The Outbreak of SARS Mirrored by Bibliometric Mapping: Combining Bibliographic Coupling with the Complete Link Cluster Method

Dr. Bo Jarneving
The Swedish School of Library and Information Science
at Göteborg University and Högskolan i Borås
E-mail: bo.jarneving@hb.se

Dr. Bo Jarneving is a senior lecturer at the Swedish School of Library and Information Science in Borås. He defended his doctoral thesis "The combined application of bibliographic coupling and the complete link cluster method in bibliometric science mapping" *in 2005 at the Göteborg University. His research fields are science mapping methodology, citation analysis and scientometrics and his ongoing projects involve evaluation of citation based mapping methods and research evaluation (visibility and production).*

Abstract

In this study a novel method of science mapping is presented which combines bibliographic coupling, as a measure of document-document similarity, with an agglomerative hierarchical cluster method. The focus in this study is on the mapping of so called 'core documents', a concept presented first in 1995 by Glänzel and Czerwon. The term 'core document' denote documents that have a central position in the research front in terms of many and strong bibliographic coupling links. The identification and mapping of core documents usually requires a large multidisciplinary research setting and in this study the 2003 volume of the Science *Citation Index* was applied. From this database, a sub-set of core documents reporting on the outbreak of SARS in 2002 was chosen for the demonstration of the application of this mapping method. It was demonstrated that the method, in this case, successfully identified interpretable research themes and that iterative clustering on two subsequent levels of cluster agglomeration may provide with useful and current information.

Keywords: *Scientometrics, Bibliometrics, Science Mapping, Bibliographic Coupling, Complete Graphs, Hierarchical Clustering*.

1. Introduction

In this study a novel method that combines bibliographic coupling with a cluster analytical method was applied. The development of this method was in part inspired by findings and suggestions by Glänzel and Czerwon (1995 and 1996) where the concept of 'core documents' was presented, and in part by graph theory. So far, the application of bibliographic coupling for scientometric mapping purposes has been very meagre in comparison with the dominating cocitation analytical approach, though valid results have been sporadically presented (Sharabchiev, 1988; Persson, 1994; Jarneving, 2001).

Bibliographic coupling was introduced by Kessler to the scientific society through a number of reports and research articles in the'60s in the context of scientific information provision and document retrieval. 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" (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 is measured by the number of coupling units between them. G^n_A 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.

In a subsequent number of articles, Kessler tested Criterion A with regard to the method's applicability in the context of information retrieval (IR) (1963a, 1963b). Later Kessler augmented the testing of this method's usefulness by comparing groups formed according to the *Analytic Subject Index* with groups generated by bibliographic coupling (1965).

In 1983 Sen and Gan published a purely theoretical paper on bibliographic coupling. With a point of departure in an *M* x *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:

$$C.A. = \frac{(D_{oj} \bullet D_{ok})}{\sqrt{(D_{oj} \cdot D_{oj})(D_{ok} \cdot D_{ok})}}, \qquad \textbf{(1.1)}$$

where the C.A. is the cosine of the angle for citing documents *j* and *k* where D_{oj} and D_{ok} are their binary vectors. The C.A. is thus a geometric interpretation where the C.A. takes on the maximum value of 1 if two binary vectors are parallel and 0 if they are rectangular (90°). Lacking a theoretical basis as well as empirical evidence for the determination of a threshold of coupling strength, Sen and Gan suggested a semi-arbitrary approach with cut off value of 0.5, which corresponds to $\theta = 60^{\circ}$.

The testing of the validity and effectiveness of the bibliographic coupling on a large scale and in a multidisciplinary environment was first performed in 1984 by Vladutz and Cook (1984) wanting to test the hypothesis that strong bibliographic coupling links imply strong subject relatedness. These authors carried out an experiment on basis of a large and randomly selected set of documents from the *Science Citation Index* (SCI) database in which bibliographically coupled publications from the entire 1981 database were sought. The goal of this research was to establish the frequency of bibliographic coupling links and the degree to which these links were meaningful. Vladutz and Cook could conclude that the application of bibliographic coupling on a large scale was feasible and that valid results can be reached.

The issue of document-document similarity through 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. Support was found for the presumption that bibliographically coupled documents were subject related.

In 1995 Glänzel and Czerwon, presented the idea that bibliographic coupling should be used for the identification of current research topics in the research front. (1) These topics were assumed to be sufficiently represented by so called "core documents". Core documents were identified through the application of appropriate thresholds for (1) the number of common references and (2) the normalized coupling strength between articles. Applying a volume of the *SCI* as the test bed, analysis of both key words in titles and indexing terms indicated the representation of important research front topics, and through expert judgements, it was found that most core documents belonged to a few high impact documents of a specialty. Most important, they also found that core documents, on the average, received significantly more citations than the average article. Their method takes its point of departure in the model suggested by Sen and Gan (1983), applying the C.A. as the measure of document-document similarity. Their definition of core documents implied that only document coupled with at least ten other documents at a minimum C.A. of 0.25 were considered. This choice of thresholds was based on theoretical considerations as well as on empirical findings. The authors concluded that documents connected by strong bibliographic coupling links may provide insights into the structure of research fronts and be applied for science mapping purposes. They also emphasized that bibliographic coupling has some advantages in comparison with cocitation clustering, the most important being the possibility to capture the early stages of a specialty's evolution.

