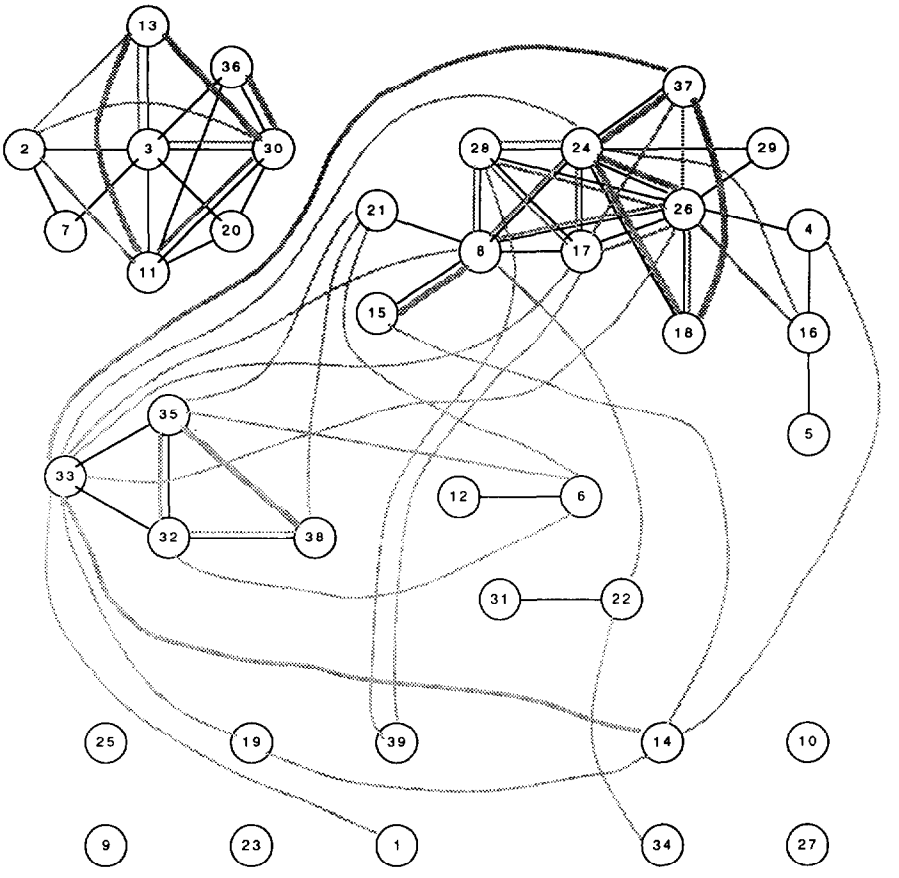
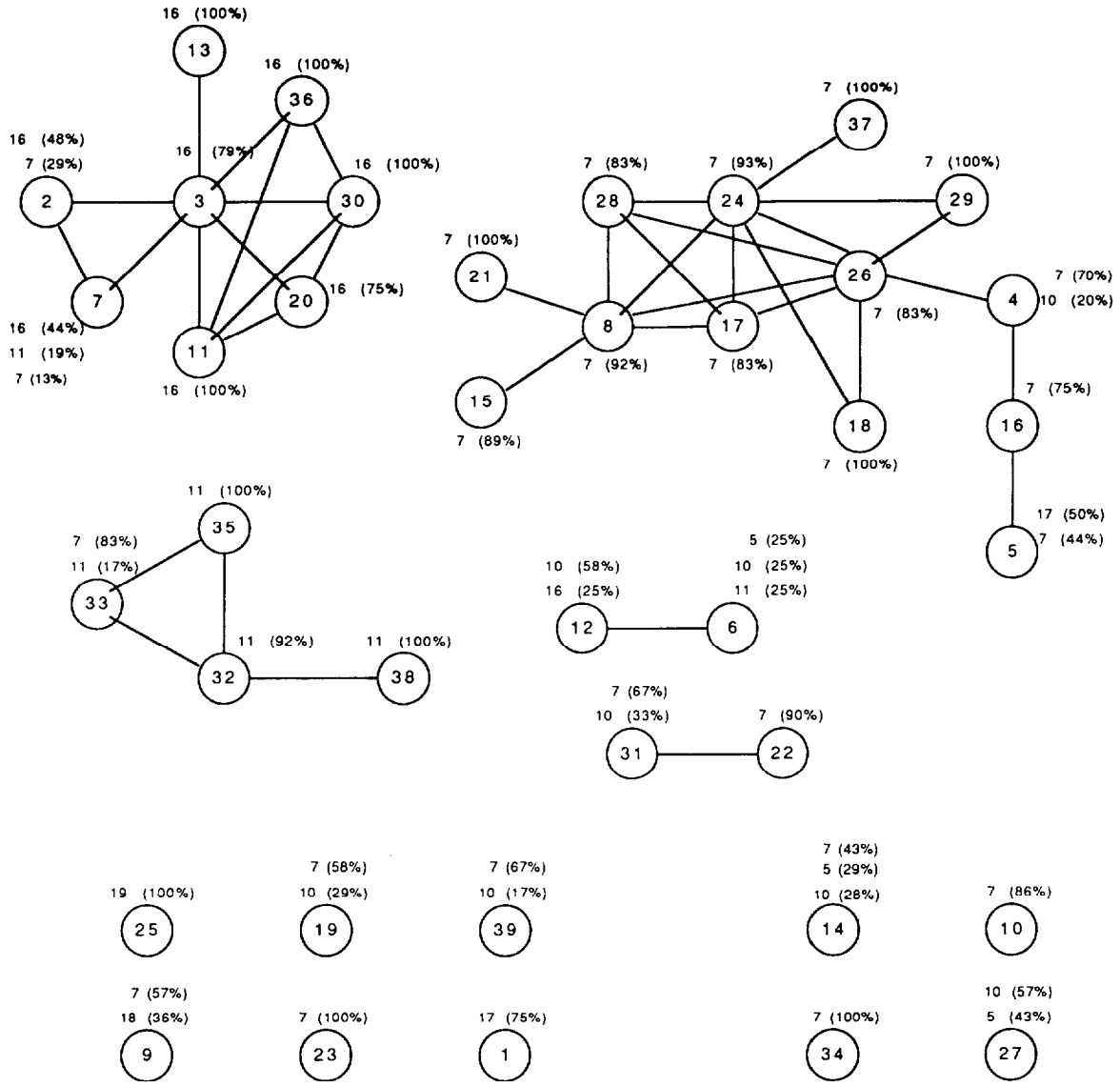


