Available online at www.sciencedirect.com

ScienceDirect

Journal of Informetrics 1 (2007) 287–307

Journal of INFORMETRICS
An International Journal

www.elsevier.com/locate/joi

Bibliographic coupling and its application to research-front and other core documents

Bo Jarneving

Swedish School of Library and Information Science, 501 90 Borås, Sweden

Received 11 June 2007; received in revised form 19 July 2007; accepted 23 July 2007

Abstract

Based on previous findings and theoretical considerations, it was suggested that bibliographic coupling could be combined with a cluster method to provide a method for science mapping, complementary to the prevailing co-citation cluster analytical method. The complete link cluster method was on theoretical grounds assumed to provide a suitable cluster method for this purpose. The objective of the study was to evaluate the proposed method's capability to identify coherent research themes. Applying a large multidisciplinary test bed comprising more than 600,000 articles and 17 million references, the proposed method was tested in accordance with two lines of mapping. In the first line of mapping, all significant (strong) links connecting 'core documents' (strongly and frequently coupled documents) in clusters with any other core document was mapped. This resulted in a depiction of all significant artificially broken links between core documents in a cluster and core documents extrinsic to that cluster. The second line of mapping involved the application of links between clusters only. They were used to successively merge clusters on two subsequent levels of fusion, where the first generation of clusters were considered objects for a second clustering, and the second generation of clusters gave rise to a final cluster fusion. Changes of cluster composition on the three levels were evaluated with regard to several variables. Findings showed that the proposed method could provide with valid depictions of current research, though some severe restrictions would adhere to its application.

Keywords: Bibliographic coupling; Science mapping; Cluster analysis

1. Introduction

This study aims at the elaboration of the appropriateness of bibliographic coupling as a form of document coupling for science mapping purposes and puts forward a method that combines bibliographic coupling with a cluster analytical method. This method was inspired by findings and suggestions in Glänzel and Czerwon (1995) where the concept of 'core documents' was introduced.

1.1. The co-citation analytical bibliometric model

The sub-field of scientometrics that focus on the mapping of science has widely been focused on the co-citation analytical approach which was independently introduced in 1973 by Small (Small, 1973) and by Marshakova (Marshakova, 1973). This form of document coupling was defined as the frequency with which two documents are cited together

E-mail address: bo.jarneving@hb.se.

1751-1577/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2007.07.004

and the co-citation strength as the number of identical citing items (Small, 1973). Through a number of subsequent pioneering articles (Griffith, Small, Stonehill, & Dey, 1974; Small & Griffith, 1974; Small & Sweeney, 1985) the method of 'co-citation bibliometric modeling' was developed to an intelligence tool, frequently used in science and policy applications. This model has been summarized as "A detailed representation of the structure and content of the international research front based on the strongly shared patterns of referencing among the current scientific literature papers" (Franklin & Johnston, 1988). Though frequently applied in the context of science mapping, this model has been criticized on both methodological as well as on theoretical grounds. Leydesdorff (1987) criticized the choice of methods preceding the model building and the exclusive focusing on the validation of the outcomes on behalf of the validation of methods. Leydesdorff meant that based on ad hoc hypotheses, which were basically wrong, Small and co-workers had assumed that ". . .the very existence of document clusters/. . ./is strong evidence for the specialty hypothesis" (Small & Griffith, 1974). Leydesdorff argued that this was a fallacious argument as cluster analysis always generates a cluster structure and that the real question was to determine what the structure represents. The methodological decisions were further criticized with regard to the choice of the single link cluster method, on grounds of not generating results consistent with results obtained by other analytical techniques.[1] The single link method is also known to produce loosely bound clusters and Leydesdorff suggested that results derived from the application of this method might well be artifacts of the method and not reflect the structure of science. From a research policy point of view, the concept of co-citation cluster analysis was also criticized by Oberski (1988), where one of several points of criticism was directed to the statistical instability resulting from both the application of the single link cluster method as well as the arbitrary application of threshold settings. Obersky concluded that ". . .it remains unclear how one could possibly distinguish between perhaps real effects from statistical effects" (Oberski, 1988, p. 448).

The claim that co-citation analysis is a useful tool to map subject-matter specialties was further examined by Braam, Moed, and van Raan (1991), who have developed a method using quantitative analysis of content-words related to articles. The authors investigated both documents grouped by the principle of co-citing documents of a particular cluster of co-cited documents, as well as the co-citation clusters themselves. The single link cluster method was applied for the grouping of co-cited articles. Based on findings, the authors concluded that the question if all topics covered by a data set can be identified by co-citation clustering can only partially be answered by comparing results for different sets of thresholds of (normalized) co-citation strength as some research areas might lack a consensual referencing. Still, findings suggested that co-citation clustering reflect research specialties, although these may be fragmented into several clusters. It was also found that co-citation clustering only partially revealed the literature relevant to identified research topics of the citing literature (a specialty's current work) and that interrelations between clusters seemed to correspond to cognitive relations on a higher level then research specialties.

The authors concluded that the method applied provides a useful instrument for the description and evaluation of co-citation analysis in terms of the cognitive content of clusters, cluster coherence and differentiation as well as the recall of specialties current citing articles.

1.2. Previous research on bibliographic coupling

Bibliographic coupling was introduced by Kessler to the scientific society through a number of reports and research articles in the 60s[2] and was primarily described as a method for grouping technical and scientific documents, facilitating scientific information provision and document retrieval. In one of the early reports, a general outline of the context in which an indexing method, concerned with countable indicators based on references, might operate was given (Kessler, 1960). In a subsequent report, the definition of bibliographic coupling was stated as: ". . .a single item of reference shared by two documents is defined as a unit of coupling between them" (Kessler, 1962). Based on this unit, two graded criteria of coupling were defined:

Criterion A: A number of articles constitute a related group G_A if each member of the group has at least one reference (one coupling unit) in common with a given test article, P_o. The coupling strength between P_o and any member of G_A

[1] "Other analytical techniques" refers to multidimensional scaling, factor analytical approaches and Ward's cluster method.

[2] The suggestion to group scientific documents on the basis of use rather than content was suggested independently by Fano (1956) and by Kessler (1958).

is measured by the number of coupling units between them. G_A^n is that portion of G_A that is linked to P_o through n coupling units (According to this criterion, there need not be any coupling between the members of G_A, only between them and P_o)
Criterion B: A number of articles constitute a related group G_B if each member of the group has at least one coupling unit with every other member of the group. The coupling strength of G_B is measured by the number of coupling units between its members. Criterion B differs from criterion A in that it forms a closed structure of interrelated articles, whereas criterion A forms an open structure of articles related to a test article (Kessler, 1962).

In a subsequent report dated 1962, Kessler applied bibliographic coupling to a test population of 40 documents from the field of radio engineering in order to test if a number of scientific documents bear meaningful relations to one another. He found that bibliographic coupling was able to partition this population into valid, related sub-groups. In order to be able to make any generalizations at all as to larger populations and other disciplines, he subsequently carried out an experiment where the automatic processing of a population of scientific documents (36 volumes of physical review) resulted in a grouping of 8521 articles in concordance with criterion A, confirming the existence of subject relatedness between bibliographically coupled documents (Kessler, 1963a).

In the same year, Kessler published another paper (Kessler, 1963b) where he further elaborates the application of bibliographic coupling in the context of information retrieval by trying to establish a factual background that could guide the design of an experimental science communication system. Bibliographic coupling was applied to a population of 8186 articles from the physical review and reported as 10 case histories, each illustrating an information retrieval problem. Different strategies of bibliographic coupling were applied, where the effects of enlarging or diminishing the search span by assigning P_os serially first (in the case of enlarging) or last (in the case of diminishing) in the list of the available literature were tested. Kessler concluded that bibliographic coupling can be applied to a large body of literature and that the process operates both in the future as well as in the past, relative to the position of P_o. This showed that bibliographic coupling could be used to identify the life span of a given literature.

In 1965, Kessler, still using data from the physical review for his experiments, compared groups formed according to the *Analytic Subject Index* and by bibliographic coupling. The aim of this experiment was to investigate how bibliographic coupling compare with results obtained by standard methods. He concluded that there was a high correlation between groups formed by bibliographic coupling and groups formed by analytical subject indexing. However, he pointed out that the report did not pass judgment on the utility of either method to any specific application (Kessler, 1965).

In a review article, Weinberg (1974) covered the major part written on bibliographic coupling up to the publication of her article. She concluded that "...at this point, bibliographic coupling does seem to be a useful tool for studying the 'science of science'—citation patterns, the useful life of literature, most cited journals etc.". However, she reflected on how citation behavior affects the standardization of the citation "unit" and put forward the meaning that the notion of "meaningful groups", claimed by those who advocates bibliographic coupling, may well constitute a problem. What she was trying to say was that since Kessler's experiments were done on documents in one field where there already was a meaningful group to begin with. Therefore, only a test on the scale of SCI would show if bibliographic coupling would work well in a complex and interdisciplinary environment.