In agreement with ideas presented by Glänzel & Czerwon (1995, 1996) and Zen & Gan (1983), the objective of this study was to present a science mapping method that combines cluster analysis with bibliographic coupling and to illustrate the application of this method on a current research problem-area. The additional theoretical aspects of this method pertain to graph theory and multivariate statistical analysis.

2. Method

First, on basis of the previous researches presented, we can conclude that bibliographic coupling provide us with a satisfactory measure of document-document similarity. It may not be the optimal measure in all contexts (i.e. various IR-scenarios), but there are strong indications that it is suitable for mapping purposes. Our next concern is that of relevance or

significance (in a general sense) of bibliographic coupling links between two documents. Clearly, some shared references may be random occurrences and need to be filter out. We may apply the metaphor of 'signal versus noise' and claim that we only need those associations that are strong enough to be considered 'signals' and that all other association should be regarded 'noise'. Applying the thresholds suggested by Glänzel & Czerwon (1995, 1996), this means that we zoom in on the most central and strongly coupled documents in the research front, hence, only a minor fraction of documents would be selected for further analysis. On the other hand, should we wish for a more comprehensive and far-reaching mapping, lower thresholds would be applied. In this study, however, we aim at the mapping of documents that fulfill the requirements for 'core documents', meaning a C.A. ≥ 0.25 and a minimum of ten such links for each document.

The next issue is to consider what type of groups of interconnected documents that would be useful from an information provision perspective. Here, we may apply a graph theoretical approach and re-connect to ideas presented by Kessler and by Sen & Gan. As previously quoted, Kessler (1962):

> …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.

Hence, Kessler elaborated on a type of document group that secures homogeneity in terms of a complete interconnectedness of documents. As mentioned, the notion of bibliographic 'cliques' elaborated on by Sen & Gan (1983), is analogous to Criterion B. In graph theoretical terms, such groups are considered 'complete graphs'. Moreover, the bibliographic coupling association between two documents is clearly symmetrical; hence, we may use the term 'undirected graphs'. We may illustrate this in detail, but first some definitions: An undirected graph *G,* is constituted by a set *V* of vertices and a set *E* of edges such that each edge $e \in E$ is associated with an unordered pair of vertices. The existence of an unique edge *e* associated with the vertices *v* and *w*, implies the existence of an edge *e* associated with the vertices *w* and *v* and this is written as $e = (v, w)$ or $e = (w, v)$ (Johnsonbaugh, 1997). In Figure 1, *G* is constituted by the set $V = \{a, b, c, d\}$ of vertices and the set $E = \{e_1, e_2 \ldots, e_5\}$ of edges. Furthermore, a graph *G′* whose vertices and edges form subsets of the vertices and graph edges of a graph *G,* is a sub graph of *G*, and *G* is said to be a super graph of *G′*. A complete graph is a graph in which each pair of vertices is connected by an edge (ibid.). In Figure 2, subsets of *V* and *E* constitute the sub graph *G′*, which also is a complete graph.

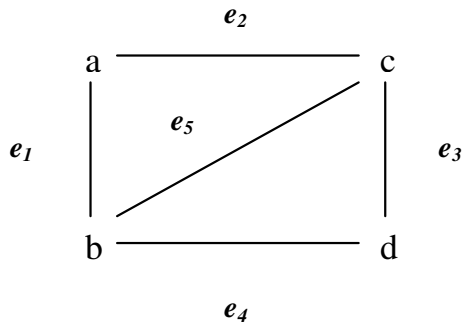

FIGURE 1: The incomplete undirected graph *G*.

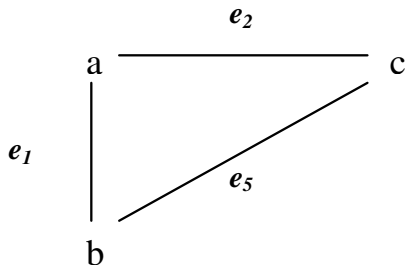

FIGURE 2: The complete sub graph G′ of the undirected Graph *G* in Figure 1

Such complete graphs would always have a maximal degree of interconnectedness, i.e. a maximal *density* (*D*), where *D* is defined as:

$$D = \frac{2 \cdot (\# L(G)}{N(N-1)}, \qquad \textbf{(2.1)}$$

where

#L(*G*)= the number of edges connecting two vertices, and
N = 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. Hence, striving for the generation of complete graphs of strongly bibliographically coupled core documents, we would aim at the identification of 'cognitive cores' in the super graphs of the total research front network. The usefulness of such cores for science information provision is quite obvious.