FIG. 2. Strength and word-profile similarity (controlled terms) relations between co-citation clusters, for a dataset of agriculture-related biochemistry literature.

and are also classified mainly (except for 33) in the CA-Section "Plant Biochemistry." This indicates that a higher-order structural hierarchy, in this case subfields, is also represented by the cluster structure.

In a discussion of the results with an expert, results were, in general, considered meaningful, both regarding the identification of research topics as well as their interrelations. Only a small number (13%) of the clusters was considered to be internally not cognitively coherent. All other clusters represent, according to the expert, typical research topics, although the character varies from very specific to quite broad. Probably, of these broad topics few publications are included in our dataset. For some clusters (18% of all clusters in the map), the absence or presence of interrelations with other clusters was considered to be incorrect, in other cases (13%) interrelations were considered to be only partially correct. Some interrelations were considered to be typically methodological by nature. The expert could easily identify the interrelated sets of clusters as representing different major research areas (subfields), such as "Biotechnology oriented towards fermentation," "Biological regulation by hormones," and "Photosynthesis." A number of different clusters, interrelated by co-citation below the applied threshold, were considered to represent one and the same research topic (e.g., clusters with index numbers 17, 24, and 37 are in the experts opinion all on beta-adrenergic receptors, while cluster numbers 18, 26, and 28 are all on alpha-adrenergic receptors). It was not clear to the expert why, in these cases, more clusters showed up. Other clusters, also interrelated by co-citations below the applied threshold, however, could be distinguished on cognitive grounds by a difference in emphasis on some aspects, in the re-



CA SECTIONS		SHARE IN DATASET
7	EMZYMES	29 %
11	PLANT BIOCHEMISTRY	15 %
10	MICROBIAL BIOCHEMISTRY	14 %
17	FOOD AND FEED CHEMISTRY	12 %
16	FERMENTATION AND BIO-INDUSTRIAL CHEMISTRY	9 %
5	AGROCHEMICAL BIOREGULATIRS	9 %
18	ANIMAL NUTRITION	6 %
19	FERTILIZERS, SOILS, AND PLANT NUTRITION	6 %
	TOTAL	100 %

FIG. 3. *Chemical Abstracts* sections covering 75% or more of "central" publications citing co-citation clusters, for a dataset on agriculture-related biochemistry.

search concerned, as indicated by some specific keywords or controlled terms.

In the discussion of results, the expert used controlled terms and keywords spontaneously in combination, so it may be fruitful to use both types of terms also together in word-profile similarity analysis.

Merely 14% (480) of all source publications in the dataset contain citations to clustered documents, i.e., 86% (2920) of the sources do not contain citations to any cluster, and are thus "lost" in the analysis! Just a small part of this "loss" is caused by the fact that a number of publications contained in *Chemical Ab-*

stracts are not present in the SCI (11%), i.e., a loss caused by a difference in scientific literature coverage between these two databases.

In Table 2 results of word-similarity analysis are shown for the field of agriculture related biochemistry. Results for both controlled terms and word-pairs from keywords, clearly indicate that a large number (883) of the "lost" source publications (containing no citations to any cluster) still are cognitively related to the identified topics involved in clusters. About half of these "lost" source publications (415) is even highly similar (see third section) to topics involved in one (32%) or more (13%) clusters, while the other half (468) is dealing with specific aspects relevant to the topics involved in one (15%) or more (40%) clusters. These figures are high compared to those for source publications related to clusters by citations. The number of source publications that are central (i.e., exclusively related to one single cluster) to the identified topics is about the same for both "citing" and "similar" sources. In Table 3 it is shown that the ISI-coverage for these "similar" sources is not much lower than the overall coverage. Our empirical results therefore indicate that co-citation analysis, though probably appropriate to identify important research topics, is indeed not appropriate to select all, or even a considerable amount, of the releva t sources in these topics. A further discussion is given in a later section. Now, if we add to the set of publications citing the co-citation clusters all the source publications not citing these clusters but having a high word-similarity to cluster topics, the above retrieval problem might be overcome. This is important in particular for evaluation purposes e.g., the participation of a country in specific clusters. The effect of this addition of source publications to clusters on a country's participation in clusters and its overall ranking is illustrated in Tables 4 and 6. Both aspects are influenced, but it is remarkable that this effect is not equally distributed over the different countries. Thus, it can be concluded that results of co-citation analysis may offer a distorted picture of a country's participation in research topics.

The Case of Chemoreception Research

Co-citation clustering has been performed for chemoreception research publications (see second section) using a threshold set for which, compared to other sets, a maximum number of 38 clusters occurred. The numbers of publications related to these clusters are in the range 2–18 for cited documents, and between 4–17 citing publications of which 10 or less are "central" to a cluster. In Figure 4 a map of chemoreception research is shown. Here, clusters are depicted by circles with an index number (these index numbers only relate to our database). Co-citation strength within and between clusters is indicated by dots and by lines, respectively. Lines indicate co-citation relations below the applied strength threshold, dots indicate co-citation relations above this threshold. The words ("supplementary terms" from BIOSIS) we used to describe research topics, are listed near the clusters concerned. These words are document title words and added words. The latter are terms selected by analysts of BIOSIS (BioSciences Information Service, Philadelphia) in order to enhance and clearify the author's meaning.