However, the first attempt to test the validity and effectiveness of the bibliographic coupling technique for detecting subject relatedness between documents on a more heterogeneous population of documents and on a large scale was not performed until more than 20 years after Kessler's 1963 reports. One reason for this was probably the technical restrictions imposed by existing (at that time) computational resources and the problems in accessing large amounts of citation data. In 1984, Vladutz and Cook carried out an experiment with 10,000 randomly selected documents from the SCI which served as test documents for which bibliographically coupled publications from the entire 1981 database were sought. The large data file covering a multitude of scientific disciplines used in this experiment, corresponded well with Weinberg's claim in 1974 of an interdisciplinary environment as a prerequisite for the evaluation of bibliographic coupling. The questions to be answered with respect to this experiment concerned the frequency of bibliographic coupling links within the file and the degree to which these links are meaningful. It was found that, 90% of the input articles that have references yielded a group of at least two coupled items. Looking back at Kessler's experiments in the '60s, Vladutz and Cook wanted to test more extensively the hypothesis that strong bibliographic coupling links imply strong subject relatedness. The evaluation of subject relatedness was performed by small groups of experts with a scientific background and trained in assigning brief subject descriptions to groups of documents generated by co-citation clustering. Lists of 300 randomly selected test documents together with their strongest coupled articles were presented to the experts. It was found that in over 85% of the cases, the articles proved to be closely related by subject to the test documents. Vladutz and Cook concluded that the utilization of bibliographic coupling in a very large citation database was practically feasible and that valid results as to subject relatedness were achieved. The hypothesis stated in this research was that bibliographical coupling "...may prove to be the easiest approximation to an algorithm for revealing the semantically closest neighbours of publications".

A year before Vladutz and Cook published their results, Sen and Gan (1983) had published a purely theoretical document on bibliographic coupling. Their point of departure was a statement by Martyn (1964) where he argues "...that bibliographic coupling is not a unit but merely an indication of the existence of the probability, value unknown, of relationship between two documents".[3] The two researchers felt that in spite of the attention that previous works on bibliographic coupling had attracted, the method had hardly been taken seriously and that there was a need for a theoretical elaboration. With a point of departure in an $M \times N$ hypothetical Boolean matrix, where elements indicated a citation relationship between rows (citing documents) and columns (cited documents), the grouping of coupled documents in bibliographic *cliques* and *clusters* was elaborated. The notion "*clique*" is here equivalent to Kessler's grouping principle G_B, and "...clusters would be formed by the populations which have at least one member having coupling with another member whereas no member of one cluster will have coupling with any member of another separate cluster".

With regard to the central issue of cognitive resemblance between bibliographically coupled documents, a measure of coupling strength, the coupling angle (C.A.) was suggested. The coupling angle was expressed as:

$$\text{C.A.} = \frac{D_{\text{o}j} D_{\text{o}k}}{\sqrt{(D_{\text{o}j} D_{\text{o}j})(D_{\text{o}k} D_{\text{o}k})}}$$

where $D_{\text{o}j}$ and $D_{\text{o}k}$ are the Boolean vectors of document j, respectively, k. More precisely, the C.A. is defined as the cosine between two Boolean vectors, d_i and d_j, and can be obtained from their scalar product (Glänzel & Czerwon, 1995):

$$\text{C.A.}(d_i d_j) = \frac{d_i d_j}{|d_i||d_j|}.$$

The C.A. takes the maximum value of 1 if two Boolean vectors are parallel and 0 if they are rectangular. Two documents may be considered to be concerned with a related topic if the angle between vectors representing documents does not exceed a given angle θ ($0° \leq \theta < 90°$) (Glänzel & Czerwon, 1995). Lacking a theoretical basis as well as empirical evidence for the determination of a threshold of coupling strength, the Sen and Gan suggested a semi-arbitrary approach with cut off value of 0.5, which corresponds to $\theta = 60°$.

The question of cognitive resemblance related to bibliographic coupling was also pursued in Peters, Braam, and van Raan (1995). These researchers tried to find out whether relatively strong cognitive resemblance within groups of documents, bibliographically coupled by one and the same highly cited item, is present in an interdisciplinary field, i.e., chemical engineering. This was done by measuring word-profile similarities between the citing documents. It was found that word profile similarity within groups sharing a citation to a highly cited publication was significantly higher than between documents without such a relationship. Hence, such cognitive resemblance was found to exist, supporting the claim that these bibliographically coupled documents represented work of the same research specialty.

In Glänzel and Czerwon (1995, 1996), it was shown that bibliographic coupling can be used to identify "hot" research topics represented by so called "core documents", which were identified through the application of appropriate thresholds for both the number of common references as well as the strength of coupling links. Using the whole annual accumulation of the 1992 volume of SCI, about 1% of all documents was found to be core documents.[4] A detailed analysis of both key words in titles and indexing terms indicated the representation of important research front topics, and through several expert questionings, it was found that most core documents belonged to a few high impact documents of a specialty. The method presented proceeds from the model suggested by Sen and Gan (1983) and uses the C.A. as

[3] The meaning of this statement in short, is that the fact that two documents have a reference in common is no guarantee that both documents are referring to the same piece of information in the cited document. Hence, bibliographic coupling is only an indication of the existence of the probability of relationship between two documents.

[4] The data comprised 511,899 articles, notes and reviews, and only the document type "letters to the editor" was excluded on grounds of generally not belonging to research fronts.

a measure of the coupling strength. Glänzel and Czerwon restricted their analysis to a subset of coupled documents where each document was coupled with at least 10 coupling links with a minimum C.A. of 0.25 to other documents. The choice of thresholds was based on both theoretical considerations as well as empirical findings. According to the researchers, a lesser number of coupling links could bring about that documents published in series might influence results, whereas a greater number of coupling links would eliminate smaller research topics. They also claimed that a certain filtering of noise is necessary in order to avoid less characteristic coupling links between documents and that a value of the C.A. considerably lower than the stipulated would not accommodate this need. Also, a too high value of the C.A. would dramatically diminish the number of coupling links, leading to a serious decrease of documents.

The researchers concluded that documents connected by strong bibliographic coupling links can provide insights into the structure of research fronts and be applied for science mapping purposes. They also highlighted that bibliographic coupling has several advantages in comparison with co-citation clustering, the most important being the possibility to capture the early stages of a specialty's evolution. From this study, one can also see that there are two distinguished properties of a core document: (1) the tendency to be highly cited by subsequent articles and (2) the many and strong associations to other articles. Both (1) and (2) should be related to the fact that progressive research generally requires a not too small group of researchers with a common research focus.[5] Based on these two properties, one could consider core documents as important articles with an impact on research and central nodes in the scientific communication system.

One could conclude from the above that making use of core documents for bibliometric mapping is a good choice in view of their perceived impact on current research. Despite its favourable features, there is a distinct lack of evaluative research concerning bibliographic coupling applied as a science mapping method. The reasons for this unobtrusive position in science mapping are not obvious and comparable and complementary results to the co-citation approach have also been reported when this measure was applied for science mapping purposes (Jarneving, 2001; Persson, 1994; Sharabchiev, 1988). Also, a novel approach was reported by Janssens, Tran Quoc, Glänzel, and De Moor (2006), where several document–document similarity approaches were tested and compared. The study comprised text analysis, bibliographic coupling and approaches integrating text analysis with bibliographic coupling. Findings showed that the integrated approaches performed significantly better, perhaps indicating a future and fruitful line of application for bibliographic coupling.

2. Statement of purpose and research objectives

Based on the findings of the various researches so far, bibliographic coupling could be combined with a cluster method to provide a method of science mapping complementary to the prevailing co-citation cluster analytical method. The complete link cluster method would on theoretical grounds be suitable for this purpose, for more coherent clusters would be generated, meaning that it would not have the drawbacks of the single link cluster method (see Section 3). Thus, based on empirical evidences and theoretical considerations, bibliographic coupling and the complete link cluster method were combined to a mapping method which was evaluated in this study. The objective was set to evaluate the proposed method's capability to identify coherent research themes on basis of 'core documents', as defined by Glänzel and Czerwon (1995, 1996). The method to be devised has the following two primary components:

- A measure for the association of documents where the association can be expressed as the similarity between two documents.
- A cluster analytical method for the partition of sets (populations) of documents.

The measure of document similarity is needed for the purpose of establishing cognitive relationships between documents. The cluster method is needed for the partition of a set of documents into subsets of reciprocally similar documents. In this study, *bibliographic coupling* was applied as the measure of document similarity and the *complete link cluster method* was used for the clustering of documents.

[5] That is, for a specific research theme (a specialty), there should be a group of researchers building on each others findings, and this would be reflected by: (i) a common intellectual base literature, i.e., a large share of common references in their publications and (ii) a large share of the common references that are inter-citations within this group of researchers.

2.1. Design and research questions

The research design aimed at a large multidisciplinary test bed, comprising an annual volume of the SCI, where the specific objective was to identify and apply core documents for the evaluation of the proposed methods applicability as a mapping method. There are three major factors which motivated the research design. They are:

(1) The incidence of core documents.
(2) The properties of core documents.
(3) The properties of the complete link cluster method.

With regard to (1), the incidence of core documents should generally be low for a single field. Glänzel and Czerwon (1996) found that less then 1% of all items (4534 documents) in the 1992 volume of SCI were core documents. These were dispersed over 42 sub-fields and assigned a total of 128 journal subject categories. This dispersion of core documents over a large number of fields and specialties underlines the necessity of a multidisciplinary research setting.

With regard to points (2) and (3), considering the severe rule for merging of objects when the complete link cluster method is applied (see Section 3) and the role of core documents as central nodes in networks of bibliographically coupled articles, one could on theoretical grounds presume that core document clusters frequently may be parts of larger groups of related articles. In order to further elaborate the implications of this presumption, a strategy of mapping was outlined and applied. The strategy has its point of departure in a set of clusters generated by a first partition of the population of core documents. Here, only strong links was used for the clustering of core documents. This partition formed a base line from which two lines of mapping were pursued:

In *the first line of mapping*, all significant (strong) links connecting core documents in clusters with any other core document were mapped. This resulted in a depiction of all significant artificially broken links between core documents in a cluster and core documents extrinsic to that cluster. The rationale for carrying out this line of mapping was that it enables one to measure the extent of fragmentation of research themes the application of the proposed method may give rise to.

The *second line of mapping* involved the application of links between clusters only. They were used to successively merge clusters on two subsequent levels of fusion, where the first generation of clusters were considered objects for a second clustering, and the second generation of clusters gave rise to a final cluster fusion. The rationale for carrying out this second line of mapping is that larger specialties with complex internal structures may be mapped when the information in links between clusters is applied.