Next, we need to find a method capable of identifying such groups in the research front. The Criterion B, the notions of 'cliques' and complete, undirected graphs correspond to a principle of agglomeration in hierarchical cluster analysis, the complete link clustering. In hierarchical agglomerative clustering the agglomeration of clusters starts with the most similar pair of objects and clusters are subsequently merged in order of similarity. How to define the similarity between clusters hence defines the cluster method. In complete linkage the similarity between two clusters, *X* and *Y*, is measured as the smallest similarity between x_i and y_i, where x_i is a member of *X* and y_i a member of *Y*. Thus, given a set minimum similarity > 0 between object-pairs, this means that all $n(n\text{-}1)/2$ object pairs in a cluster generated by the complete link method are bibliographically coupled, constituting a complete graph.

From a mere practical point of view, the exclusive identification of such coherent document groups implies cutting of informative links between documents in different groups, possibly indicating the continuation of a research theme. Therefore, in order to more comprehensively map also this aspect, a method of iterated clustering (clustering of clusters) was applied. This means that we need to construct an additional measure that approximates the similarity between the generated clusters. Such a measure is the average coupling strength between two clusters, *C* and *C'*, AvgCS(*C*, *C'*): Let *C* and *C'* be clusters of sizes *k* and *m*, respectively, then we can define this measure as follows:

$$AvgCS(C,C') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} CS(d_i,d_j)}{k \times m}, \qquad \textbf{(2.2)}$$

where

CS = number of bibliographic coupling units between two articles,
d_i, d_j and $d_i \in C, d_j \in C'$

This measure of cluster similarity was applied for two subsequent clusterings. For the sake of clarity, we label the different levels of clustering as follows:

- **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 distances (the inverse similarity) between clusters, or between articles in clusters, generated by clustering are not comprehensible when presented as mere figures in a table. A better understanding of the pattern of distances between clusters is arrived at when all distances are configured in a more than one-dimensional space. A method that is able to generate such displays of distances is MDS.

MDS could be summarised as a method for solving the problem of how to represent *n* objects geometrically by *n* points, so that the distances between points correspond to experimental dissimilarities or similarities between objects (Kruskal, 1964). By locating objects as points in a spatial configuration, one seeks to determine the theoretical meaning of this representation.

Briefly, this is how MDS works. The input for the analysis should be an N·N symmetrical matrix containing proximity data. First, an object is indexed primarily by the letter I and secondarily by the letter j, and one assumes objects to run from 1 to N if there are N objects. Proximity data values connecting object i with object j are represented by δ_{ij} and distance data values between objects will be noted as d_{ij}. The central motivating concept, then, is that the distances d_{ij} should correspond to the proximities δ_{ij}, for example, by a linear function f where $f(\delta_{ij}) = d_{ij}$. As this correspondence may not be perfect, meaning a perfect monotone relationship between proximities and distances, such discrepancies, $f(\delta_{ij}) - d_{ij}$, are measured by a goodness of fit function (Kruskal & Wish, 1978, p. 24). The scaling starts with a random configuration and through a number of iterations the configuration is changed in order to find the optimal fit with the experimental proximities. Measuring how well the fitted distances match the experimental proximities, a so called "stress value" is arrived at. The stress ranges from 0-100% and a stress value of zero consequently mean that for every δ_{ij}, $f(\delta_{ij}) = d_{ij}$. A stress ranging from "excellent" to "good" is then expressed as a value from 0.025 to 0.05 inclusive, according to Kruskal (1964). Different *n* dimensional solutions are possible to choose from, but usually a two dimensional configuration is selected, given reasonable stress values.

Conclusively, MDS is a systematic procedure for obtaining a geometric configuration, or a "map", which has a certain relationship to the proximity data (Kruskal & Wish, 1978, p. 12). It is applied in this study with the intention of visualizing clusters' internal structures in two dimensional spaces.

3. Data

The method of bibliographic coupling is somewhat sensitive to the length of the publication period, and an increase of the distance in time between bibliographically coupled articles leads to a diminishing pool of shared references due to the tendency to cite the more current articles. In Glänzel & Czerwon (1996), an observation period of ½ - 2 years was suggested on these grounds. From the SCI volume 2003 on CDROM, 619,570 items of the document type "articles" were downloaded. Next, the number of links was delimited to only comprise links with a NCS of ≥ 0.25, which resulted in a reduction to 6,060 core documents, constituting a final set. Conclusively, given applied thresholds, we aim at the identification of cognitive cores in the approximate 2003 research front, as mirrored by data in the SCI.

4. The Empirical Investigation

Enormous amounts of information can be derived from the huge original data set applied. In this study we delimit the investigation to comprise one example of iterated clustering of core documents over three levels of agglomeration (C1-C3). Some background information should be provided though.

