INFORMETRICS 87/88
L. Egghe and R. Rousseau (Editors)
Elsevier Science Publishers B.V., 1988

CO-WORDS AND CITATIONS RELATIONS BETWEEN DOCUMENT SETS AND ENVIRONMENTS

L. LEYDESDORFF and R. ZAAL

Department of Science Dynamics, Nieuwe Achtergracht 166,
1018 WV Amsterdam, The Netherlands

Abstract

Co-word linkages among documents have been proposed as an alternative to citations .(and co-citations) for the study of how documents are related. In this study we examine co-word linkages among a set of biochemistry documents which were subjected to a citation analysis in another context. The co-word structure in the titles of this document set is compared with that of the titles of the documents which cited at least one of these documents, and with that in all other documents of the Science Citation Index in the same period. A similar analysis of the words used in the abstracts of these documents has been made, to check whether one should start the analysis from titles or abstracts.

The structure of co-words in the titles of the citing documents significantly resembles that of the co-word structure of the cited documents. These co-word structures in the first place seem to reflect the internal structure of the document set and the related groups of authors, as can be validated against the results of an independent document analysis.

Since it is possible to search in the SCI for both title-words and index-words (cluster terms), as a side-effect of our research, we are able to draw systematic conclusions about the so-called 'indexer effect' in co-word analysis.

1. INTRODUCTION

Despite the popularity of citations as a means of measuring performance in research, and the continuous call for a theory of citation in the literature, no theory of citation is yet available ([1], see also [2]). It has been suggested that citations reflect 'impact', 'indebtedness', 'quality', or at least partially also institutional affiliation. The problem of what citations mean becomes even more urgent when we move to the study of aggregates of citations or specific (logical) combinations such as co-citations. In the latter case it has been strongly argued that co-citations reflect the world of science as practicing scientists perceive it

The research reported here was executed within the framework of a larger project conducted in cooperation with Dr. Olga Amsterdamska, to whom we are grateful for her comments and criticism. The authors also want to thank Nero Matthias for developing the necessary software.

[3]. In the case of journal-journal citations we have argued elsewhere that in a certain sense these aggregate numbers may provide us with a baseline against which to measure changes at other levels of aggregation [4].

Of course, the attention to citations in science studies has been induced by the availability of the Science Citation Index in machine readable form. However, this Index contains not only citations but also titles and index words--the latter based on cluster analysis of co-citations [3]. Other important databases on various sciences provide abstracts as well, and with the new technologies full texts are or will soon be available in machine readable form. (For a substantive argument against the use of words and their co-occurrences, and in favour of citations, see [5]).

In a scientific text the title, the abstract, the addresses, and the references are all indicators of the text which should not be disconnected from each other, nor from the knowledge claims which are made in the article. The problem of how to validate particular indicators, which now tends to be treated rather crudely by asking some important scientists whether computer-plots make sense to them (see [6] among others), can only be solved theoretically when we are able to understand the functions of specific indicators with respect to the knowledge claim involved, and to the overall structure of the argument. What does it mean in terms of the linking elements that one method leads to results different from another? And what does the appreciation of the results of one type of analysis or another by practicing scientists tell us about these elements considered important in various aspects of the scientific enterprise? How do scientists link their own papers to other document sets, and how do they and their colleagues perceive of this process?

Callon, Courtial, Turner and Rip [7] have argued that the analysis of co-occurrences of words (co-words) may lead to a more cognitive indicator for 'the dynamics of science' than citation-analysis. Indeed, it is wellknown that practicing scientists are very careful in choosing the titles of their articles, also with respect to the classification of their articles in abstracts and indexing services under the appropriate headings. Therefore, we decided to look more carefully at title-words and their co-occurrences.

2. CO-WORDS

In an extensive validation study based on interviews with practicing scientists, the conclusions proved to be more favorable to citation and co-citation analysis than to the results of co-word analysis with respect to the question of how to 'map' science [6]. Moreover, at the methodological level, the serious problem of the so-called 'indexer effect' (resulting from the fact that keywords are selected by an indexer who is not a practicing scientist) has yet to be solved. More recently, Courtial claims to have been able to circumvent this problem technically ([8]).

To prevent the 'indexer effect' from occurring in this study, we will use the original title and abstract words of the document set instead of index words. We are interested here in the differences and similarities between the two ways by which scientific articles are linked to the scientific literature: by sharing words in titles (and abstracts), in contrast and as an alternative to the linking of documents by citations.

Therefore, we should not select title-words on a priori grounds, or subsume them under broader terms, but use all the title words of the documents involved as keywords. Because titles may be selected with respect to an audience-- and hence, we might have an 'audience effect' here-- we repeated the analysis using words occurring with a certain frequency in the collection of the abstracts of these documents.

However, since we will have to conclude from the latter comparison-- between title-words and abstract-words-- that there are certain advantages in using titles instead of abstracts, we will pursue the analysis with the title-words in the citing documents, and in the whole Science Citation Index database.

3. METHODS

3.1 Samples

Citations from all the documents of one research group at the Biochemistry Laboratory of Amsterdam University during the period 1979-1982 were stored in a database for other research purposes [9]. Of these 57 documents, which were cited 639 times by December 31, 1985, only the 47 full-length articles were selected. This set was divided into two equal parts (23+24) randomly, to keep the second half for later validation if necessary.

From the set of 311 articles which cited these original 23 articles 121 self-citations or 'in-group' citations were removed. Hence, we have a document set of 23 original cited documents and one of 190 external citing documents.

In the 23 cited articles, 45 title-words occurred more than once and were not trivial. (A word which occurs only once cannot form a co-word linkage.) Abstracts of these documents were downloaded from the Index Medicus as installed on DIALOG, and when we could not find them there, we checked the original document. In the few cases with no abstracts, we used the abstract as provided by Chemical Abstracts.