The impact of iterated clustering was regarded with respect to the overall cluster structure, with a starting point at the base line (first clustering). Changes of cluster composition on the three levels were evaluated with regard to the following variables:

(1) The internal coherence of clusters.
(2) The external isolation of clusters.
(3) The reduction of the number of clusters.
(4) The increment of cluster sizes.
(5) The number of isolated clusters.
(6) The number of singleton clusters.

Concerning points (1) and (2), the internal coherence and the external isolation of a cluster reflect the extent to which a cluster is consistent and demarcated with regard to the definition of similarity applied for the merging of articles to clusters. Points (3)–(4) reflect effects that a priori could be expected when applying iterated clustering.

With regard to the multidisciplinary aspect of this research setting, a comprehensive expert evaluation of cluster relevance would be impracticable. Therefore, the assessment of cluster relevance (i.e. cluster subject coherence) was grounded on statistical assessment of cluster properties. At the base line, clusters were assumed to be subject consistent. This was deemed reasonable as previous researches (e.g. Peters et al., 1995; Vladutz & Cook, 1984) have shown that strong bibliographic coupling links between research articles generally imply subject relatedness, only strong links between core documents was applied in the clustering and the use of the complete link cluster method will exclusively generate completely interconnected clusters. It was therefore presumed that changes of cluster coherence generally would mirror changes of cluster subject coherence. Likewise, changes of the external isolation are presumed to mirror the continuation or discontinuation of a specialty over levels of cluster fusion.

In order to complement findings, four cases of iterated clustering was presented to field experts, who were invited to evaluate and comment on the subject coherence and separation of clusters in terms of cluster relevance on different levels of cluster fusion. The selection of these cases aimed at finding examples from the dominant scientific fields, namely physics, chemistry and bio-medical science.

The following research questions were stated:

(1) To what extent does the proposed method impose a fragmentation of specialties, when applied for core document mapping?
(2) What is the impact of iterated clustering on the overall cluster structure?
(3) Is there an optimal level of cluster fusion?
(4) What are the implications of the results in 1–3 with regard to the application of the proposed method on core document data?

3. Methods and data

3.1. Measurement of proximity

The first task when the objective is to partition sets of objects for mapping purposes is to find a method for deciding the proximity between objects. The original definition of bibliographic coupling strength may not indicate the optimal measure of document similarity as the significance of a bibliographic coupling unit associating two articles should be inversely related to the combined lengths of the reference lists of both documents (Vladutz & Cook, 1984). Therefore, a function that normalizes for the length of reference lists is needed. This calls for the use of the C.A. The C.A. has been applied by several other researchers in the past (e.g. Glänzel & Czerwon, 1995, 1996; Mubeen, 1995; Sharada & Sharma, 1993). As shown by Glänzel and Czerwon (1995), the C.A. is a geometrical interpretation of *Salton's measure*, which was applied and defined as:

$$\mathrm{NCS}_{ij} = \frac{r_{ij}}{(n_i n_j)^{1/2}}, \tag{3.1}$$

where NCS_{ij} is the normalized coupling strength between article i and article j, r_{ij} the number of references common to both i and j, n_i the number of references in the reference list of article i and n_j is the number of references in the reference list of article j.

The interval is [0,1] and $n_i = n_j = r_{ij}$ gives the maximum value.

This function will be referred to as the normalized coupling strength (NCS) henceforth in this study.

3.2. The cluster analysis

When selecting an appropriate method of clustering, experience achieved and recorded by researchers within the field could to some extent be used as a guide. Originally, when the co-citation clustering method was developed by Henry Small and colleagues at the ISI (Griffith et al., 1974; Small, 1973; Small & Griffith, 1974; Small & Sweeney, 1985), the *single link* cluster method was applied. The defining feature of this method is that the distance between groups is defined as that of the closest pair of individuals. Single link clustering, is known to produce straggling and loosely bound clusters, especially in large data sets, and this problem might show up as a less clear structure due to this "chaining" phenomenon. Still, single link applications seem to have been successfully used by many researchers in the context of document co-citation analysis (e.g. Braam et al., 1991; Griffith et al., 1974; Small & Griffith, 1974, 1983; Small & Sweeney, 1985) and a variant was used by Persson performing an author co-citation analysis (1994). The single link method is easy to implement and use especially when large amounts of data is to be clustered. Comparing complete link clustering with single link clustering, the difference is how the distance between an existing cluster and a candidate for fusion with that cluster is defined. In complete link clustering, the largest distance between the candidate object and any object of the existing cluster is sought. This means that any candidate must be within a certain level of similarity to *all* members of that cluster. As mentioned, in single link clustering the shortest distance between clusters is sought. Hence, single link clustering and complete link clustering could be seen as each other's opposites. In addition to these methods, the *between group*s *average link* appears as an alternative. For this method, the distance between two clusters is the average of the distance between all pairs of individuals that are made up by one individual from each group. It was developed as an "antidote" to the extremes of both single and complete link (Aldenderfer & Blashfield, 1984, p. 40). Some general assumptions concerning differences between these methods can be made. The single link cluster method would generate more loosely bound clusters whereas the complete link cluster method would produce compact clusters and the group average link method something in-between.

In order to evaluate the extent to which significant links between clusters generated by iterated clustering still remained on the last level of cluster fusion, the between groups average cluster method was applied. The reason for this was that the use of the complete link cluster method implied too severe conditions to be fulfilled.

3.2.1. Choice and application of cluster method

If one can assume that a strong similarity between document A and document B and a strong similarity between document B and document C generally would suggest a strong similarity between document A and document C, a method of clustering with less severe conditions to fulfill may be appropriate (e.g. the single link method). This assumption, however, should be considered unconfirmed as previous research on co-citation clustering has shown that the chaining effect of the single link method has caused some undesirable effects like large subject inconsistent clusters when applied for co-citation cluster analysis. Therefore, a method that ensures that all objects in a cluster are within a set maximal distance to each other would be preferred in order to secure coherent clusters.

From a graph theoretical viewpoint, such groups could be considered complete undirected graphs. Such graphs would always have a maximal degree of interconnectedness, i.e., a maximal *density* (*D*), where *D* is defined as:

$$D = \frac{2(\#L(G))}{N(N-1)}, \tag{3.2}$$

where $\#L(G)$ is the number of edges connecting two vertices and N is the number of vertices (Otte & Rousseau, 2002).

The interval is [0,1] and the maximum value is reached when the value of $\#L(G)$ equals the value of $N(N-1)/2$. In this context, this means that the maximal value is reached when all possible document pairs in a cluster are bibliographically coupled. Applying the complete link cluster method, one will reach this objective as each cluster member is associated with every other member in a cluster, given that fusions of clusters at a level of zero association are disregarded. As the maximal interconnectedness is given by the method, only the strength of association (the distance) between documents varies.

Since all agglomerative hierarchical techniques reduce data to a single cluster containing all objects, the search for an optimal number of clusters demands a decision of when to stop. Usually partitions are achieved by cutting a dendrogram at a particular height, a "best cut" (Everitt, Landau, & Leese, 2001, p. 76). This requires that clear shifts of fusion levels are discernable. However, when the variance of proximity values between objects in a cluster is large, there may be no clear hierarchical structure, but in spite of that, clusters may still be subject coherent. A way to get around this problem is to set a minimal NCS, and when the complete link cluster method is applied, this means that all objects of a cluster have at least that strength of association with every other object in that clusters. This approach was applied in this study, and in line with the definition of a 'core document', all links with a NCS < 0.25 were filtered out. When a choice of partition has been accomplished, the distribution of documents over clusters may be skewed, with a majority of clusters constituted by one or two documents. As the goal of clustering is the arrival at some kind of meaningful summation of data in a smaller number of groups of objects, a confused pattern of numerous single objects and pairs would not contribute to such a goal. Hence, in this study, clusters containing less than three documents were excluded from further analysis.

3.3. Methods of evaluation

3.3.1. The qualitative assessment of cluster compositions

In this study, field experts were presented with data on Excel spreadsheets. On these spreadsheets, titles of articles in clusters were given as well as the hierarchical structure of clusters on three levels of cluster fusion. For each cluster on the first level of clustering, any article not in line with the identified research focus of the cluster was marked. Next, the relevance of the compound cluster on the next level of fusion was assessed. When all constituent clusters on the second level of cluster fusion had been evaluated, the relevance of the merging of these clusters on the last level of cluster fusion was assessed. In this way, not only the relevance in terms of misplaced articles was assessed, but also the relevance of the fusion of sub-clusters. This was deemed important as a common research theme may emerge by the fusion clusters, otherwise not clearly discernable on sub-cluster level. The field experts were asked to provide appropriate comments on these issues.

In order to visualize cluster compositions and provide with illustrations facilitating the comprehension of field experts interpretations, maps based on multidimensional scaling (MDS) were produced. MDS could briefly be described as set of mathematical techniques that enable a researcher to uncover the "hidden structure" in a data set or a data base. In essence, by locating analyzed objects in a spatial configuration (a 'map'), one seeks to determine the theoretical meaning of this representation. For each map presented in this study, a statistic called "stress" is presented. Stress is a measure of how well the configuration represents the data of the original matrix. A high value of stress is due to a low degree of correspondence between the original data of the matrix and the configuration on the map.

3.3.2. The quantitative assessment of cluster compositions

Many authors have attempted to define a cluster in terms of its internal cohesion and external isolation (Everitt et al., 2001, p. 6). Ideally, a cluster should therefore be internally coherent and externally well separated, meaning that it should contain articles that are reciprocally, strongly bibliographically coupled, but lacking (strong) bibliographic couplings with articles in other clusters.