A total of 1,761 clusters were generated of which 228 were singleton clusters. Of the original 6,060 core documents a total of 5,543 core documents were merged to 1,533 clusters varying in size between 2 and 22. This is in itself an interesting result, as one can not per see assume that core documents are linked to each other, though the setting of a minimum of ten links makes this plausible. Hence, only 8.5 percent of the original document population exclusively connected outside this set of selected documents. It was found that 1000 clusters, corresponding to a total of 4,477 core documents, were of a size ≥ 3. These were selected for further analysis.

Summing up the effects of the iterated clusterings (C1-C3), the first partition resulted in a set of relatively small clusters with a median size of 4 for the selected set of clusters with a size ≥ 3. On each subsequent fusion level, a share of clusters that did not fulfil 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.

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

Level of agglomeration	No. of clustered core documents	No. of clusters	Median cluster size
C1	4,477	1,000	4
C2	3,524	212	24
C3	1,763	38	37

Note: 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).

4.1 The expansion of a cognitive core: the mapping of SARS

Severe Acute Respiratory Syndrome (SARS) is a respiratory disorder caused by the SARS corona virus (SARS-CoV). The first case of this disease was probably seen in the Guangdon province in Mainland China and there has been one major epidemic to date, between November 2002 and July 2003.

In the following we will mirror this outbreak by mapping bibliographic coupling associations in the network of the formal scientific communication reporting on this, and further, we will expand this delimited research theme in order to map its associative context. The narration has its point of departure in one of the generated C1-clusters, **C1/1391**. In this cluster, the focus is on SARS and all eleven constituent papers consistently treat this subject. The core documents in this cluster were published in ten different journals and assigned twelve different journal subject categories. The constituent articles are presented with article number, article title, journal title and journal subject category as follows:

1) 27178/ Chest-X-Ray Imaging of Patients with SARs / *Chinese Medical Journal* / **Medicine, General & Internal**
2) 28525/ Reovirus, Isolated from SARs Patients/ *Chinese Science Bulletin* / **Multidisciplinary Sciences**
3) 110241/ Infection-Control for SARs in a Tertiary Neonatal Center/ *Archives of Disease in Childhood*/ **Pediatrics**
4) 219617/ Description and Clinical Treatment of an Early Outbreak of Severe-Acute-Respiratory-Syndrome (SARs) in Guangzhou, Pr- China/ *Journal of Medical Microbiology*/ **Microbiology**
5) 275806/ A Clinicopathological Study of 3 Cases of Severe Acute Respiratory Syndrome (SARs)/ *Pathology*/ **Pathology**
6) 333109/ Severe Acute Respiratory Syndrome (SARs) - The Questions Raised by the Management of a Patient in Besancon and Strasbourg/ *Presse Medicale*/ **Medicine, General & Internal**

7) 383006/ Evaluation of WHO Criteria for Identifying Patients with Severe Acute Respiratory Syndrome Out-of-Hospital - Prospective Observational Study/ *British Medical Journal*/ **Medicine, General & Internal**
8) 400512/ Severe Acute Respiratory Syndrome-Associated Coronavirus Infection/ *Emerging Infectious Diseases*/ **Immunology; Infectious diseases**
9) 400574/ Microbiologic Characteristics, Serologic Responses, and Clinical-Manifestations in Severe Acute Respiratory Syndrome, Taiwan/ *Emerging Infectious Diseases*/ **Immunology; Infectious diseases**
10) 490401/ Safe Tracheostomy for Patients with Severe Acute Respiratory Syndrome/ *Laryngoscope*/ **Medicine, Research & Experimental; Otorhinolaryngology**
11) 527101/ Severe Acute Respiratory Syndrome in Hemodialysis-Patients-A Report of 2 Cases/ *Nephrology Dialysis Transplantation*/ **Transplantation; Urology & Nephrology**

It can be seen from the above list of articles that the problem of SARS is treated with a point of departure in several problem areas like *diagnosis* (both in vitro as well as in vivo), *pathology*, *clinical treatment* and specific *clinical problems* associated with this syndrome. It would also be interesting to know which cited works connect the core documents of this cluster. As the size of this cluster is 11, cited works with a frequency of 11 would be common to all articles in this cluster (Table 2).

TABLE 2: *The frequency of works cited by articles in C1/1391.*

Frequency	Cited Work
11	Drosten C, 2003, V348, P1967, New Engl J Med
11	Lee N, 2003, V348, P1986, New Engl J Med
9	Peiris JSM, 2003, V361, P1319, Lancet
9	Tsang KW, 2003, V348, P1977, New Engl J Med
8	Ksiazek TG, 2003, V348, P1953, New Engl J Med
7	Poutanen SM, 2003, V348, P1995, New Engl J Med
2	Ho W, 2003, V361, P1313, Lancet
2	Hon KLE, 2003, V361, P1701, Lancet
2	Li TST, 2003, V361, P1386, Lancet
2	Peiris JSM, 2003, V361, P1767, Lancet
2	WHO, 0000, CAS DEF Surv SEV AC
2	WHO, 0000, Cum Numb Rep Prob CA

Note: Only cited works with a frequency > 1 are shown in the table.