These abstracts contained 825 different words, of which only 57 occurred more than three times and only 29 more than four times. We took the cut-off at the latter level of 29 words because for reasons of cost-effectiveness we had decided arbitrarily to keep the possible combinations as close to 1000 as possible. (The total number of combinations is N*(N-1)/2.)

The 190 citing documents contained 724 different words which showed the same type of skewed distribution.

The DIMDI-installation of the Science Citation Index was used to compare the co-word structure of the document sets with the co-word structure in titles and index terms in that database.

3.2 Statistics

Co-word analysis generates a symmetrical matrix with an empty diagonal, i.e., A AND B happens as many times as B AND A. The matrices were factor-analyzed using both orthogonal and oblique rotations (to check for inter-factorial relations). For graphic representation cluster analysis was pursued using Ward's mode of analysis. Elsewhere, we have argued that Wards' mode of analysis is better suited for matrices like these than single linkage clustering because of the large number of zero hits which may lead to 'chaining' in the first cluster and 'isolates' ([4]).

In this case, we have the additional problem of the choice of a similarity coefficient from which to cluster. For reasons of comparison with the factor analysis, the choice of the Pearson correlation seems straightforward. Others, however, have argued for the use of the Jaccard coefficient with co-word analysis ([10], see also [11]). In his study of Computer Science Literature Salton [12] has proposed using the so-called cosine formula because it deals with links between high and low cited papers more effectively than the Jaccard Index. This argument has been taken over by Small and Sweeney [13], and may be a valid argument in our case, too.

However, in practice the choice of coefficient does not make much difference; the solutions from a Pearson correlation matrix and from a similarity matrix based on the cosine formula are almost identical. In the first case, we will give the dendrogram which results from clustering with all four coefficients discussed. In latter cases, we will comply to the use of the cosine formula.

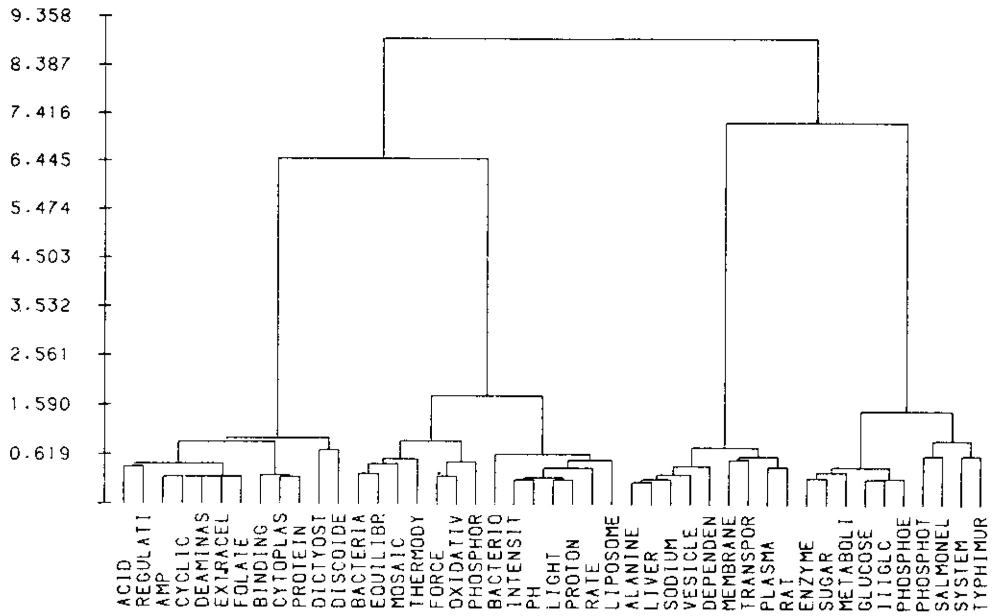

Figure 1: Co-word patterns of 45 title-words in 23 biochemistry articles Wards' mode of analysis; Euclidean distances.

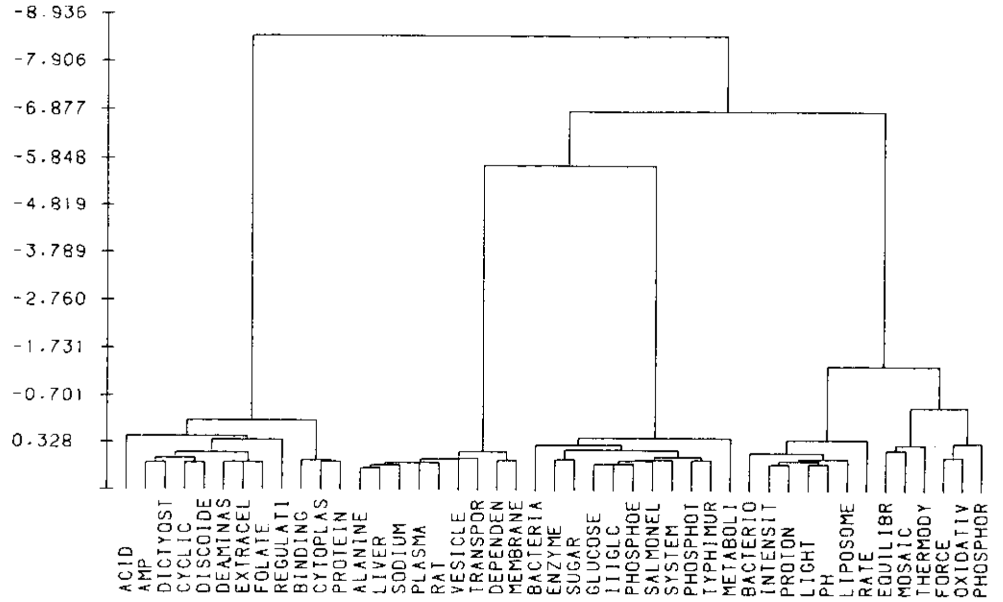

Figure 2: As figure 1 but on the base of Pearson Correlations.