Average wo d-profile similarity *within* clusters (coherence) is remarkably high for source publications citing clusters, as compared to average similarity to cluster topics for source publications not citing any cluster (Table 1). This finding shows that co-citation analysis selects publications that are cognitively more related then expected from random selection of publications.

As in the case of agriculture-related biochemical research, we also find a positive and significant, though again weak, correlation ($r = 0.45$ at level 0.025) between average word-similarity and average co-citation strength within clusters. Thus, also for this subfield of

TABLE 2. Types of relations between source publications and co-citation cluster, for datasets on agriculture-related biochemistry and chemoreception research.

	Source Publications Citing Clusters (and thus constituting cluster topics)				Source Publications with High Word-Similarity to Cluster Topics, But Not Citing Clusters[a]				All Source Publications In Dataset		Source Publications Citing Clusters as Percentage of All Publications Related to Clusters (Recall)	
Dataset	All	(%)	Central[b]	(%)	All	(%)	Central	(%)	Total	(%)	All	Central
Agriculture-related biochemistry	480	(14)	414	(12)	883	(26)	399	(12)	3400	(100)	35%	51%
Chemoreception research	221	(16)	167	(12)	99	(7)	54	(4)	1384	(100)	69%	76%

[a]Similarity higher than average similarity within clusters or 100% inclusion.
[b]Source publications related exclusively to one cluster.

TABLE 3. ISI-coverage of source publications per type of relation to clusters, for datasets on agriculture-related biochemistry and chemoreception research.

Dataset	Source Publications Citing Clusters	Source Publications Not Citing to, but Similar to Topics Involved in Clusters	All Source Publications
Agriculture-related biochemistry	100%	84%	89%
Chemoreception research	100%	85%	75%

science there seems to be a convergence of citing practices and use of words for important concepts and problems in the specialty.

With respect to relations between clusters, we find that some clusters are strongly interrelated by co-citations and also have highly similar word-profiles (e.g., clusters 9 and 28 in Figure 5). These clusters probably represent parts of one larger research specialty. However, there are also cases where co-citation relations between clusters are absent, although these clusters have highly similar word-profiles (e.g., clusters 9 and 26 in Figure 5). In these cases there is no shared intellectual focus on earlier literature that corresponds to similarity in research problems. But according to the definition of specialties as "sets of related research problems and concepts studied by a number of researchers," these latter clusters should also be seen as parts of one larger specialty.

Inspection of similarity relations between clusters (measured by the mean similarity of source publications from one cluster to word-profiles of other clusters) reveals that not all sets of source publications that cite clusters are, on the average, most relevant to the research topics (represented by word-profiles) of the clusters they cite, but are more relevant to research topics of other clusters. In these cases different clusters probably represent parts of the same specialty.

An expert in the field considered nearly all (92%) of the clusters to be coherent, as far as the associated word-profiles are concerned. However, in the expert's opinion, not all of these clusters represent identifiable research specialties. Twenty-five clusters (66%) were

TABLE 4. Presence in clusters for countries,[a] for dataset on agriculture related biochemical research (only first 20 countries shown).

	(A) Number of Clusters Cited		(B) Number of Clusters Cited Or Similar To.		Rank In Case (A)		Rank In Case (B)	
Country	All	Central	All	Central	All	Central	All	Central
USA	36	35	39	38	1	1	2	1
Japan	23	20	38	31	2	2	5	2
Great Britain	21	20	39	30	3	3	1	3
Fed Rep Ger	19	18	35	24	4	4	7	5
Canada	17	16	36	24	5	5	6	4
France	14	11	38	17	6	7	3	6
Sweden	14	12	24	17	7	6	18	7
India	11	11	38	15	8	8	4	9
Australia	10	10	31	15	9	9	11	8
Italy	7	5	32	10	10	12	10	13
USSR	7	7	31	14	11	10	13	10
Belgium	6	3	31	5	12	16	12	16
Switzerland	6	4	26	12	13	14	17	12
Netherlands	5	5	34	12	14	11	8	11
Hungary	5	3	13	4	15	17	22	22
Israel	5	5	13	8	16	13	23	14
South Africa	5	2	8	5	17	19	29	18
Poland	4	4	16	5	18	15	21	17
Denmark	3	—	30	—	19	—	14	—
South Korea	3	—	7	—	20	—	31	—

[a]At least one publication of a country should cite a cluster or should be similar to research topics involved in a cluster.

TABLE 5. Presence in clusters for countries,[a] for dataset on chemoreception research (only first 15 countries shown).

Country	(A) Number of Clusters Cited		(B) Number of Clusters Cited Or Similar To.		Rank In Case (A)		Rank In Case (B)	
	All	Central	All	Central	All	Central	All	Central
USA	36	29	37	30	1	1	1	1
Japan	14	8	16	11	2	4	2	3
Fed Rep Ger	12	8	12	8	3	5	5	5
France	10	9	14	10	4	2	4	4
Great Britain	9	5	12	7	5	6	6	6
Canada	8	8	14	11	6	3	3	2
Australia	5	3	7	3	7	8	8	8
Sweden	5	3	7	3	8	9	9	13
Spain	4	2	5	3	9	12	10	12
Netherlands	4	4	4	4	10	7	13	7
India	3	1	8	3	11	14	7	9
Israel	3	2	3	3	12	10	14	10
Italy	2	2	4	3	13	11	12	11
Poland	2	0	3	1	14	20	15	20
Argentina	2	2	2	2	15	13	16	14

[a]At least one publication of a country should cite a cluster or should be similar to research topics involved in a cluster.

considered to represent research specialties, among which eighteen were considered different, so that—in the expert's opinion—a number of clusters in fact represent segments of the same research specialty. Clusters interrelated by co-citations (below the applied co-citation strength threshold) were in most cases considered to be cognitively related as well. However, some clusters with similar word-profiles, and thus considered to be cognitively related, are isolated as far as co-citations are concerned. In general, to the expert's opinion, topics were too much fragmented. This split-up of specialties into different clusters possibly reflects

TABLE 6. Overall contribution of countries to clusters and influence of including source publications having word-similarity with cluster topics, for dataset on agriculture-related biochemical research (only first 20 countries shown).