As can be noted, three of the cited works with the highest frequencies are all published in the same journal and in the same issue, *New England Journal of Medicine*, 2003, Volume 348, issue number 20. In addition, they are all about the outbreak of SARS in Hong-Kong in the year 2003:

Drosten C et al/ *Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome*

Lee N et al/ *A Major Outbreak of Severe Acute Respiratory Syndrome in Hong-Kong*

Tsang KW et al/ *A Cluster of Cases of Severe Acute Respiratory Syndrome in Hong-Kong*

This list of references is a clear indication of a research front topic where strongly connected source documents share references to currently published articles. The novelty value of the identified research theme on SARS is clear.

The next task is to study how this cluster possibly connect with other C1-clusters on the next level of cluster agglomeration. Here we find that C1/1391 is associated with four other clusters and the merging of these clusters leads to the construction of the C-2 cluster 170. This merging is depicted in Table 3.

TABLE 3: *The merging of C1-1391 with four C1-Clusters to C2/170.*

AvgCS(C, C')	C1-Cluster	Cluster Size	C1-Cluster	Cluster Size
5.42	1501	4	1513	3
5.14	1390	9	1501	4
5.00	1390	9	1513	3
4.27	1391	11	1501	4
4.03	1391	11	1513	3
4.00	1390	9	1391	11
3.60	1389	5	1501	4
3.56	1389	5	1390	9
3.47	1389	5	1513	3
3.11	1389	5	1391	11

This iterated clustering (clustering of clusters) resulted in a C2-cluster containing 32 documents. Hence, the research theme 'outbreak of SARS' has been considerably augmented. This raises the question of the homogeneity of the compound C2-cluster. One aspect of cluster

homogeneity is of course the requirement of a complete graph given by the method as such. Another would be the mean coupling strength of a cluster. This is easily computed as the ratio between the number of bibliographic couplings and the possible number pairs in a cluster, the Average Coupling Strength, AvgCS(C), for a cluster C. We define this average as:

$$AvgCS(C) = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} CS(d_i d_j)}{\binom{n}{2}}, \qquad \textbf{(4.1)}$$

where

n = number of articles in a cluster c,
CS = number of bibliographic coupling units between two articles, d_i, d_j and
$d_i d_j (\in C)$

Hence, the equation 2.1 as a measure of coherence, is complementary to equation 4.1. In this case we arrive at a value of 3.13 AvgCS(C). This may be compared with the mean AvgCS(C) of 4.25 for the constituent C1-clusters. Hence, the internal coherence, from a mere statistical point of view, has decreased. However, clear subject relatedness between constituent core documents and C1-clusters is seen when tabulating and sorting core document titles in accordance to cluster affiliation (Table 4).

TABLE 4: *Titles of articles in five C1-clusters merged to C2/170. The first column holds the numbers of constituent C1-clusters in C2/170.*

1389	Clinical Analysis of 45 Patients with Severe Acute Respiratory Syndrome
1389	The Role of Radiological Imaging in Diagnosis and Treatment of Severe Acute Respiratory Syndrome
1389	Initial Otolaryngological Manifestations of Severe Acute Respiratory Syndrome in Taiwan
1389	A Young Infant with Severe Acute Respiratory Syndrome
1389	Clinical Presentation and Outcome of Severe Acute Respiratory Syndrome in Dialysis Patients
1390	Zcurve-Cov - A New System to Recognize Protein-Coding Genes in Coronavirus Genomes, and Its Applications in Analyzing SARs-Cov Genomes
1390	Prediction of Proteinase Cleavage Sites in Polyproteins of Coronaviruses and Its Applications in Analyzing SARs-Cov Genomes
1390	Maintaining Dental Education and Specialist Dental-Care During an Outbreak of a New Coronavirus Infection - Part 1 - A Deadly Viral Epidemic