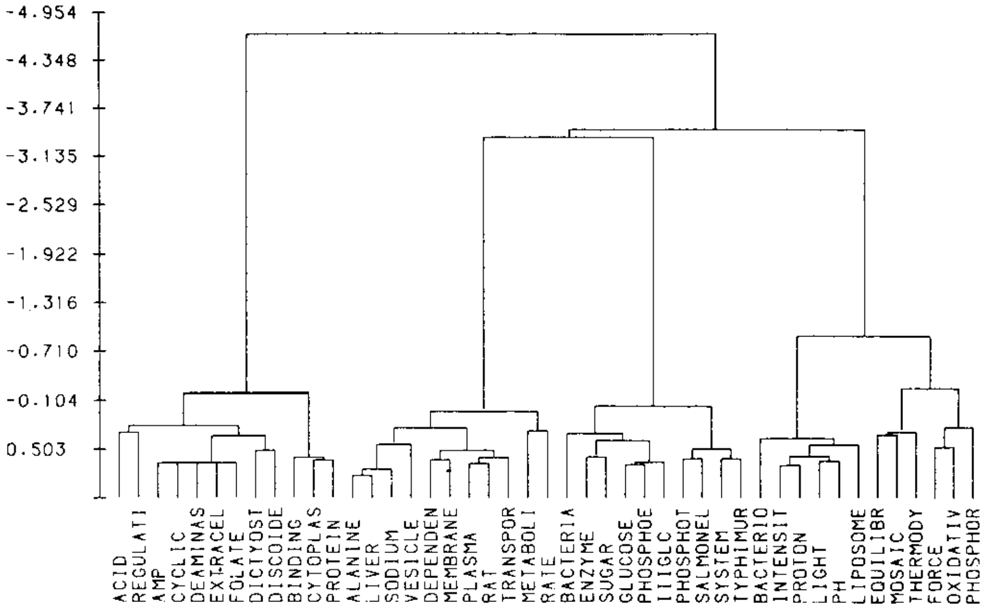

Figure 3: As figure 1 but on the base of the Jaccard Index

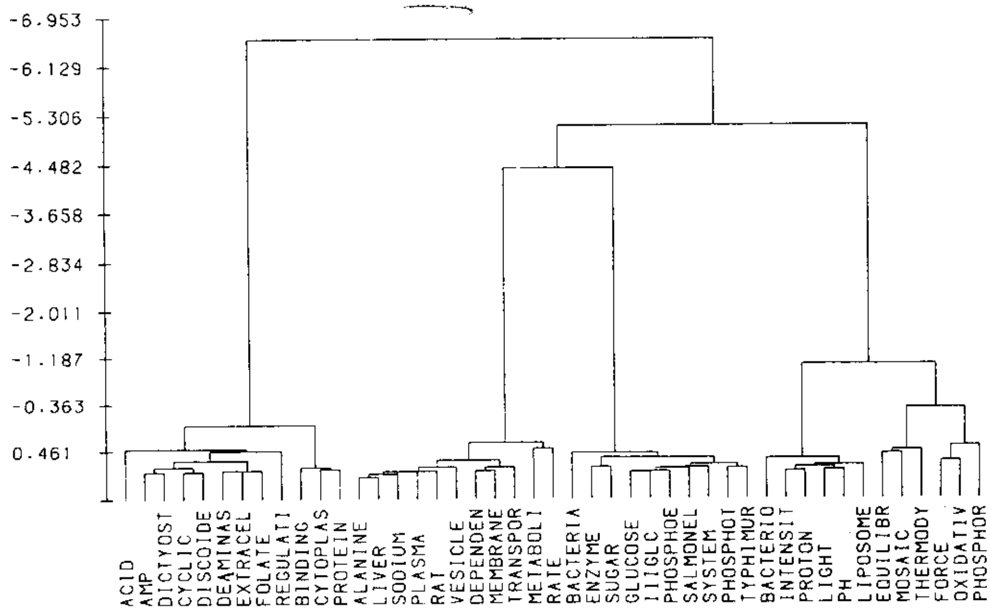

Figure 4: As figure 1 but on the base of the Cosine

Table 1: 45 title words from 23 biochemistry documents

Acid
Alanine
Amp
Bacterial
Bacteriorhodopsin
Binding
Cyclic
Cytoplasmic
Deaminase
Dependent
Dictyostelium
Discoideum
Enzyme
Equilibrium
Extracellular
Folate
Force
Glucose
IIIGLC
Intensity
Light
Liposomes
Liver
Membrane
Metabolism
Mosaic
Oxidative
Ph
Phosphoenolpyruvate
Phosphorylation
Phosphotransferase
Plasma
Protein
Proton
Rat
Rate
Regulation
Salmonella
Sodium
Sugar
System
Thermodynamics
Transport
Typhimurium
Vesicles

4. RESULTS

4.1 Titles of the original articles (N=23)

As noted above, the document set of original biochemistry documents contained 45 title-words occurring in at least two documents. All 23 documents are linked by at least one of these words, which occur in the set a total of 122 times (i.e., 5.4 per document). The 45 title words are listed in Table 1.

These 45 words form 333 co-words which indicates the internal coherence of the set (which, of course, was generated from one research group). Analysis of this matrix leads to a clear division into four major clusters. In general, this results is insensitive to the choise of similarity coefficient. Figures 1-4 give the dendrograms for various coefficients as announced in section 3; note the striking similarity of figure 2 (Pearson) and figure 4 (cosine).