A measure of the internal coherence is the average coupling strength, AvgCS(C), for a cluster C. It is defined as:

$$\mathrm{AvgCS}(C) = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathrm{CS}(d_i d_j)}{\binom{n}{2}}, \tag{3.3}$$

where n is the number of articles in a cluster C, CS the number of bibliographic coupling units between two articles, d_i, d_j and $d_i d_j$ ($\in C$).

Complementary to this measure is the aforementioned D. These two measures of cluster coherence reflect different aspects of internal cohesion and it is possible that a cluster could have an average coupling strength that is relatively high and a relatively low score of cluster density, and vice versa. The first case would occur if a cluster contained a number of articles coupled with strong links but a large share of all possible pairs of articles were not coupled. The second case, i.e., a high-density cluster with a relatively low average coupling strength would occur if all, or most articles, were coupled but with a weak coupling strength. Hence, the need for both measures was acknowledged, as they do not substitute for one another.

In order to elaborate on the aspect of external isolation of clusters, a third measure is needed. Let C and C' be clusters of sizes k and m, respectively. *The average coupling strength between two clusters*, C and C', AvgCS(C, C'), is defined as:

$$\mathrm{AvgCS}(C, C') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m}\mathrm{CS}(d_i, d_j)}{k \times m} \tag{3.4}$$

where CS is the number of bibliographic coupling units between two articles, d_i, d_j and d_i, d_j and $d_i \in C$, $d_j \in C'$.

All three measures are needed for the establishment of cluster relevance from a quantitative viewpoint. The cluster coherence provided by (3.2) and (3.3) is needed for the identification of coherent research themes, whereas the separation between clusters provided by (3.4), is needed for the mapping of the discontinuation or continuation of a research theme. In addition, (3.4) was applied for the iterated clustering, where links between clusters were used to successively merge clusters on the two subsequent levels of fusion (see Section 2.1). The methods chosen in this study for the assessment of cluster relevance may be considered robust and traditional, focusing on coherence and separation. However, new, more advanced methods have been developed (see Lamirel, Francois, Shehabi, & Hoffmann, 2004), where concepts from mathematics (Galois lattice) and information retrieval (recall and precision) are applied. Such methods may also be suited for the evaluation of document clusters generated by the method presented here.

4. Results

A multidisciplinary research setting was constructed in order to identify the crop of core documents in a year's accumulation of research articles. From the SCI volume 2003 on CDROM, 619,570 items of the document type "articles" were downloaded. The average number of references in a core document was 28. A total of 17,674,944 references were processed, resulting in 149,198,407 bibliographic coupling units distributed over 121,968,904 links. The number of links was next delimited to only comprise links with a NCS of ≥0.25, which resulted in a reduction to 267,034 links. In these links, 6060 unique core documents were identified and constituted a final set for the analysis. The following notation will be used when referring to the different levels at which clusters were generated:

- C1 denotes the level at which clusters were generated by the first clustering.
- C2 denotes the level at which clusters were generated by the second clustering.
- C3 denotes the last clustering between C2-clusters.

The result section consists of three sub-sections: Section 4.1 describes and summarizes the statistical findings with regard to cluster fusions on three consecutive levels, C1–C3. In Section 4.2, the assumed effect of fragmentation is elaborated by mapping links external to clusters on the C1-level. In Section 4.3, the expert evaluation of four different C3-clusters and their compositions is presented. Though Section 4.2 (first line of mapping) empirically preceded Section 4.1, the order of presentation is reversed for clarity.

4.1. Cluster fusions

In order to assess the association through bibliographic coupling between core documents, the NCS between all 6060 core documents of the original population was computed. Links with a NCS lower than 0.25 were filtered out and 5771 articles were clustered by the complete link cluster method. A total of 1761 clusters were generated of which 228 were singleton clusters. In all, 5543 core documents were merged to 1533 clusters varying in size between 2 and 22. One thousand clusters had a size ≥3 and contained in total 4477 core documents. The median cluster size was 4 and approximately 78% of all core documents were contained within these clusters. These 1000 clusters were selected for further analysis and fusion.

For the 1000 C1-clusters containing at least three articles, the distribution of AvgCS(C) was near symmetrical with a mean AvgCS(C) of 10.59. With regard to the aspect of isolation 49 core document clusters were isolated. The distribution of AvgCS(C, C') was positively skewed with a median of 0.65, isolated clusters excluded. Though the median was relatively low, a large number of clusters were strongly connected. Counting links where the distance between C1-clusters > Q3[6] (where the AvgCS(C, C') > 3.11), 1628 links connecting 706 C1-clusters are found. This shows that a large share of clusters on the C1-level was connected by relatively strong links.

4.1.1. The first cluster iteration, the C2-level

On basis of the AvgCS(C, C') between C1-clusters containing at least three articles, 6537 links connecting 951 C1-clusters were applied for an iterated clustering. The clustering of the 951 C1-clusters resulted in 153 singleton clusters and 212 clusters varying in size between 5 and 97 articles. In total, 3524 core documents were contained in the set of 212 C2-clusters. These were selected for further analysis and fusion. The distribution of core documents over 212 clusters was positively skewed with a median cluster size of 13. At the upper tail of the distribution, nine macro clusters with a size >50 was found. These were foremost large physics clusters, but two were from the bio-medical sciences.[7]

The distribution of AvgCS(C) was still rather symmetrical, though the mean AvgCS(C) had dropped from 10.59 on the C1-level to 7.95. With regard to D, calculated as the density of links between articles in C2-clusters, the distribution was extremely negatively skewed and most clusters on the C2-level still formed complete graphs. The median of D was 1.0 (mean = 0.98) and the range of D was set by a lowest value of D of 0.60.

[6] Q3 denotes the third quartile, i.e., the score that divides the bottom three quarters of the distribution from the top quarter.

[7] The following fields were presented by macro cluster in accordance to their sizes: particle physics, condensed matter, crystallography, applied physics and, endocrinology and oncology.

Focusing on the extent to which C2-clusters were separated from one another, the AvgCS(C, C') between C2-clusters were calculated and a total of 23 isolated clusters were found. The distribution was positively skewed with a median of 0.02. At the tail of the distribution, a few links between C2-clusters indicated that a few relatively strong associations between clusters remained also on this level. Counting links where the distance between C2-clusters > Q3 (where the AvgCS(C, C') > 0.0625), 179 links connecting 145 C2-clusters were found. In all, this means that there was a drastic reduction of strong links and number of connected clusters on this second level of cluster fusion.

4.1.2. The last cluster iteration, the C3-level

The application of the complete link cluster method on the last level of cluster fusion resulted in a partition where numerous singleton clusters and a few clusters containing two objects were generated only. This means that an upper limit for the application of iterated clustering was found for the proposed method. Still, the question if links between clusters generated on the C2-level were able to form relevant clusters on the last level of cluster fusion needed to be answered. In order to be able to map such links, the between groups average cluster method was applied. On basis of the computed AvgCS(C, C') between C2-clusters, 189 C2-clusters were partitioned into 92 singleton clusters and 38 clusters contained more than 1 C2-cluster. The total sum of articles in the 38 clusters was 1763. This distribution was positively skewed, with most cluster sizes gathered at the lower range of the scale. The median cluster size was 37. Macro clusters ($N \geq 86$) are seen at the higher range of the scale and at the upper tail. The macro clusters are again from the field of physics (condensed matter, particle physics, crystallography and applied physics) and from the bio-medical sciences (oncology–haematology). One physics cluster (condensed matter) builds partly on one of the macro clusters formed at the second fusion level, and one cluster from the bio-medical sciences (oncology–haematology) on two of the macro clusters formed at the second fusion level. Otherwise, macro clusters were generated by merging medium and smaller sized C2-clusters.

For the 38 C3-clusters, the AvgCS(C) was calculated and an almost rectangular distribution was generated. Moving up to this level of cluster fusion, the mean AvgCS(C) drops from 7.95 to 3.64. Likewise, the density of links between core documents D makes a drop from near the maximum value to a mean of 0.66 (md = 0.64).

4.1.3. Summary of cluster fusions

Applying the complete link cluster method on core document data resulted in a first partition where the majority of clusters were relatively small and the median cluster size was 4 for the selected set of clusters with a size ≥3. On each fusion level, a share of clusters that did not fulfill the requirements for cluster fusion emerged. This way, by each level of cluster fusion, C1–C3, the original set of core documents was reduced as the sizes of clusters increased. The stepwise loss of core documents and simultaneous increase in cluster size is presented in Table 1.

Concerning the aspect of external cluster isolation, by each level of fusion, the share of isolated clusters was increased and the strength of association between clusters was weakened. At the same time, the internal cluster coherence was weakened too. Comparing levels, the most drastic change with regard to the separation between clusters take place when moving up to the C2-level, while the most drastic change with regard to cluster coherence takes place when moving up to the C3-level (Table 2).

On the C1-level, clusters were internally coherent but many clusters were still associated with other clusters through relatively strong links. On the second level of cluster fusion, the internal coherence remained strong and at the same time clusters were generally more isolated, while on the last level of cluster fusion, the drop of cluster coherence was considerable, indicating the generation of more subject inconsistent clusters.

Table 1
Three levels of cluster fusion: effects on document populations, frequency of clusters and cluster sizes

Fusion level	No. of clustered core documents	No. of clusters	Median cluster size
C1	4477	1000	4
C2	3524	212	24
C3	1763	38	37

The calculation of median cluster size does not include singleton clusters. On the C1-level, clusters have a minimal size of three articles and on the C2- and C3-levels, clusters are composed by at least two objects (clusters from earlier fusion levels).