	Begins
1390	Role of China in the Quest to Define and Control Severe- Acute-Respiratory-Syndrome
1390	A Hospital Outbreak of Severe Acute Respiratory Syndrome in Guangzhou, China
1390	An Outbreak of Severe Acute Respiratory Syndrome Among Hospital Workers in a Community-Hospital in Hong-Kong
1390	Epidemiology and Cause of Severe Acute Respiratory Syndrome (SARs) in Guangdong, Peoples-Republic-of-China, in February, 2003
1390	Transmission Dynamics of the Etiologic Agent of SARs in Hong- Kong - Impact of Public-Health Interventions
1390	Children Hospitalized with Severe Acute Respiratory Syndrome- Related Illness in Toronto
1391	Severe Acute Respiratory Syndrome-Associated Coronavirus Infection
1391	Microbiologic Characteristics, Serologic Responses, and Clinical-Manifestations in Severe Acute Respiratory Syndrome, Taiwan
1391	Chest-X-Ray Imaging of Patients with SARs
1391	Severe Acute Respiratory Syndrome (SARs) - The Questions Raised by the Management of a Patient in Besancon and Strasbourg
1391	Evaluation of Who Criteria for Identifying Patients with Severe Acute Respiratory Syndrome Out-of-Hospital - Prospective Observational Study
1391	Safe Tracheostomy for Patients with Severe Acute Respiratory Syndrome
1391	Description and Clinical Treatment of an Early Outbreak of Severe-Acute-Respiratory-Syndrome (SARs) in Guangzhou, Pr- China
1391	Reovirus, Isolated from SARs Patients
1391	A Clinicopathological Study of 3 Cases of Severe Acute Respiratory Syndrome (SARs)
1391	Infection-Control for SARs in a Tertiary Neonatal Center
1391	Severe Acute Respiratory Syndrome in Haemodialysis-Patients - A Report of 2 Cases
1501	Clinical-Course and Management of SARs in Health-Care Workers in Toronto - A Case Series
1501	Outcomes and Prognostic-Factors in 267 Patients with Severe Acute Respiratory Syndrome in Hong-Kong
1501	Newly Discovered Coronavirus as the Primary Cause of Severe Acute Respiratory Syndrome
1501	Severe Acute Respiratory Syndrome in a Hemodialysis-Patient
1513	Severe Acute Respiratory-Distress-Syndrome (SARs) - A Critical-Care Perspective
1513	Enteric Involvement of Severe Acute Respiratory Syndrome- Associated Coronavirus Infection
1513	Investigation of a Nosocomial Outbreak of Severe Acute Respiratory Syndrome (SARs) in Toronto, Canada

In order to better appreciate the intra-cluster relations in C2/170 a MDS was performed (Figure 3).

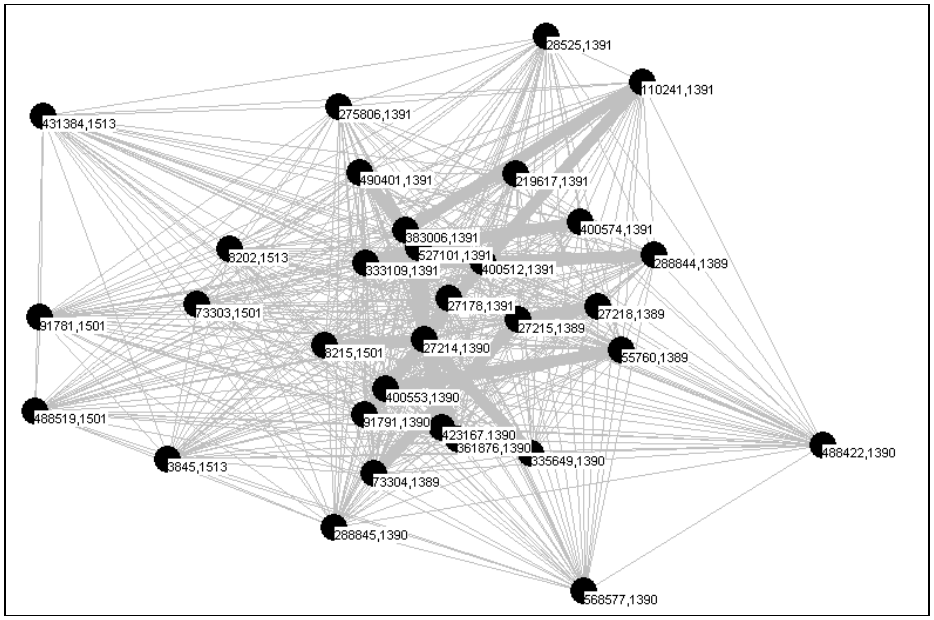

FIGURE 3: MDS of 32 core documents in C2/170. Commas separate document numbers from C1-cluster numbers. Width of links connecting data points, representing core documents, correspond to the linkage strength between the documents. Kruskal's stress 0.22.

Though Kruskal's stress was as high as 0.22, the map in Figure 3 still gives a comprehension of the internal structure of C2/170. We can see that documents basically cohere cluster wise, with the strongest associative structures indicated by the width of connecting links. However, the agreement on the map between documents and clusters is not perfect, indicating that significant associations exist between documents in different clusters, which is in line with the idea of iterative clustering.

So far, we have arrived at coherent and clearly interpretable cluster structures. Though the internal coherency diminishes somewhat when merging C1 clusters, we are still provided with pertinent information, as demonstrated. The question now arises if a final merging of clusters generates useful information. At the C2-level, however, it was found that the application of the complete link cluster method resulted in a partition with numerous singleton clusters and a few clusters containing only two documents. In praxis, this means that an upper limit of application for the

method is reached. Still, the question if the remaining links between clusters generated on the C2-level may give rise to an interpretable cluster solution on a higher level needs to be answered. In order to be able to map links, the between groups average cluster method had to be applied. For this method, the similarity between two clusters is the average of the similarity between all pairs of individuals that are made up by one individual from each group. Hence, clustering on basis of averages implies that we have to abandon the appealing concept of complete graphs.