Country	Source Publications Citing to Clusters		Source Publications Similar to Cluster Topics		All Source Publications Related to Clusters		Rank for Citing Source Publications		Rank for Citing and "Similar" Source Publications	
	All	Central	All	Central	All	Central	All	Central	All	Central
USA	240	174	915	145	1155	319	1	1	1	1
Japan	61	41	298	42	359	83	2	2	3	2
Great Britain	44	40	353	26	397	66	3	3	2	3
Fed Rep Ger	41	36	118	22	159	58	4	4	5	4
Canada	40	31	111	19	151	50	5	5	6	5
France	25	17	211	15	236	32	6	6	4	6
Sweden	21	15	27	6	48	21	7	8	15	8
India	20	16	123	11	143	27	8	7	7	7
Italy	11	7	84	8	95	15	9	12	8	11
Australia	10	10	47	8	57	18	10	9	11	9
Belgium	9	4	49	2	58	6	11	16	10	17
USSR	9	9	43	9	52	18	12	10	14	10
Switzerland	9	5	32	8	41	13	13	14	17	14
Israel	8	8	16	6	24	14	14	11	20	13
South Africa	7	2	3	3	10	5	15	19	27	19
Netherlands	6	6	66	9	72	15	16	13	9	12
Hungary	6	4	12	2	18	6	17	17	23	18
Poland	5	5	17	3	22	8	18	15	21	15
Denmark	3	—	38	—	41	—	19	—	16	—
Czechoslovakia	3	3	33	4	36	7	20	18	18	16

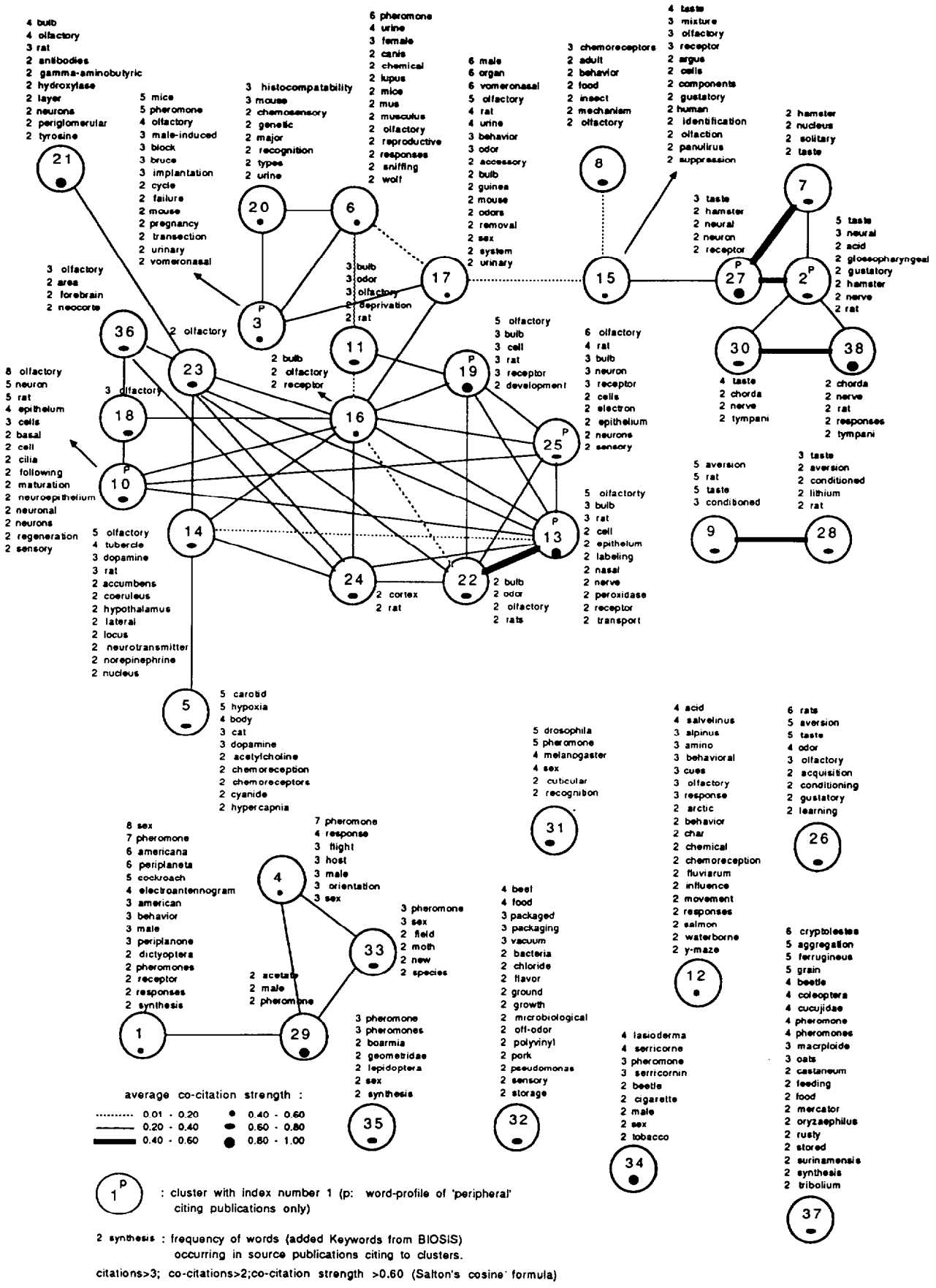


FIG. 4. Chemoreception literature co-citation clusters and related keywords.