The four clusters are immediately recognizable by researchers as belonging to the four research lines pursued in the research programme of the laboratory group under study. As we knew from document analysis and interviewing, the group as a whole works on energy-dependent transport systems in biological membranes; but four specific lines of research can be distinguished, among others in terms of the differences in their research objects, namely:

1. The regulation of binding in membranes of Dictyostelium Discoideum;
2. The thermodynamic study of oxidative phosphorylation (e.g. in Bacteriorhodopson);
3. Membrane transport in rat liver cells and vesicles;
4. Sugar transports in Salmonella Thyphimurium (phospho-transferase).

Table 2 presents the factor analytic solution in four dimensions-- four factors account for 65.6% of the common variance. (The orthogonal solution is given since in the oblique solution no substantial factor relations were found.) When more factors are allowed (7 factors have an eigenvalue > 1) internal differentiations of the research lines are also revealed at lower levels, as is also visible in the dendrogram. As can be expected from the substance of the field, words like 'regulation' have some factorial complexity.

4.2 Abstracts of the original articles (N=23)

The abstracts of the 23 documents contain 29 words which occur more than four times, and 170 times in total (7.4 per abstract). In the 23 documents these 29 words generate 800 co-word links (as against 333 co-word links by 45 words in the former case!).

Table 2 : Varimax Rotated Factor Matrix of Co-Word Occurrences of 45 Title Words in 23 Biochemistry Documents

	FACTOR 1 (Van Driel)	FACTOR 2 (Postma)	FACTOR 3 (Sips)	FACTOR 4 (Van Dam)
ACID	.20058 *	-.06096	.20189 *	-.04024
ALANINE	-.02181	-.03869	.30709 *	-.00798
AMP	.28720 *	-.04913	-.05136	-.03498
BACTERIA	-.09004	.18929 *	-.05608	-.01650
BACTERIO	-.03018	-.02612	-.03609	.31598 *
BINDING	.20323 ●	-.05780	-.06421	-.02650
CYCLIC	.28720 *	-.04913	-.05136	-.03498
CYTOPLAS	.21754 ●	-.05123	-.05604	-.03486
DEAMINAS	.27179 *	-.02493	-.02971	-.00801
DEPENDEN	-.01657	.09864	.26731 *	-.02649
DICTYOST	.25336 ●	-.05868	-.05960	-.04886
DISCOIDE	.25336 *	-.05868	-.05960	-.04886
ENZYME	-.06510	.24029 *	-.01058	-.06392
EQUILIBR	-.12365	-.07579	-.08826	-.01933
EXTRACEL	.27179 *	-.02493	-.02971	-.00801
FOLATE	.27179 *	-.02493	-.02971	-.00801
FORCE	-.23864	-.16833	-.14437	-.24291
GLUCOSE	-.01185	.31317 *	-.01221	.00840
IIIGLC	.01603	.31140 *	-.01937	.04070
INTENSIT	-.02916	-.02058	-.03951	.37829 *
LIGHT	-.02916	-.02058	-.03951	.37829 *
LIPOSOME	-.02642	-.02135	-.03507	.33940 *
LIVER	-.01828	-.05018	.31075 *	-.01993
MEMBRANE	-.03853	.03951	.27358 *	-.04014
METABOLI	-.00484	.18895 *	.17359 *	.01472
MOSAIC	-.12432	-.09439	-.10098	.07480
OXIDATIV	-.21894	-.15155	-.13288	-.22783
PH	.02191	.01721	-.01294	.41275 *
PHOSPHOE	.00960	.32177 *	-.03075	.01623
PHOSPHOR	-.21388	-.07026	-.11246	-.22367
PHOSPHOT	-.03998	.27905 *	-.02760	-.04184
PLASMA	-.02292	-.04653	.29592 *	-.02706
PROTEIN	.21754 *	-.05123	-.05604	-.03486
PROTON	-.12253	-.08776	-.09432	.24327 *
RAT	-.02292	-.04653	.29592 *	-.02706
RATE	-.06669	-.06313	.16152 *	.24171 *
REGULATI	.22279 *	.19228 *	-.05090	-.00056
SALMONEL	-.01076	.29058 *	-.02937	-.00252
SODIUM	-.00931	-.04209	.30797 *	-.02561
SUGAR	-.03947	.25070 *	-.03246	-.05426
SYSTEM	-.03998	.27905 *	-.02760	-.04184
THERMODY	-.17277	-.11883	-.12406	.06792
TRANSPOR	-.05411	-.01698	.27099 *	-.06652
TYPHIMUR	-.01076	.29058 *	-.02937	-.00252
VESICLES	.01626	-.03402	.29347 *	.06895