Table 2
Three levels of cluster fusion: effects on cluster coherence and cluster separation

Fusion level	Percentage isolated clusters	Median AvgCS(C, C')	Mean AvgCS(C)	Median D
C1	5	0.65	10.58	1.00
C2	11	0.02	7.95	1.00
C3	16	0.00	3.64	0.64

4.2. Extrinsic links and fragmentation

The finding of coherent clusters on the C2-level indicated the breaking up of specialties when this method was applied solely on the C1-level. In addition, the generation of C2-clusters did not cover for all associations between core documents in a C1-cluster and core documents extrinsic to it as the partition in clusters itself breaks up links. Mapping all links between core documents in a cluster and core documents extrinsic to the cluster with a minimal NCS of 0.25, it can be illustrated that core document clusters on the C1-level constitute fragments of larger research themes. Computing all such links with a NCS ≥ 0.25, the ability of C1-clusters to expand was assessed. In this experiment, the expansion of clusters was tried on all clusters with a size >1.[8] It was found that on the average, a cluster could expand by eight times its size, consequently 12.5% of the articles in an expanded cluster typically constituted the original cluster (Fig. 1).

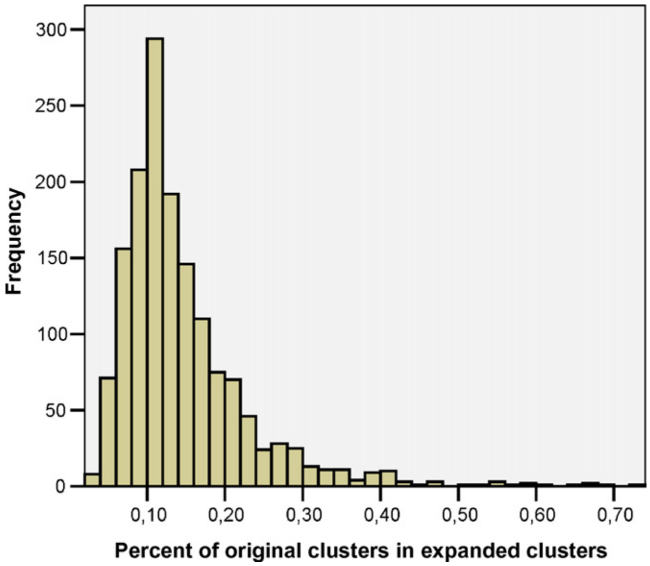

Fig. 1. Percent of original (C1) clusters in expanded (C2) clusters.

As can be seen from Fig. 1, only a few core document clusters constitute 50% or more of the expanded clusters. The median of the distribution is 0.125, in line with the mentioned factor 8 when calculating the size of an original cluster. The product-moment correlation between original cluster sizes and share of original articles in expanded clusters was strongly positive with a value of +0.64. This means that articles in larger original clusters to a lesser extent were associated with articles extrinsic to the original cluster. Conclusively, it has been shown that a large number of core documents can be added to core document clusters on the C1-level of fusion by tracking strong links between core documents. This is in line with the suggestion that the mapping of articles linked to core documents would facilitate the coverage of whole research fronts (Glänzel & Czerwon, 1995, 1996). In this case, only strong links to other core documents were applied and a strong subject relationship between core document clusters and the added core

[8] This means links between core documents in C1-clusters with a size >1 and all other 5771 core documents from the first level of cluster fusion.

documents could be presumed. The examination of a several samples of such expanded clusters did not contradict this presumption. As an example, cluster C1/203 could be expanded. Originally this cluster was composed of three articles, all on bio-rhythms. Articles are presented by article number, title, journal title and journal subject category as follows:

- 321110/Light and Circadian Regulation in the Expression of Lhy and Lhcb Genes in Phaseolus-Vulgaris/*Plant Molecular Biology*/biochemistry & molecular biology; plant sciences.
- 321536/The Circadian Clock—A Plants Best Friend in a Spinning World/*Plant Physiology*/plant sciences.
- 401249/Light-Regulated Translation Mediates Gated Induction of the Arabidopsis Clock Protein Lhy/*EMBO Journal*/biochemistry & molecular biology.

In this cluster, between 14 and 16 common references connect the bibliographically coupled pairs of papers and a total of 10 references are common to all papers (the total number of references for a pair in brackets). They are:

- 15 (88) 321110-321536;
- 16 (88) 321110-401249;
- 14 (99) 321536-401249.

This cluster is linked to 16 other papers extrinsic to the cluster with a NCS of at least 0.25 through a total of 28 links as follows:

- 8/321110;
- 10/321536;
- 10/401249.

Expanding the cluster on basis of these links, the cluster could be depicted as an incomplete graph, where the density, *D*, is decreased to 0.18 from the default value of 1.0 (see Fig. 2).

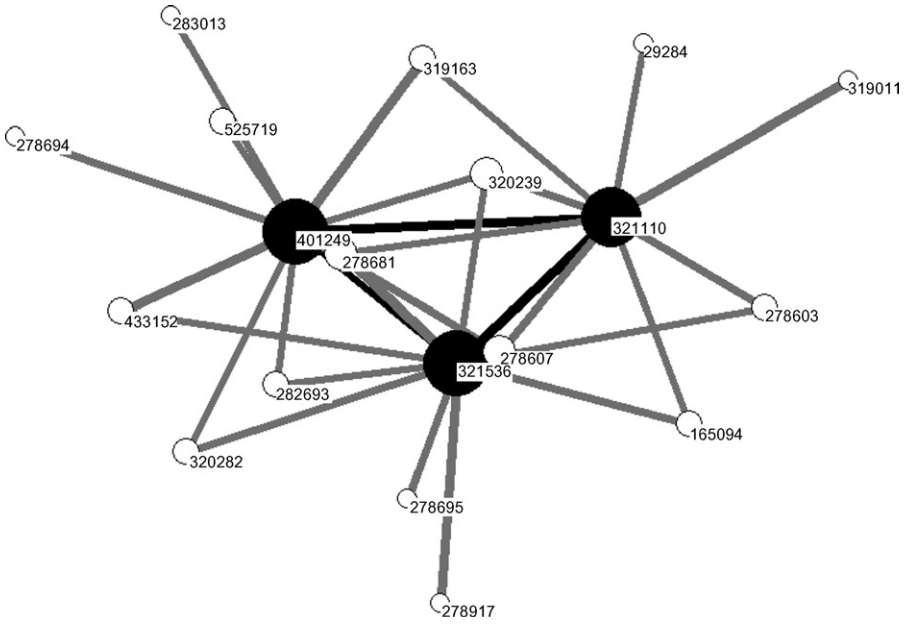

Fig. 2. The expansion of C1/203 depicted by MDS. Cluster C1/203 expanded with 16 unique links to an incomplete graph of 31 edges and 19 vertices. Sizes of circles representing clusters are proportional to the number of links in which a core document occurred (in the expanded cluster) and the width of connecting lines to the NCS. Darker lines connecting darker circles depict the original complete sub graph (cluster C1/203). D for the incomplete graph (with regard to the applied threshold of NCS) C1/203 was 0.18 and Kruskal's stress 0.06.

Table 3
The 16 core documents by which cluster 203 was expanded

29284/Surface-Plasmon Resonance Spectroscopy (Spr) Interaction Studies of the Circadian-Controlled Tomato Lhca4-Asterisk-1 (Cab-11) Protein with Its Promoter/**biology; physiology**
165094/Suite of Photoreceptors Entrains the Plant Circadian Clock/**biochemistry & molecular biology; plant sciences; cell biology**
278603/Arabidopsis Pseudo-Response-Regulator7 Is a Signaling Intermediate in Phytochrome-Regulated Seedling Deetiolation and Phasing of the Circadian Clock/**biochemistry & molecular biology; plant sciences; cell biology**
278607/The Time-for-Coffee Gene Maintains the Amplitude and Timing of Arabidopsis Circadian Clocks/**biochemistry & molecular biology; plant sciences; cell biology**
278681/Comparative Genetic-Studies on the Aprr5 and Aprr7 Genes Belonging to the Aprr1/Toc1 Quintet Implicated in Circadian-Rhythm, Control of Flowering Time, and Early Photomorphogenesis/**plant sciences; cell biology**
278694/The Evolutionarily Conserved Osprr Quintet—Rice Pseudo-Response Regulators Implicated in Circadian-Rhythm/**plant sciences; cell biology**
278695/Characterization of the Aprr9 Pseudo-Response Regulator Belonging to the Aprr1/Toc1 Quintet in Arabidopsis-Thaliana/**plant sciences; cell biology**
278917/Response Regulator Homologs Have Complementary, Light-Dependent Functions in the Arabidopsis Circadian Clock/**plant sciences**
282693/2 Arabidopsis Circadian Oscillators Can Be Distinguished by Differential Temperature Sensitivity/**multidisciplinary sciences**
283013/Circadian Phase-Specific Degradation of the F-Box Protein Ztl Is Mediated by the Proteasome/**multidisciplinary sciences**
319011/The Novel Myb Protein Early-Phytochrome-Responsive1 Is a Component of a Slave Circadian Oscillator in Arabidopsis/**biochemistry & molecular biology; plant sciences; cell biology**
319163/Dual Role of Toc1 in the Control of Circadian and Photomorphogenic Responses in Arabidopsis/**biochemistry & molecular biology; plant sciences; cell biology**
320239/A Link Between Circadian-Controlled Bhlh Factors and the Aprr1/Toc1 Quintet in Arabidopsis-Thaliana/**plant sciences; cell biology**
320282/Cell Autonomous Circadian Waves of the Aprr1/Toc1 Quintet in an Established Cell-Line of Arabidopsis-Thaliana/**plant sciences; cell biology**
433152/The Arabidopsis-Srr1 Gene Mediates Phyb Signaling and Is Required for Normal Circadian Clock Function/**developmental biology; genetics & heredity**
525719/Fkf1 Is Essential for Photoperiodic-Specific Light Signaling in Arabidopsis/**multidisciplinary sciences**

Core documents are presented with article numbers, article titles and journal subject categories. Subject categories are in bold style.