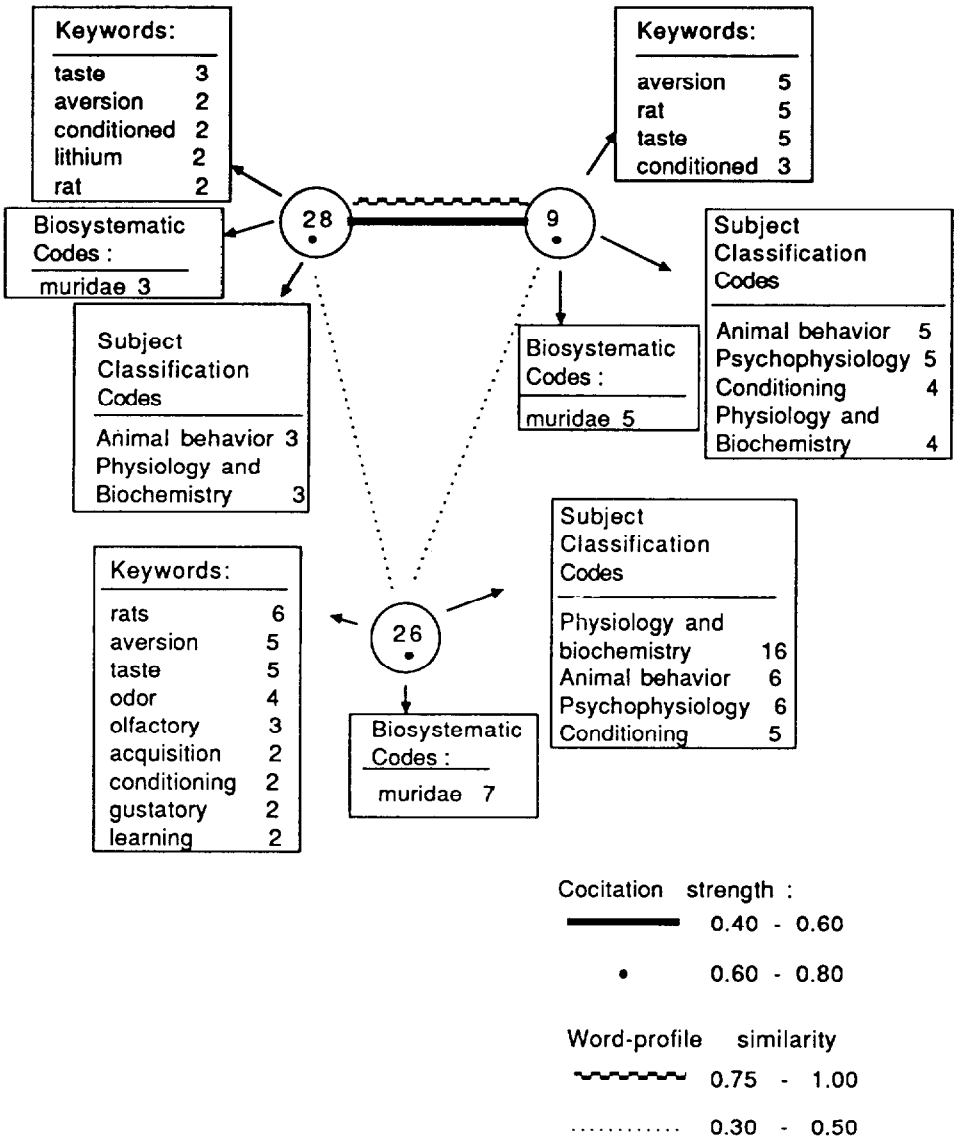


FIG. 5. Split up of a specialty: three different co-citation clusters indicating the same specialty.

differences in intellectual focus expressed in references of publications relevant to the research topics involved in the specialty. The level of the applied co-citation strength threshold then indicates the importance of such difference in intellectual focus. On the other hand, this fragmentation may be largely the result of the maximization of the number of clusters that has been applied in our clustering algorithm. On a higher hierarchical level, groups of interrelated clusters correspond—in the opinion of the expert—to broad research themes, such as pheromone research on insects and on mammals, taste research, taste aversion as a tool in other research, research on olfactory bulb.

One particular cluster, on food-technology research, was considered by the expert to be much too small in relation to the estimated share of publications on this topic in the dataset. This most probably is the result of infrequent referencing and/or absence of reference lists in articles on this topic.

Clusters containing combinations of words considered to be less coherent might, however, still indicate interesting, though unexpected, relations between research specialties (e.g., human taste research, and taste-mixture research on malacostraca ("panulirus"), in cluster 15, Figure 4).

Also, in the case of chemoreception research, results of the co-citation analysis are not "complete," as quite a number (99) of publications not citing any co-citation cluster are highly similar (Table 2) to topics involved in one cluster (54 publications) or more clusters (45 publications) (see earlier section). Again, this can not be explained by a lack of coverage of these publications by the ISI/SCI database (Table 3). The explanation is the same as for the foregoing case: many source publications do (cognitively) belong to a specialty, but they do not have the citing characteristics that relate them to a specific co-citation cluster. The effect of adding these cognitively related but not citing source publications to

clusters, on the participation of countries to clusters and overall ranking is illustrated in Tables 5 and 7. The influence of this effect, though less important than in the case of agriculture-related biochemistry, is not negligibly small. Thus, also in this case, co-citation results offer a distorted picture of a country's participation to clusters, in particular for smaller countries.