● Means : factor loading

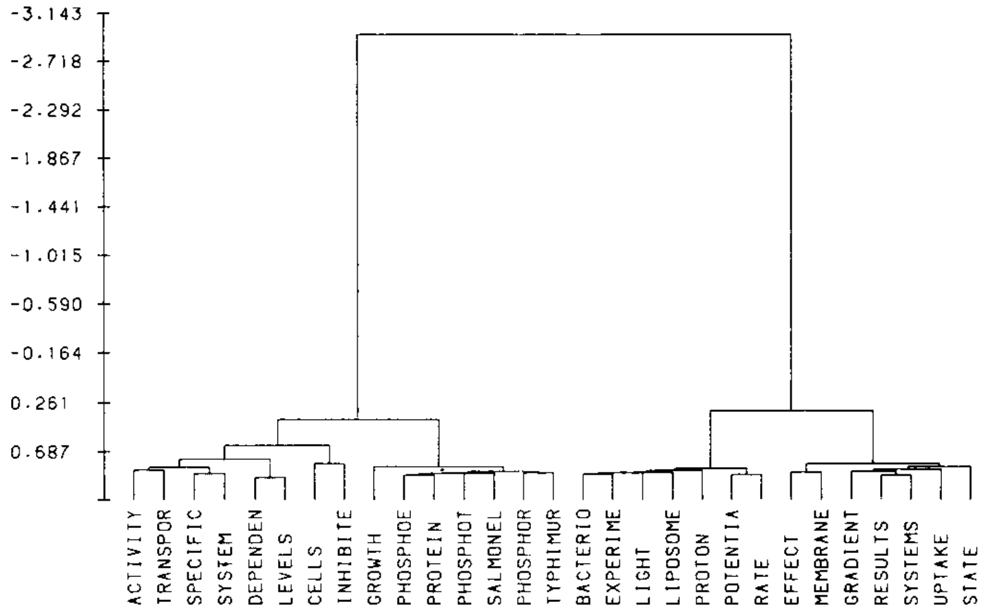

Figure 5: Co-word patterns of 29 abstract-words from 23 biochemistry articles. Ward's mode of analysis; Cosine as similarity coefficient

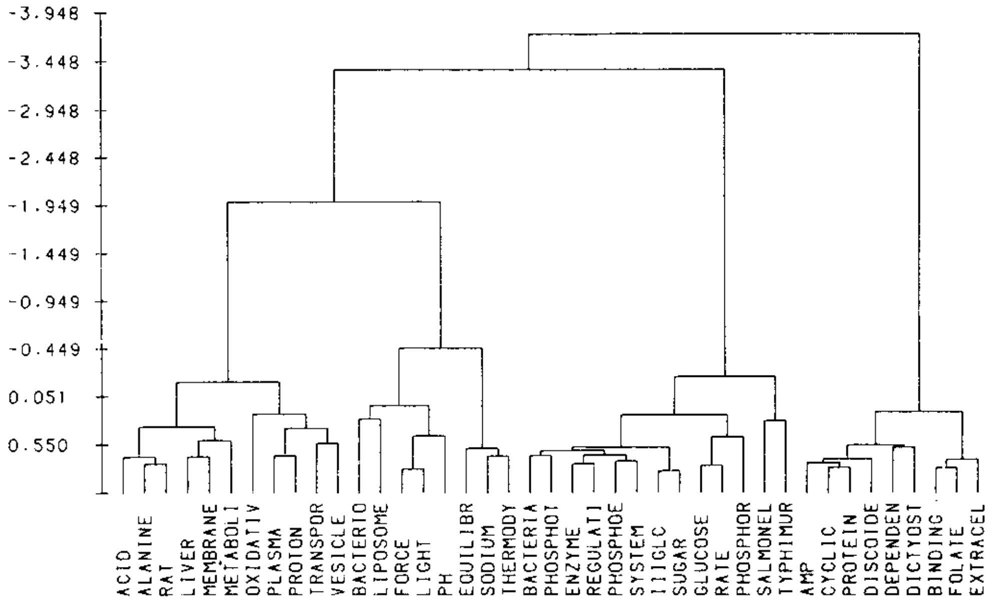

Figure 6

Table 3 : Varimax Rotated Factor Matrix of Co-Word Occurrences of 45 Title Words in 190 Citing Documents

	FACTOR 1 (Van Dam)	FACTOR 2 (Postma)	FACTOR 3 (Sips)	FACTOR 4 (Van Driel)
ACID	-.01653	.05281	.30571 *	.15160 *
ALANINE	-.08689	-.03393	.35231 *	.00588
AMP	.01567	.01181	.00191	.34152 *
BACTERIA	.13803 *	.30996 *	-.06270	-.03420
BACTERIO	.19554 *	-.02668	-.01752	-.01917
BINDING	-.02382	-.05382	-.02754	.33172 *
CYCLIC	-.01592	.01282	.01027	.33118 *
DEPENDEN	.01423	.16712 *	.00013	.25811 *
DICTYOST	-.01631	-.02791	-.01800	.28405 *
DISCOIDE	.00297	-.01134	-.00996	.29399 *
ENZYME	-.01214	.32223 *	-.03895	.04198
EQUILIBR	.26624 *	.02394	-.08301	-.10578
EXTRACEL	.09902	-.06869	-.02792	.27599 *
FOLATE	.07124	-.05443	-.01735	.32643 *
FORCE	.40410 *	.04317	-.01719	.00069
GLUCOSE	.01466	.30586 *	.05627	-.00372
IIIGLC	-.05613	.29607 *	-.08985	-.03826
LIGHT	.43901 *	-.00941	.02517	.11508
LIPOSOME	.19182 *	-.05115	-.05898	-.01562
LIVER	-.00272	-.03616	.33432 *	-.01251
MEMBRANE	.17435	-.03031	.28147 *	-.05144
METABOLI	-.12933	-.03926	.30345 *	.00596
OXIDATIV	.06072	.02080	.13026 *	-.07075
PH	.27679 *	-.07710	-.00312	.17742
PHOSPHOE	.04002	.31585 *	-.00305	-.00533
PHOSPHOR	.06161	.17115 *	.08554	-.06692
PHOSPHOT	.01794	.29772 *	-.03274	-.01634
PLASMA	.05774	-.06093	.33249 *	-.03244 *
PROTEIN	.00666	.09078	.04447	.31169
PROTON	.12569 *	-.11835	.11568 *	-.11712
RAT	.01789	.01553	.34647 *	-.02700
RATE	-.06263	.20535 *	.18966 *	.00636
REGULATI	-.06955	.25466 *	.10953 *	.12896 *
SALMONEL	-.13494	.07426	-.11954	-.02172
SODIUM	.34787 *	.02377	.01021	-.08152
SUGAR	-.00948	.30093 *	-.07966	-.02026
SYSTEM	.00356	.29367 *	.05767	-.02340
THERMODY	.29320 *	.02465	-.09532	-.09923
TRANSPOR	-.08423	.06298	.23225 *	-.02650
TYPHIMUR	-.13494	.07426	-.11954	-.02172
VESICLES	.19809 *	.12389 *	.21909 *	-.04684

• Means : factor loading

However, the dendrogram (Fig. 5) reveals two main clusters corresponding to the two main senior researchers in the group. The smaller clusters disappear in this analysis. The probable reason for this is that the specific words in the few articles of these smaller 'research lines' drop out at a word-frequency level of four times per word; these smaller research lines are then only connected through the words they share with the other members of the group to the main lines. Moreover, it is clear from inspection of the words that these terms are less specific than the title words. Hence, we pursued analysis further using title words.