On the final level of cluster fusion, cluster C2/170 is merged with two other C2-clusters to C3/3, with a total number of 61 core documents. The by now decreased average coupling strength between clusters should be noted (Table 5; cf.Table 3).

TABLE 5: *The fusion of three C2-Clusters to C3/3.*

AvgCS(C, C')	*No. Shared References*	*C2-Cluster*	*Cluster Size*	*C2-Cluster*	*Cluster Size*
0.25	129	87	16	170	32
0.10	20	87	16	171	13
2.48	1030	170	32	171	13

As can be concluded from Table 5 above, cluster C2/87 is the most distant (dissimilar) vertice in a three-edged graph (Figure 4).

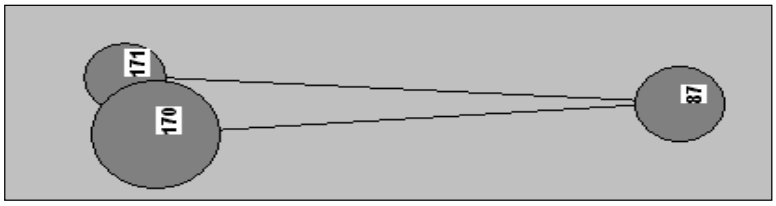

FIGURE 4: MDS of three C2 clusters on basis of the AvgCS(*C*, *C´*).

From this we can assume that C2/87, from some aspect of similarity, should deviate from a research topic common to other two C2-clusters. This is, however, not clearly reflected by the mix of journal subject categories assigned to the core documents in the C2-clusters:

C2/ 87: infectious diseases, clinical microbiology

C2/ 170: general & internal medicine, infectious diseases, biochemistry ; pediatrics; urology & nephrology

C2/ 171: biochemistry & molecular biology, clinical chemistry, microbiology, virology