Co-Citation clusters, Research Fronts, and Research Specialties

As shown in the foregoing sections, more publications may be involved in a specialty than are citing into co-citation clusters. Subject-similarity between these publications was indicated by relatively high word-profile similarity. In as far as co-citation clusters identify research fronts, one might raise the question whether differences exist between the source publications in a specialty that do participate in the constitution of these clusters, and those publications that do not. Co-citation analysis measures the extent to which researchers use the *same* earlier literature, i.e., it builds on "intellectual focus" as a characteristic of the research front. Another characteristic of the research front, the "immediacy effect," is measured by Price's Index: the percentage of 1 to 5-year-old cited references (De Solla Price, 1970). Price's Index is a measure of the extent to which researchers concentrate their references to the most *recent* earlier literature. As indicators of the research front, Price's Index and co-citation analysis are independent of each other. Of course, both these indicators of the research front (Price Index, and co-citation) are, as indicators, independent of word usage in research publications which reflects in an independent fashion subject-matter of the research involved in current publications of a specialty.

According to Cozzens, co-citation studies have shown that "immediacy" and "intellectual focus" are related phenomena (Cozzens, 1985). If both Price's Index and co-citation are good indicators of the research front, it is interesting to compare Price's Index for the above two groups of source publications: one group consisting of publications citing into co-citation clusters, and another group of publications not citing these clusters but similar in subject-matter to the publications of the first group. In particular such comparison is interesting in relation to the value of Price's Index for all other source publications in the dataset.

In the present context, it is of interest that, for both agriculture-related biochemical research and chemoreception research, source publications related to clusters by citations *or* word similarity, do have a significantly higher value for Price Index than all other sources (differences among means were compared using Tukey's studentized range test at 0.05 significance level). Between them, "citing" source publications have a somewhat higher value for this index than "similar" (though not citing) source publications in case of agriculture-related biochemical research, but in case of chemoreception research these index values are not significantly different ($p = 0.05$). Thus, "citing" and "similar" source publications do share a focus on the *most recent* earlier literature, but consensus about *what* most recent earlier literature is important (and "should be cited") only exists among the publications citing co-citation clusters. Further, analysis of the number of references included in publications reveals that source publications citing clusters contain significantly more references than all

TABLE 7. Overall contribution of countries to clusters and influence of including source publications having word-similarity with cluster topics, for dataset on chemoreception research (only first 15 countries shown).

	Source Publications Citing to Clusters		Source Publications Similar to Cluster Topics		All Source Publications Related to Clusters		Rank for Citing Source Publications		Rank for Citing and "Similar" Source Publications	
Country	All	Central	All	Central	All	Central	All	Central	All	Central
USA	173	93	40	18	213	111	1	1	1	1
Japan	24	16	7	5	31	21	2	3	3	3
Canada	20	16	14	6	34	22	3	2	2	2
Fed Rep Ger	16	10	1	1	17	11	4	5	6	5
Great Britain	14	8	5	3	19	11	5	6	5	6
France	12	10	8	4	20	14	6	4	4	4
Sweden	10	5	3	1	13	6	7	7	8	7
India	7	1	8	4	15	5	8	16	7	8
Spain	6	2	1	1	7	3	9	13	10	14
Australia	5	3	2	0	7	3	10	9	9	10
Netherlands	5	5	0	0	5	5	11	8	13	9
Israel	4	2	1	1	5	3	12	10	11	11
Italy	2	2	3	1	5	3	13	11	12	12
Mexico	2	2	1	1	3	3	14	12	15	13
Poland	2	0	1	1	3	1	15	20	16	20

other source publications, including those source publications similar to (but not citing) clusters. Thus it seems that part of the literature involved in a specialty is characterized by a (partial) consensus concerning *what* select part of the most recent earlier literature is of importance. These publications contain also relatively more references and constitute a co-citation cluster. Another part of the specialty-literature consists of publications not participating in this (partial) consensus, or at least not expressing such consensus in their references. These publications contain, on the average, less references. A schematic overview of these results is given in Figure 6.

Thus, according to our results, there seems to be a discrepancy between "intellectual focus on base literature" and "immediacy" within scientific specialties. If, according to Cozzens (Cozzens, 1985), co-citation studies confirm the relationship between immediacy and intellectual focus on base literature, this confirmation may be largely due to the fact that such studies analyze only the literature citing co-citation clusters, and disregard papers similar in content though not citing any of the documents in these clusters.

Further, for example, validation of results of a co-citation study of the collagen specialty carried out by Small (Small, 1977) was based on questionnaires sent only to authors *citing* one or more papers in the 1973-collagen cluster. Thus, validation was based only on authors involved in the "consensus part" of the specialty. The question whether collagen research publications *not citing* the 1973-cluster could provide significant information on the development of the specialty, was not examined in that study. It may therefore be interesting to examine in what part of a specialty the "critical processes" of knowledge growth in fact take place. Perhaps, the most interesting, or even "crucial" processes take place in the part of a specialty that is not captured by co-citation analysis! For example, possibly, in the "nonconsensus" part of a specialty fresh ideas are generated by researchers drawing on alternative basic concepts, methods, or experiments, expressed in "non-standard" literature! An answer to these latter questions, however, awaits further research.

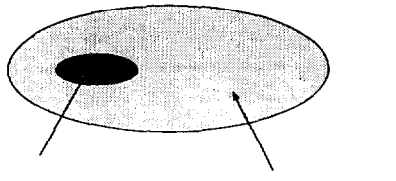


FIG. 6. Publications in a subject-related research area.

Conclusions

Combining word-analysis and co-citation analysis offers a useful instrument to describe, evaluate, and compare results of co-citation analysis in a systematic and clear way, in particular concerning aspects related to the cognitive content of publications. It can now be more accurately established what topics are involved in clusters, and whether clusters are, in a quantitative sense, sufficiently (internal) coherent and (external) different one from another in order to speak of "separate" research specialties. Moreover, it can now be established whether all sources relevant to identified research topics are indeed retrieved using co-citation analysis, and whether retrieval can be improved by adding to these topics all relevant source publications not citing the clusters involved but having high word-similarity with research topics related to these clusters. The question whether all topics covered by a dataset are identified by co-citation clusters can, only *partially,* be answered by comparing results for different sets of thresholds. Of course, research specialties lacking "focused" referencing can never be identified using co-citation analysis. Perhaps, such topics are covered by co-word analysis. For such comparison we refer to our forthcoming publication (Braam et al., 1989).