4.3 Title words of citing articles (N=190)

Of the 190 citing articles (after substracting 121 'in-group' and self-citations), 168 or 88% use at least one of the 45 title-words of the original document set in their title. Actually, they use only 41 title-words, since four of the words on the original list ('cytoplasmic', 'deaminase', 'intensity', and 'mosaic') are not used in the titles of the citing documents at all.

The distribution of the total of 556 uses of these words is rather different from that of the 122 times these words occurred in the cited sample: the distributions correlate at 0.40 ($p<.01$). However, the distribution of the 971 co-words among the citing documents correlates much higher with that of the 333 co-words in the cited documents (0.56). This suggests a specificity for co-words over single word-occurences when comparing these two sets of documents.

The cluster analysis of the resulting co-word matrix in this case again shows a four-cluster structure, although some words have changed positions (Fig. 6). The first division in the two main groups of words now corresponds to the distinction between studies at the molecular level and studies at the cellular level. As can be expected, since some citations are shared among documents, factor analysis of the citing documents (Table 3) reveals more factorial complexity, i.e., more variables (words) are relevant to more clusters of words than in the case of the original documents. The striking point is that again four factors emerge as independent--oblique rotation gives factor pattern correlations below 0.07-- and that these factors correspond significantly to those in the originally cited set. (When the two factor matrices are compared as presented in Tables 2 and 3, they relate (not significantly) negatively (-.18) to each other. However, when the factors are ordered according to our designation of four research lines, this relation is raised to .75 ($p=0.001$). This relation is stronger than that between the co-word distributions (.63; see also Table 5.)) Only some words which are typically at the object level-- such as Salmonella Typhimurium or Bacteriorhodopsin-- now show none or only small factor loadings. In the case of 'Salmonella Typhimurium', for example, the work on the relevant enzyme system is being done by other researchers usings other micro-organisms, uncluding E. Coli.

We may now conclude that in the co-word structure of their titles, the citing documents are to a considerable extent congruent with the originally cited documents.

4.4 Title words in the Science Citation Index

Is this co-word structure a particular feature of this citing document set, or is it more generally the case that, for example, "sugar" and "enzyme" form a co-word ? To answer this question we have run the 45 words and their 990 possible combinations through the titles in the Science Citation Index for the same period (1979-1982). In these four years the SCI contained about 2.4 million titles, of which 337,234 (appr. 14.2%) contained one of the 45 title words used in our analyses. In these documents these 45 words occurred 410,144 times (1.2 per document), and formed 92,342 co-word linkages among them. Hence, we find considerably fewer co-word occurences than word-occurences, and thus one could argue that the use of co-words indicates a certain specificity. However, one should keep in mind that there are about 10^5 documents linked to the original set through co-words, against 190 through citations. Hence, there is a difference of more than two orders of magnitude in specificity between these two types of

Table 4 : Pearson correlation coefficients among word frequency distributions of 45 title words in :
- 23 biochemistry articles (TI23)
- 190 citing articles (TI190)
- the Science Citation Index title words 1979-1982 (SCI)
- the Science Citation Index cluster terms 1979-1982 (SCI)

	TI23	TI190	SCI	CT
TI23	1.0000 (0) p=****	.4046 (45) p= .003	-.0808 (45) p= .299	-.1176 (45) p= .221
TI190		1.0000 (0) p=****	.3077 (45) p= .020	.3197 (45) p= .016
SCI			1.0000 (0) p=****	.9724 (45) p= .001
CT				1.0000 (0) p=****

Table 5 : Pearson correlation coefficients among co-word distribution of 45 title words in :
- 23 biochemistry articles (TI23)
- 190 citing articles (TI190)
- the Science Citation Index title word 1979-1982 (SCI)
- the Science Citation Index cluster terms 1979-1982 (SCI)

	TI23	TI190	SCI	CT
TI23	1.0000 (0) p=*****	.5601 (45)	.2825 (45)	.2847 (45)
TI190		1.0000 (0) p=*****	.3576 (45) p= .008	.3866 (45) p= .004
SCI			1.0000 (0) p=*****	.9218 (45) p= .001
CT				1.0000 (0) p=*****

linkages, and it is a fair guess to state that the citing documents are more or less a subset of the co-word linked docuements. Thus, there is no "either ... or" between citation as a linkage and co-word linkage; citation is rather a further restriction to the set of documents which share occurrences of words and co-words.
The specific position of the citing documents can be further illustrated by the distributions of word- and co-word occurrences over resp. the original documents (N = 23), the citing documents (N = 190), and the SCI documents (N = 337,234). The correlations between these distributions are given in tables 4 and 5. (I will come back to the additional "CT" column in these tables in a moment.) In the first place, one notices the higher correlations among co-word distributions than among word-distributions between cited and citing document sets, indicating the specificity of co-word linkages. Secondly, the word distribution of the titles of the cited document set is not or is negatively (not significantly!) related to the overall document set of the SCI, while the citing documents have an intermediate position between the original set and the overall database. However, the co-word structure as found in the cited and citing sets cannot be retrieved in the database.
In summary, the conclusions are that title-words seem to offer a means of making visible the specific internal cognitive structure in a research group : in science, co-words are specific to the research lines which a particular research group may share with other research groups which cite their articles (see also [14]). Hence, they seem to allow us to partition research activities along intellectual lines. The external co-word linkages to the overall database, however, seem to be much less specific than citation linkages. We found in this case a difference of more than two orders of magnitude. Since we do not yet know what either of the operations precisely means, it may be more effective to stick to citation as a means of linking the original document set to the overall universe of scientific documents.