The titles and journal subject categories of the added core documents are presented in Table 3, which clearly presents a shared focus with the original research theme in cluster C1/203.

4.3. Experts' evaluations

As a complement to the statistical study, four cases of iterated clustering were presented to four different field experts for evaluation. The design of this experiment aimed at finding clusters on the last level of cluster fusion from the three major science fields: physics, chemistry and bio-medicine for the evaluation. C3-clusters from these fields were then matched against profiles of researchers and when a match occurred, a preliminary choice of cluster was made. If a researcher was available to do the evaluation, a final choice of cluster was made. In order to approximate the greater impact of physics on the composition of the database underlying the experiments, two cases from the field of physics and one each from the other two fields were selected. In all, a total of 154 core documents were evaluated. Corresponding to each case is a C3-cluster, which is assumed to reflect research themes with cognitive linkages to one another. The field experts were asked to assess the relevance of cluster composition on all three levels of cluster fusion, i.e., C1–C3. The maps based on MDS presented here were not made available for the evaluators.

4.3.1. Cluster C3/12, "human genetics and disease"

This C3-cluster contained 53 core documents distributed over 3 C2-clusters and 11 C1-clusters as follows:

- C2/45: C1/616; C1/1003;
- C2/46: C1/1171; C1/1297; C1/1170; C1/1172;
- C2/210: C1/9; C1/1163; C1/1168; C1/1184; C1/1286.

The focus in this cluster is generally on human genetics and disease. A total of 21 different journal subject categories were assigned to the journals in which core documents in this cluster were published, the most frequent category being

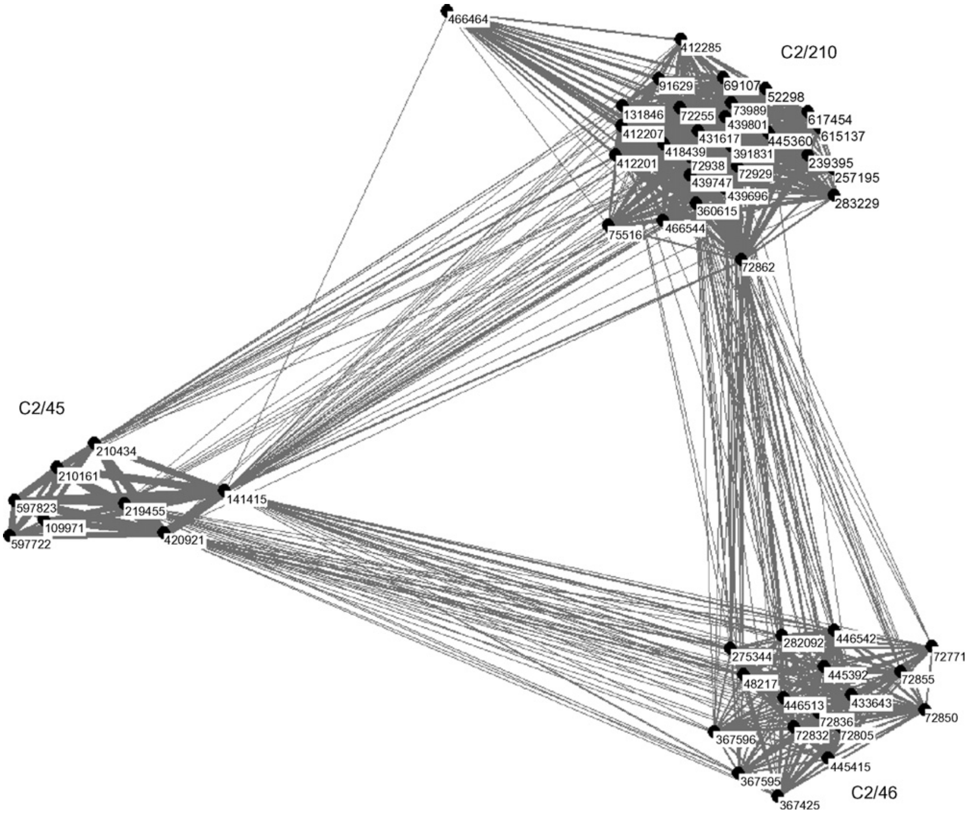

Fig. 3. The configuration of core documents in cluster C3/12. Kruskal's stress was 0.06.

Genetics & Heredity (22), followed by Gastroenterology & Hepatology (13). In the graph representing C3/12, each of the three C2-clusters are clearly depicted and demarcated as can be seen from Fig. 3. The density *D* of the graph was 0.52 and the AvgCS(*C*) 2.85. Hence, both values of cluster coherence were below the average.

The field expert's opinion was that core documents in C1- and C2-clusters were subject related, with two exceptions. The first exception is article 466464 in C2/210/C1/1184 which seemed to have too general subject content in relation to the pronounced focus in C2/210 on inflammatory bowel-diseases. This was in agreement with its more peripheral position on the map. The second exception was core document 283229 in C2/210/C1/1163, which the expert assumed to be relevant, but with some uncertainty as the title was not exhaustive enough. The field expert renounced judging the relevance of merging disease-gene-mapping methods (C2/46), with research on genetic aspects of psoriasis (C2/45) and inflammatory bowel-disease (C2/210).

4.3.2. Cluster C3/19: "Chemistry"

This cluster contained 25 core documents distributed over two C2-clusters and five C1-clusters as follows:

- C2/111: C1/1189; C1/1263;
- C2/191: C1/81; C1/406; C1/1394.

All core documents but one in C3/19 pertained to the field of chemistry and the composition of contributing disciplines varied, though with an emphasis on organic chemistry. A total of five different journal subject categories were assigned the journals in which core documents in this cluster were published the most frequent being Organic Chemistry (17) followed by Multidisciplinary Chemistry (6). In the graph depicting C3/19, the composition of C2/191 is visualized as a compact cluster whereas C2/111 is a looser construct, and there exists no links between C1/1263 and C2/111 (see Fig. 4). In spite of the latter, the coherence of C3/19 is above the average with 5.32 for the AvgCS(*C*) and 0.67 for *D*.

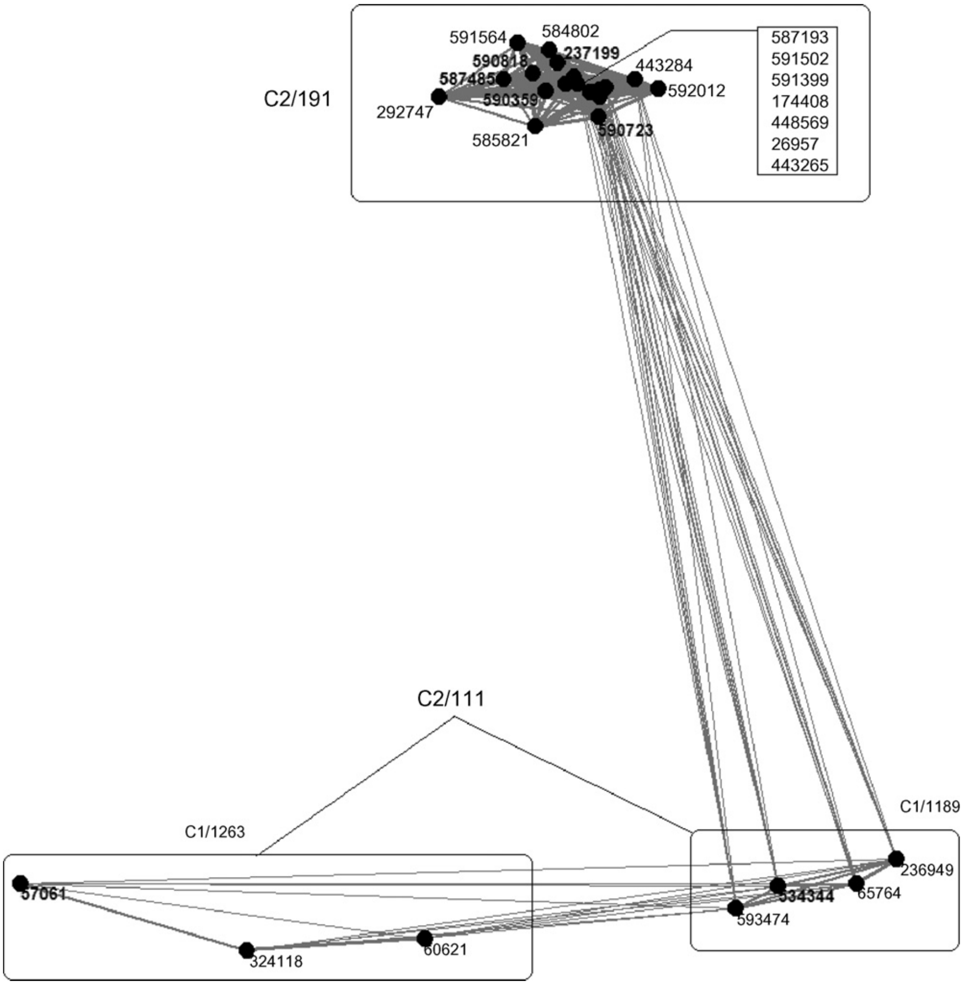

Fig. 4. The configuration of core documents in cluster C3/19. Due to the compactness of C2/191, seven labels representing articles could not be fitted to mark corresponding circles of cluster C2/191 and are therefore presented in the nearby table in the map. Kruskal's stress was 0.02.

According to the field expert, no misplaced core documents were found on the C1-level. However, C2/111/C1/1263 was found to be more diverse than cluster C2/111/C1/1189, which is reflected by the configuration in the map where C1/1263 forms a looser structure. On the C2-level, C2/111 was considered to be subject consistent in terms of a research focus common to the constituent C1-clusters. As for C2/191, the partition in C1-clusters appeared artificial to the field expert and C2/191 was better regarded as one cluster, which is reflected by this cluster's compactness, as seen in the map. Regarding the merging of the C2-clusters, no clear subject relationship between them was obvious.