Our results suggest that co-citation analysis indeed displays research specialties, although these may be fragmented into several different clusters. However, we found that the analysis only partially reveals the literature relevant to identified research topics involved in the current work of these specialties. Further, interrelations between clusters seem to correspond to cognitive relations on a higher level than specialties, such as fields defined by conventional subject classification (CA section-codes).

Thus, co-citation clusters are certainly not, at least not entirely, artifacts of an applied technique, but on the other hand they do not represent the entire published current work comprised in a specialty. Therefore, as a document retrieval tool, co-citation analysis is not appropriate to select all, or even a major part of the source publications relevant to these specialties (low recall).

It can be concluded from our results that the relative number of source publications citing clustered documents, as compared to the total number of source publications relevant to topics involved in clusters, is related to the degree of consensus concerning the intellectual base literature existing in a specialty, i.e., consensus about *what* earlier publications are of importance.

However, an alternative explanation might be the following, taking into account the difference in the number of references per publication. In a number of publications much attention is paid to inclusion of these earlier important publications in the reference lists, whereas these are left out in the reference lists of other publications. In this latter case consensus on intellectual base literature may not be entirely absent, but no consensus exists on the need to refer to these publications.

Nevertheless, according to our results, a discrepancy seems to be present within specialties between "immediacy" and "intellectual focus on base literature," indicating differences in the processes of knowledge growth within these specialties. It will be interesting to study in more detail these differences in processes of knowledge growth within specialties, in order to shed more light on the relationship between the concept of "research specialties" and Price's concept of "research fronts." In particular, it is interesting to examine in which part of a specialty the "critical" processes (e.g., "discoveries") take place, because these processes might be taking place in the part not covered by co-citation analysis.

Discussions with experts point in the same direction as our evaluation by word analysis: results may probably be improved by aggregating highly similar clusters, and by adding to clusters all source publications relevant to topics involved in these clusters (via word-similarity) although not citing the clusters concerned.

Support for such a strategy is also found in the fact that source publications, whether or not citing clusters, that are highly similar to topics involved in clusters, also have a high Price Index compared to other sources. This indicates a shared concentration on the most recent scientific literature (De Solla Price, 1965), although *not* a shared concentration on the *same* recent literature, and therefore only part of these publications also constitute co-citation clusters. Apparently, a number of current researchers do not share consensus regarding which recent earlier publications are of special importance or usefulness for the research in the specialty, or at least they do not include these publications in the reference lists of their publications.

Thus, in as far as co-citation clusters identify research specialties, it is clear that they do so only partially, and probably only cover the part where consensus exists about important or useful earlier contributions. The remaining part lacks such consensus, but still uses to a large extent the same words for problems and concepts in the research involved and is also interested in most recent literature. As researchers involved in this part of a research specialty cite in an unconventional, or uncommon way, they also constitute no separate co-citation cluster.

This "lack of citations" to co-citation clusters in source publications that are otherwise relevant to research topics related to these clusters is not regularly distributed over countries. Therefore, the use of co-citation analysis does not offer a representative picture of participation in research topics by countries. Combination of co-citation analysis with word analysis offers a possibility to improve the applicability of such results, since a more complete set of publications can be retrieved.

It should be noted that such an additional word analysis can be used at any level of the co-citation strength. The applied strength level defines size and scope of a cluster, a low level indicating a broad, a high level a narrow focus (ISI/DIMDI, 1987). According to our view, an additional word analysis should be used at any applied strength level in order to see whether at that level size and scope of a cluster is sufficient to comprise coherent and complete groups of current publications concerning the research topics indicated by these clusters.

The question remains, also in case of combined co-citation and word analysis, whether this approach can cover all research topics in a dataset; though results may be very complete for specific identified topics, there may still be topics that are not identified using co-citation clustering. Further work is needed to examine whether there are more appropriate criteria to specify threshold levels used in the clustering process. For instance, it could be useful to vary the threshold level in order to maximize not only the number of clusters, but also their size and coherence. In addition, it might be interesting to compare results of co-citation analysis to other clustering routines, e.g., routines based on complete linkage clustering, or clustering routines using other aspects of publications than co-cited references such as co-words.

We agree with Mullins et al. (1988) that, in order to generate significant results in the field of "Mapping of science," it will be necessary to analyze different structural aspects of publications in combination in a quantitative fashion in stead of the exclusive use of single aspects. In fact, the present study, combining co-citation and word analysis is an example of such approach. At the present we explore (Braam et al., 1989) the fruitfulness of combining co-citation clustering, and co-word clustering.

References

ABRC (Advisory Board for the Research Councils), The Royal Society Policy Studies Unit, and the ESRC (Economic and Social Research Council), (1986). Evaluation of National Performance in Basic Research: a review of techniques for evaluating national performance in basic research, with case studies in genetics and solid state physics. In *ABRC Science Policy Studies,* (pp. 137–139). No. 1, London: Department of Education and Science.

Braam, R.R., Moed, H.F., & Van Raan, A.F.J. (1988a). Mapping of Science: Critical Elaboration and New Approaches, a Case Study in Agricultural Related Biochemistry, In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88* (pp. 15–28). Amsterdam: Elsevier Science Publishers. And also, in a more extended version: Braam, R.R., Moed, H.F., & Van Raan, A.F.J. (1987, November). *Mapping of science: Critical elaboration and new ap-*

proaches, a case study in agricultural Biochemistry. Research Report to the Netherlands Advisory Council for Science Policy (RAWB), Leiden.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1988b). Mapping research specialties and their interrelations: a combined co-citation and word analysis of chemoreception literature. *News Letter of the European Chemoreception Research Organization (ECRO), 38,* 360–363.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1989, November). Comparison of co-citation and co-word clustering. In A. F. J. Van Raan, Nederhof, A. J., & Moed, H. F. (Eds.), *Science and technology indicators their use in science policy and their role in science studies: Select proceedings of the first international workshop on science and technology indicators Leiden, The Netherlands, 14–16 November 1988* (pp. 307–337), DSWO Press, University of Leiden, The Netherlands.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1990). Mapping of Science by Combined Co-Citation and Word Analysis: II Dynamical Aspects. *JASIS,* Ed. update.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: A introduction to co-word analysis. *Social Science Information, 22,* 191–235.