4.5 The indexer effect

Because the SCI is indexed with so-called control terms based on rather sophisticated clustering techniques [3], we saw an opportunity to do an experiment about what has been called "the indexer effect", to examine how much the influence of an external indexer intervens in the co-word structure when one uses index words instead of original words from titles. The results of the comparison are presented in the fourth column of tables 4 and 5 above.
In the SCI the use of index words seems to lead to about the same results as title words. The major difference is the quantitative amount of total co-words involved. Although about as many documents are involved (386,228 in this case as against 337,234 in the former) and proportionally about the same amount of word- occurrences (520,444 vs. 410,144), the total number of co-words generated among these documents with our 45 words is more than twice as large (202,781 vs. 92,342).
In an attempt to be more specific about the differences between the use of title- words and index-words, we checked the differences for three specific co-words, namely those composed of (i) "bacteriorhodopsin", (ii) "cyclic", and (iii) "thermo- dynamics.
In the subject field of the original document set the combination of "bacteriorhodopsin" and "thermodynamics" is meaningful and does occur once in the original document set (N =23). There is one more occurrence of this combination in the title-words of the whole SCI for this period. This occurrence happens to be an article from the other half of our original sample of 47 documents. This specific combination of title-words does not exist outside our group, and therefore not among the citing documents. However, when using index-words 7 other occurences can be found, of which two more belong to our original document set. The other documents are also closely related, but deal for example with "mitochondria" instead of "bacteriorhodopsin", thus illustrating the loss of

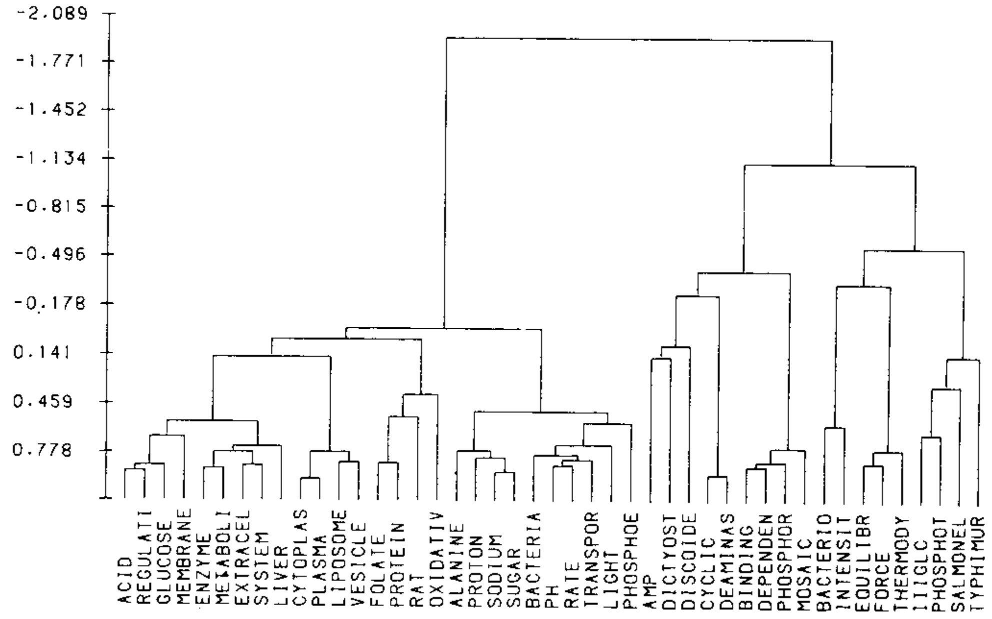

Figure 7 : Dendrogram for the co-words in the Science Citation Index, based on the cosine-formula.

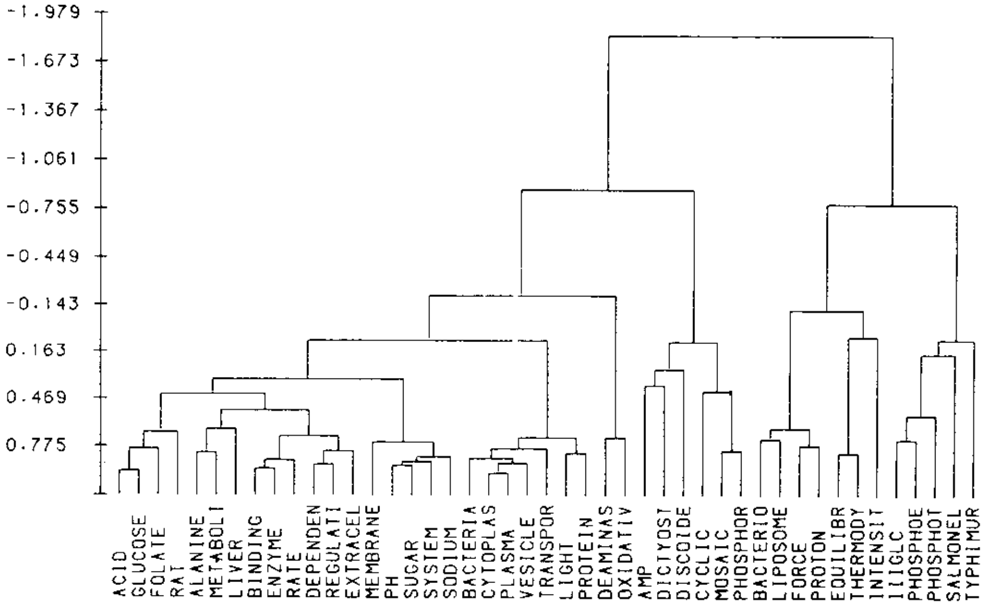

Figure 8 : dendrogram for the co-word of 45 key-words in the Science Citation Index

specificity when using index words.