4.3.3. Cluster C3/27: "Bose-Einstein condensation"

This cluster contained 54 core documents distributed over 4 C2-clusters and 13 C1-clusters as follows:

- C2/140: C1/367; C1/555;
- C2/141: C1/459; C1/557; C1/578;
- C2/143: C1/353; C1/362; C1/439; C1/552;
- C2/144: C1/352; C1/359; C1/454; C1/643.

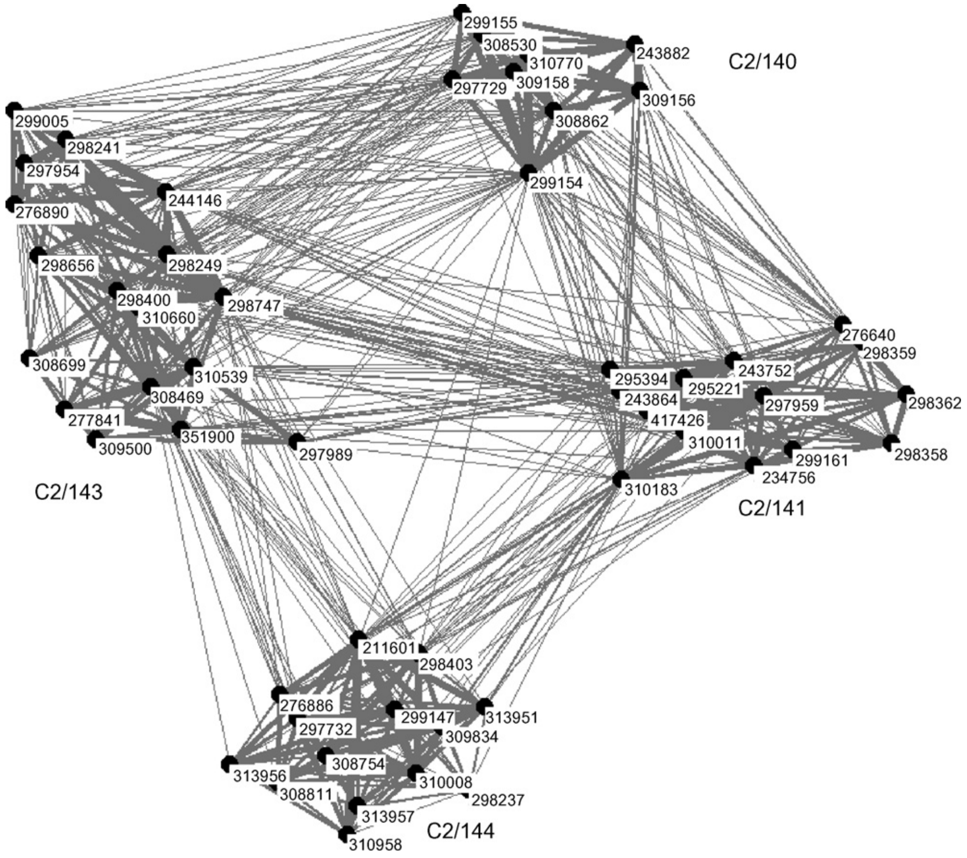

Fig. 5. The configuration of core documents in cluster C3/27. Kruskal's stress was 0.08.

Articles in this cluster pertain to research areas of optical, atomic and molecular physics and the major focus is on Bose-Einstein condensation.[9] A total of five different journal subject categories were assigned to the journals in which core documents in this cluster were published, the three most frequent were optics (19), atomic, molecular and chemical physics (19) and multidisciplinary physics (16). In the graph depicting C3/27, each of the four C2-clusters is clearly demarcated (Fig. 5). The density D is 0.43 and the AvgCS(C) 1.97, hence both values are clearly below the average.

The field expert presented in this case an elaborated evaluation where not only misplaced core documents on the C1-level were considered, but also minor deviations between their research foci:

- In cluster C2/140/C1/367, core document 308862 caused some uncertainty as the subject content as reflected by its title was not completely transparent.
- In cluster C2/140/C1/155, core document 299154 had a somewhat deviating focus in comparison with other cluster members.
- In cluster C2/143/C1/353, core document 298656 had a slightly deviating focus in comparison with other cluster members. Also, core document 308469 and core document 308699 cohered, but were considered somewhat deviating in relation to core document 277841 and core document 308699, which formed a coherent pair. Hence, this cluster "sprawled" slightly in terms of cluster coherence.
- In cluster C2/143/C1/362, core document 297989 deviated somewhat from the other core documents in C3/27.
- No core document deviated to the extent that it should be considered as clearly misplaced.

[9] Bose-Einstein condensation is the collapse of atoms into a single quantum state.

Concerning C2-clusters, the field expert's opinion was that all constituent C1-clusters shared the same research focus, hence, all C2-clusters belonged to the same area of research. Conclusively, some deviations on the C1-level were detected and when core documents were aggregated to higher levels, a common research theme for all core documents in C3/27 is seen.

4.3.4. Cluster C3/29: "Carbon-nano-tubes"

This cluster contained 22 core documents distributed over two C2-clusters and five C1-clusters:

- C2/27: C1/1018; C1/1416;
- C2/28: C1/549; C1/1072; C1/1137.

Core documents in C3/29 focus on carbon-nano-tubes (CNTs) from different angles.[10] A total of 11 different journal subject categories were assigned the journals in which core documents in this cluster were published, the more frequent was analytical chemistry (5), applied physics (5) and condensed matter physics (4). In the graph depicting C3/29, a complex and less clear cluster structure is reflected. Hence, the division of the map in C2-clusters and the subdivision in C1-clusters are not clearly mirrored by the configuration of the graph representing C3/29 (see Fig. 6). The density *D* was 0.70 and the AvgCS(*C*) was 2.45. Hence, the general level of interconnectedness is above the average, but the average strength of links is below the average.

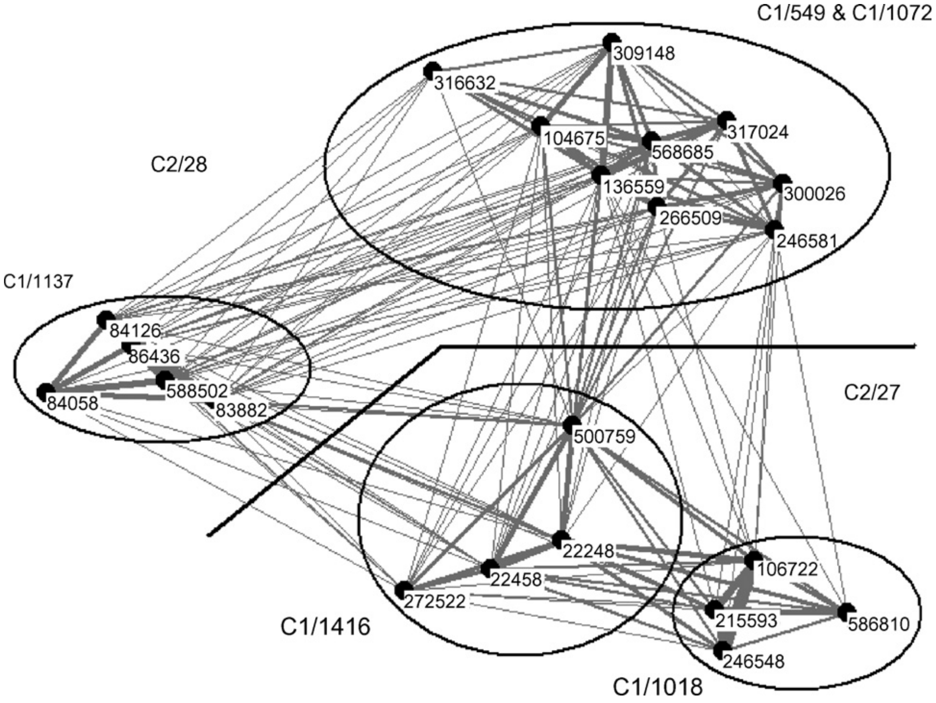

Fig. 6. The configuration of articles in cluster C3/29. The angled line dividing the map indicates the border between cluster C2/27 and cluster C2/28. Kruskal's stress was 0.07.

The more complicated structure was also reflected in the field expert's evaluation. To begin with, cluster C1/1416 contained one misplaced core document (500759) as did cluster C1/549 (core document 246581). Moving to the C2-level, in C2/27, both C1/1018 and C1/1416 handled CNT growth, though C1/1018 focused on the growth of aligned CNT on patterned substrates whereas C1/1416 was about non-aligned growth. In C2/28 (containing C1/549, C1/1072

[10] Carbon-nano-tubes are cylindrical carbon molecules with properties that make them potentially useful in extremely small scale electronic and mechanical applications. They exhibit unusual strength and unique electrical properties, and are efficient conductors of heat.

and C1/1137), C1-clusters focus on CNTs from divergent perspectives with no obvious common theme which would justify their fusion to a C2-cluster. Concerning the subject relationship between cluster C2/27 and cluster C2/28, all C1-clusters explicitly focused on CNTs except for C1/1137 where the interest in CNTs was deemed secondary.[11] However, the field expert renounced the evaluation of the relevance of merging the C2-clusters.

These examples can be related to the statistical evaluation cluster fusion. On the C1-level, clusters were generally considered subject coherent. On the C2-level, one C2-cluster (C3/29) was considered artificial but the others relevant. On the C3-level, the evaluation of the merging of two (out of four) C3-clusters was renounced, and one of was considered irrelevant. Hence, field experts' evaluations were in line with statistical findings.