Cozzens, S. E. (1985). Using the archive: Derek Price's theory of differences among the sciences. *Scientometrics, 7,* 431–441.

De Solla Price, D. J. (1965). Networks of scientific papers. *Science, 149,* 510–515.

De Solla Price, D. J. (1970). Citation measures of hard science, soft science, technology, and non-science. In C. E. Nelson, & D. K. Pollack (Eds), *Communication among scientists and engineers* (pp. 1–12). Lexington, Mass.: Heath Lexington.

Franklin, J. J. (1988). Testing and using quantitative methods in science policy contexts: A response to Hicks. *Social Studies of Science, 18,* 365–374.

Garfield, E., Malin, M.V., & Small, H. (1978). Citation data as science indicators. In Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray, & H. Zuckerman (Eds.), *Toward a Metric of Science: The Advent of Science Indicators* (pp. 179–207). New York: John Wiley & Sons.

Griffith, B. C., Small, H. G., Stonehill, J. A., & Dey, S. (1974). The structure of scientific literatures II: Toward a macro- and microstructure for science. *Science Studies, 4,* 339–365.

Healey, P., Rothman, H., & Hoch, P. K. (1986). An experiment in science mapping for research planning. *Research Policy, 15,* 233–251.

Hicks, D. (1987). Limitations of co-citation analysis as a tool for science policy. *Social Studies of Science, 17,* 295–316.

Hicks, D. (1988). Limitations and more limitations of co-citation analysis/bibliometric modeling: A reply to Franklin. *Social Studies of Science, 18,* 375–384.

ISI/DIMDI, (1987, August). *SCISEARCH–user manual English version,* Edition 01.00 (pp. 8–10). Köln: DIMDI-Deutsches Institut für Medizinische Dokumentation.

Jones, W. P., & Furnas, G.W. (1987). Pictures of Relevance: A Geometric Analysis of Similarity Measures. *Journal of the American Society for Information Science, 38,* 420–442.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of Information Science, 13,* 261–276.

Kuhn, T. S. (1970). *The Structure of scientific revolutions,* Second Enlarged Edition. Chicago: University of Chicago Press.

Laudan, L. (1977). *Progress and its problems: Towards a theory of scientific growth.* California: University of California Press.

Moed, H. F. (1988). The use of online databases for bibliometric analysis. In L. Egghe & R. Rousseau (Eds.), *Informetrics 87/88 (pp. 133–146). Amsterdam: Elsevier Science Publishers.*

Mombers, C., Van Heeringen, A., Van Venetië, R., & Le Pair, C. (1985). Displaying strengths and weaknesses in national R&D performance through document citation. Scientometrics, 7, 341–355.

Mullins, N., Snizek, W., &Oehler, K. (1988). The structural analysis of a scientific paper. In A. F. J. Van Raan (Ed.), *Handbook of Quantitative Studies of Science & Technology* (pp. 85–101). Amsterdam: Elsevier Science Publishers.

Oberski, J. E. J. (1988). Some statistical aspects of co-citation cluster analysis and a judgement by physicists. In Van Raan A. F. J., (Ed.), *Handbook of Quantitative Studies of Science & Technology (pp. 431–462). Amsterdam: Elsevier Science Publishers.*

Rip, A. J., & Courtial, J. (1984). Co-word maps of biotechnology: an example of cognitive scientometrics. Scientometrics, 6, 381–400.

Rip, A. J., (1988). Mapping of Science: Possibilities and Limitations. In A. F. J. Van Raan, (Ed.). *Handbook of Quantitative Studies of Science & Technology* (pp. 253–273). Amsterdam: Elsevier Science Publishers.

Small, H. G. (1973). Co-citation in the scientific literature: A new measure of the relationship between publications. *Journal of the American Society of Information Science, 24,* 265–269.

Small, H. G., & Griffith, B. C. (1974). The structure of scientific literatures, I: Identifying and graphing specialties. *Science Studies, 4,* 17–40.

Small, H. G. A (1977). Co-citation model of a scientific specialty: A longitudinal study of collagen research. *Social Studies of Science, 7,* 139–166.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science, 8,* 327–340.

Small, H. G. & Crane, D. (1979). Specialties and disciplines in science and social science: An examination of their structure using citation indexes, *Scientometrics, 1,* 445–461.

Small, H., & Sweeney, E. (1985). Clustering the science citation index using co-citations: I. A comparison of methods. *Scientometrics, 7,* 393–409.

Small, H., Sweeney, E., & Greenlee, E. (1985). Clustering the science citation index using co-citations: II. Mapping science. *Scientometrics, 8,* 321–340.

Small, H. (1986). The synthesis of specialty narratives from co-citation clusters. *Journal of the American Society of Information Science, 37,* 97–110.

Small, H. G. & Garfield, E. (1986). The geograpphy of science: Disciplinary and national mappings. *Journal of Information Science, 11,* 147–159.

Sullivan, D., White, D. H., & Barboni, E. J. (1977). Co-citation analysis of science: An evaluation. *Social Studies of Science, 7,* 223–240.

Van Der Starre, H. (Ed.), (1985, and 1986). *Chemoreception references,* Vol. 1, nr. 4 and Vol. 2, nrs 1 and 2. European Chemoreception Research Organization (ECRO).