The combination "bacteriorhodopsin" and "cyclic", which is conceptually impossible in our original document set, does not lead to any co-word linkages at the level of the ISI-database if one restricts the analysis to title-words. However, using control terms 15 hits are generated, of which most relate to completely different photochemical problems. In the case of the combination of "cyclic" and "thermodynamics",also highly unlikely in the field of our original documents, both control terms and title words give 7 hits - in this case the same - which all relate to the thermodynamic properties of "cyclic" compounds other than the "cyclic AMP" which is relevant for our group.

In our opinion, the main indexer effect is not that the indexing words would be inadequate or obsolete, but the quantitative effect produced by the fact that indexing essentially reduces the number of words involved by subsuming them under more general categories, so that the indexer increases the number of co-word linkages substantially because the smaller set is more strongly tied together than the larger. In addition to this, new and artifactual co-word linkages are generated.

Whether one considers this quantitative effect of the use of indexes as an advantage or a disadvantage of course depends on what one is after. From a scientometric point of view, however, we think that the introduction of bibliographic artifacts can hardly be justified. Given the lack of specificity of co-word linkages as compared to citations as noted in the former section, in most cases a further expansion of the universe of linked documents will be undesirable, and hence one should preferably stick to title words instead of key-words for co-word analysis.

Conclusions

1. Since co-occurrences of words seem to be highly specific for research lines, co-word linkages seem a good indicator of the internal structure of already coherent, i.e. selected, document sets.
2. However, as an instrument for bibliographic search, i.e. a search for the external links of a document or a document set, we found co-words to be more than 100 times less specific than citations. Particularly when index-words are used, one can expect a lot of noise in the results.
3. Indexing packs the word sets, so that more co-word linkages are found. Hence, searching with index words generates even more noise than searching with title words.
4. On the other hand, the citing document set has the same co-word structure as the cited document set. It can therefore grosso modo be regarded as a specific selection of the co-word linked document set.
5. Abstract-words have been found to be less specific than title-words, and hence less suited to the purpose of dividing the document set into its intellectual constituencies, or of linking the document set to the wider environment.

The use of co-word linkages among scientific documents seems particularly fruitful when one wants to distinguish among the intellectual specificities in a cognitively related document set. The structure which we found is probably maintained since authors are very much aware of the significance of picking the right words in their titles in order for their articles to be noticed by the intended audiences. From a scientometric point of view, indexes create bibliographic artifacts in this structure.

In terms of scientometrics, the structure discussed here is important since it allows us to describe - probably also dynamically - the structure of science at lower levels of aggregation than the structures of journal-journal citations, which we have discussed extensively elsewhere ([4], also [15]). However, when the main use of this indicator is indeed to partition document sets which belong to specific groups of documents, and not to search for external linkages, there is no reason to stick to co-words and to neglect lower- and higher-order correspondences; in such cases the analysis of word occurrences in documents seems more straightforward ([16]).

References

[1] S. Cozzens, Taking the measure of science : a review of citation theories. Newsletter of the International Society for the Sociology of Knowledge 7 (1981, May) 16-21.

[2] L. Leydesdorff, Towards a theory of citation ? A reaction to MacRoberts and MacRoberts. Scientometrics (forthcoming).

[3] H. Small, E. Sweeney, E. Greenlee, Clustering the Science Citation Index using co-citation. II Mapping science, Scientometrics 8 (1985) 321-340.

[4] L. Leydesdorff, Various methods for the mapping of science. Scientometrics 11 (1987) 281-320.

[5] E. Garfield, M.V. Malin, H. Small, Citation data as science indicators, in : Y. Elkana, J. Lederberg, R.K. Merton, A. Thackray, H. Zuckerman (eds), Towards a metric of science, (Wiley, New York,1978, p. 184).

[6] P. Healey, H. Rothman, ABRC science policy study 1983/4. Evaluative summary report. (ABRG, London, 1984).

[7] M. Callon, J. Law, A. Rip, Mapping the dynamics of science and technology. (MacMillan, London, 1986).

[8] J.-P. Courtial, Technical issues and developments in methodology, in : M. Callon, J. Law, A. Rip, Mapping the dynamics of science and technology. (Macmillan,London, 1986).

[9] O. Amsterdamska, L. Leydesdorff, What makes an article a significant contribution? (Science Dynamics, Amsterdam, 1986) (internal paper).

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

[11] H. Small, Structural dynamics of scientific literature. International Classification 3 (1976) 67.

[12] G. Salton, D. Bergmark, A citation study of computer science literature. IEEE transactions of professional communications 22 (1979) 146.

[13] H. Small, E. Sweeney, Clustering the Science Citation Index using co-citations. I A comparison of methods. Scientometrics 7 (1985) 391-409.

[14] S. Zeldenrust, Lines of research in organized science. Paper presented at the XIth annual meeting of the society of the social studies of science. Pittsburgh PA, (October 1986).

[15] L. Leydesdorff, The development of frames of references. Scientometrics 9 (1986) 103-125.

[16] L. Leydesdorff, Intellectual foci in research programmes. The use of word linkages as an indicator. Paper to be presented at the XIIth annual meeting of the society of the social studies of science (4S), Worcester, (November 1987).