5. Discussion

The point of departure in the following discussion is in the final set of core documents containing 4477 articles. This set was generated by a gradual reduction of the original set of 6060 core documents by 26% when thresholds of NCS and cluster size were applied.

5.1. The extent of fragmentation imposed by the applied method

It was shown that the applied method leads to a fragmentation of research themes. On the average, a core document cluster could increase its size by a factor of eight and only a few clusters were expanded by less than half their sizes. The effect of fragmentation was illustrated by example where it also was shown that the adding of core documents brings about a decrease of cluster coherence, measured as *D*. Hence, the expansion of clusters is at a cost of a presumably diminished relevance.

5.2. The impact of iterated clustering on the overall cluster structure

With the starting point in a large set of smaller clusters, the fusion of clusters at two subsequent levels showed an increasing loss of core documents as larger aggregations of core documents were formed. This loss was due to the generation of singleton clusters and isolated clusters emerging at each level of cluster fusion. At each subsequent level of cluster fusion, the general tendency was that an increase of cluster size was followed by a decrease of cluster coherence as well as an increase of the isolation of clusters. This means that increasingly less relevant clusters were formed but also that clusters got more isolated.

5.3. The optimal level of cluster fusion

It was found that the second level of cluster fusion (C2) should be the optimal level. This could be concluded on the following grounds:

- On the first level of cluster fusion (C1), a large share of clusters was associated with other clusters through relatively strong links.
- On the second level of cluster fusion (C2), the internal coherence remained strong and at the same time clusters were generally more isolated.
- On the last level of cluster fusion (C3), the drop of cluster coherence was considerable, indicating the generation of more subject inconsistent clusters.

Field experts' evaluations were in line with these findings.

[11] Cluster C2/1137 focused primarily on film-electrodes though all but one core document title had the term "carbon nano tubes" in the title.

5.4. Implication of findings

Though empirical findings speak in favor for the second level of cluster fusion as the most appropriate one, it was shown that some clusters on the C1-level were near complete in terms of extrinsic associations and that a few clusters on the C3-level may be relevant. The proposed method is, however, not likely to be applicable on the last level of cluster fusion (C3), though cluster fusion on the third level may be tried for the mapping of links between disciplines. Moreover, the quite severe loss of core documents generated by iterated clustering would require data from the preceding levels if a more comprehensive mapping should be accomplished. Hence, it is strongly suggested that at least the two first levels of fusion are applied, including singleton clusters and isolated clusters and that mapping results be interpreted from bottom to top (or top bottom) as the cluster merging itself contain important information. The assessed effects of fragmentation implies that the proposed method when applied for core document clustering do not identify and map research themes exhaustively, but rather smaller cores of referencing consensus. Also, findings showed that approximately a quarter of the final population of core documents where lost when clustered, given the applied minimum size of clusters.

It could be suggested that the proposed method may be used as a navigating and information seeking tool. Several applications may be successful, and one can directly be outlined on basis of the findings in this study. With a starting point in a complete graph, the additional expansion by significant links could be used to monitor the radiating associations of articles related to the core of a specific research theme. When additional information of cluster affiliation of such associated articles is added, the navigation in and between scientific structures would be facilitated. The navigation could be geared by varying threshold settings, deciding the maximum radius from each core.

5.5. Reflections on the results in relation to previous research

Several results connect to previous findings and theoretical considerations in the literature on bibliographic coupling and co-citation cluster analysis. First, claims that the method of bibliographic coupling is capable of associating documents that have a similar research focus (e.g. Peters et al., 1995; Vladutz & Cook, 1984) was confirmed by the perceived high degree of relevance in clusters generated by the complete link method. The application of the complete link method, in line with coupling criterion B, originally suggested by Kessler (1960) and the notion of cliques put forward by Sen and Gan as one type of bibliographically coupled document groups (1983), resulted in small but compact and generally subject consistent clusters. Hence, the problems of macro clusters encountered in co-citation cluster analysis (cf. Griffith et al., 1974) were avoided. The effect of fragmentation, or more precisely, the split up of research themes in smaller clusters, also encountered in co-citation cluster analysis (Braam et al., 1991), was conspicuous. The issue of fragmentation is also related to the setting of thresholds of coupling strength and the higher the threshold, the more severe the effect of fragmentation. These problems were approached by Small and co-workers (Small & Sweeney, 1985) by implementing variable level clustering in order to find the best cluster solution. The problem of fragmentation was also recognized by Braam, Moed and van Raan in their 1991 article and they recognized that the question of coverage of topics demands a comparison of cluster solutions on different levels of thresholds.

The problems of threshold setting were avoided in the case of core document mapping, as previous empirical findings guided the setting of these (cf. Glänzel & Czerwon, 1995, 1996). However, the conditions to fulfill for a core document implies the mapping of larger fields with a high publication and citation activity, hence the applicability of the proposed method for core document mapping when smaller research fields are analyzed should be limited. Also, it would be most desirable with more empirical research concerning the impact, i.e., the subsequent frequent citation, of core documents as well as theoretical elaborations of such a mechanism (cf. Glänzel & Czerwon, 1996).

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). Cluster analysis. In *Quantitative applications in the social sciences, Vol. 44*. London: Sage Publications Inc..

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping science by combined co-citation and word analysis 1: Structural aspects. *Journal of the American Society for Information Science*, *42*(4), 233–251.

Everitt, B. S., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). London: Arnold.

Fano, R. M. (1956). *Document in action*. New York: Reinhold Publishing Corporation.

Franklin, J. J., & Johnston, R. (1988). Co-citation bibliometric modeling as a tool for S&T and R&D management: Issues, applications, and developments. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology*. Amsterdam: North Holland.

Glänzel, W., & Czerwon, H. J. (1995). A new methodological approach to bibliographic coupling and its application to research-front and other core documents. In *Proceedings of 5th International Conference on Scientometrics and Informetrics* (pp. 167–176).

Glänzel, W., & Czerwon, H. J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, *37*(2), 195–221.

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

Janssens, F., Tran Quoc, V., Glänzel, W., & De Moor, B. (2006). Integration of textual content and link information for accurate clustering of science fields. In *InSCit2006, current research in information sciences and technologies: Multidisciplinary approaches to global information systems, Vol. I*. Badajoz: Open Institute of Knowledge., pp. 615–619.

Jarneving, B. (2001). The cognitive structure of current cardiovascular research. *Scientometrics*, *50*(3), 365–389.

Kessler, M. M. (1958). *Concerning some problems of intrascience communication*. Massachusetts Institute for Technology, Lincoln Laboratory.

Kessler, M. M. (1960). *An experimental communication center for scientific and technical information*. Massachusetts Institute for Technology, Lincoln Laboratory.

Kessler, M. M. (1962). *An experimental study of bibliographic coupling between technical papers*. Massachusetts Institute for Technology, Lincoln Laboratory.

Kessler, M. M. (1963a). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25.

Kessler, M. M. (1963b). Bibliographic coupling extended in time: Ten case histories. *Information Storage and Retrieval*, *1*, 169–187.

Kessler, M. M. (1965). Comparison of the results of bibliographic coupling and analytic subject indexing. *American Documentation*, *16*(3), 223–233.

Lamirel, J.-C., Francois, C., Shehabi, S. A., & Hoffmann, M. (2004). New classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. *Scientometrics*, *60*(3), 445–462.

Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics*, *11*, 295–324.

Marshakova, I. V. (1973). System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya Seriya 2-Informatsionnye Protsessy I Sistemy*, *2*(6), 3–8.

Martyn, J. (1964). Bibliographic coupling. *Journal of Documentation*, *20*, 236.

Mubeen, M. A. (1995). Bibliographic coupling: An empirical study of economics. *Annals of Library Science and Documentation*, *42*(2), 41–53.

Oberski, J. E. J. (1988). Some statistical aspects of co-citation cluster analysis and a judgment by physicists. In A. F. J. van Raan (Ed.), *Handbook of quantitative studies of science and technology*. Amsterdam: North Holland.

Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for the information sciences. *Journal of Information Science*, *28*(6), 441–453.

Persson, O. (1994). The intellectual base and research front of JASIS 1986–1990. *Journal of the American Society for Information Science*, *45*(1), 31–38.

Peters, H. P. F., Braam, R. R., & van Raan, A. F. J. (1995). Cognitive resemblance and citation relations in chemical engineering publications. *Journal of the American Society for Information Science*, *46*(1), 9–21.

Sen, S. K., & Gan, S. K. (1983). A mathematical extension of the idea of bibliographic coupling and its applications. *Annals of Library Science and Documentation*, *30*(2), 78–82.

Sharabchiev, Y. T. (1988). Comparative analysis of 2 methods of cluster analysis of bibliographic citation. *Naucno-techniceskaja informacija/2* (in Russian).

Sharada, B. A., & Sharma, J. S. (1993). A study of bibliographic coupling in linguistic research. *Annals of Library Science and Documentation*, *40*(4), 25–137.

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

Small, H., & Griffith, B. (1974). The structure of scientific literatures I: Identifying and graphing specialties. *Science studies*, *4*(1), 17–40.

Small, H., & Griffith, B. (1983, July). The structure of the social and behavioural sciences literature. In S. Schwarz (Ed.). *Stockholm papers in library and information science*. Stockholm: Royal Institute of Technology Library, TRITA-LIB-6021.

Small, H., & Sweeney, E. (1985). Clustering the Science Citation Index using cocitations I. A comparison of methods. *Scientometrics*, *7*(3–6), 391–409.

Vladutz, G., & Cook, J. (1984). Bibliographic coupling and subject relatedness. *Proceedings of the ASIS Annual Meeting*, *47*, 204–207.

Weinberg, B. H. (1974). Bibliographic coupling: A review. *Information Storage and Retrieval*, *10*(5/&), 189–195.