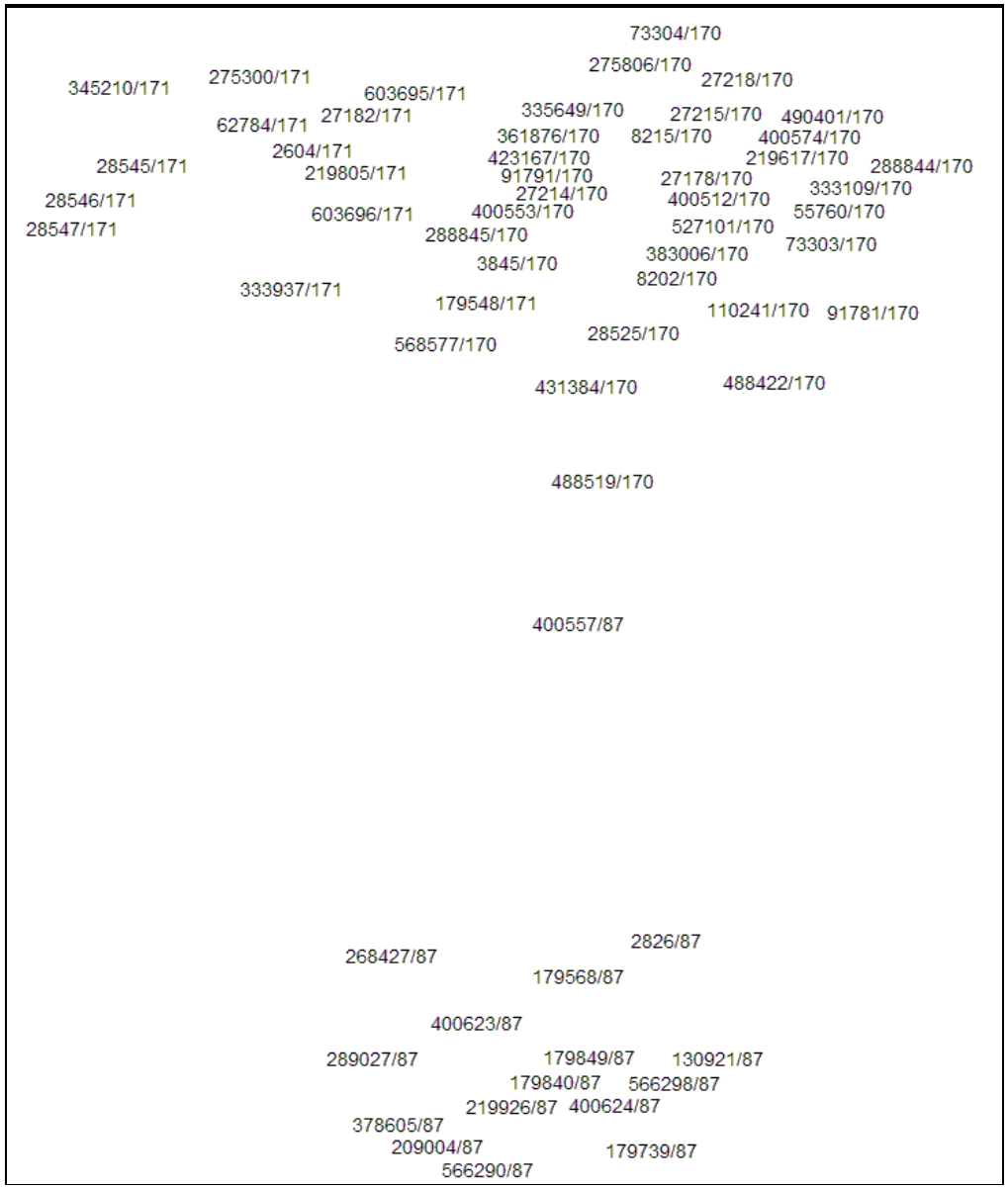

FIGURE 5: MDS of C3/3 constituted by C2/87, C2/ 170 & C2/ 171. Numbers on the map correspond to documents and C2-clusters. Kruskal's stress is 0.11

Approximately 11 different disciplines are more or less associated with the research theme(s) of C3/3 (2) and this could be considered a very raw approximation of the homogeneity of the C3-cluster. From a statistical point of view, the internal coherency of the generated cluster on the C3 level is weakened, with a value of 2.30 for the AvgCS(*C*) and for *D* 0.62. At this level of cluster fusion, the structure is complex and requires a thorough analysis. Applying MDS for this purpose, we may apply a "top-bottom" interpretation (Figure 5).

Beginning this analysis from the "top", all links between core documents in C3/3 are displayed in a two dimensional plane by MDS with an acceptable value of stress. In this graph, the constituent C2 clusters are clearly discernable. Clusters C2/171 and C2/170 are configured in the upper part of the map and cluster C2/87 in the lower part with the core document 400557 in an intermediate position.

Applying MDS to display the associations between significant terms (title words) according to their co-occurrence in titles, an overview sketch of the topic content in cluster C3/3 is arrived at (Figure 6).

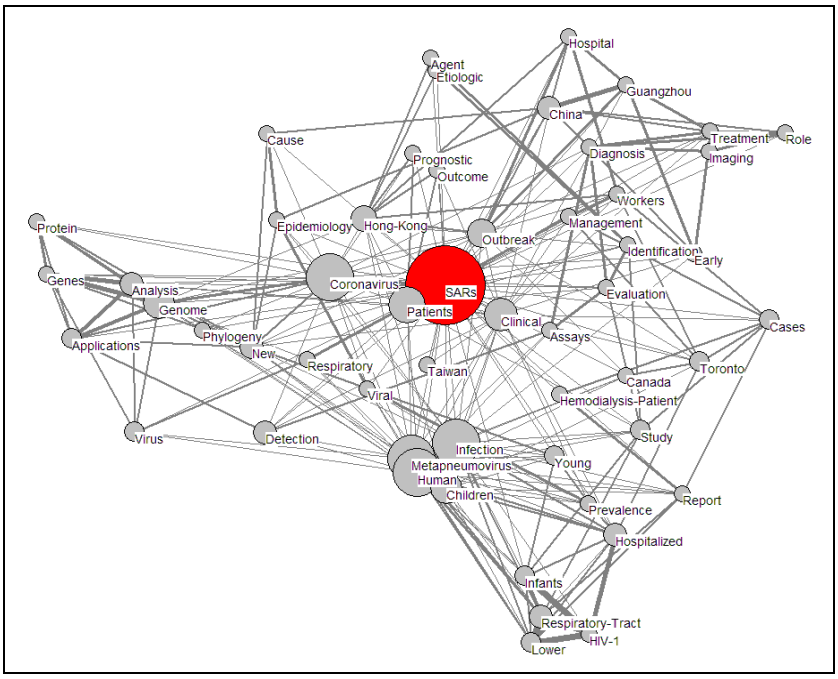

FIGURE 6: MDS Display of the co-occurrence of title words from core documents of C3/3. The lowest frequency of term occurrence allowed for was 2. The width of links is corresponding to the strength of similarity between terms as measured by the Jaccard index. Circle sizes correspond to the frequency of occurrence. Kruskal's stress is 0.10.

As can be seen from Figure 6, the focus of core documents in cluster C3/3 is on infectious diseases caused by viruses (in particular SARS). To begin with, the configuration in terms of term size-position could be examined. The term "SARs" has a central position and, as indicated by the circle size, it is the most frequent term. Radiating out from "SARs", different dimensions associated with diseases caused by viruses can be discerned:

- a time-geography dimension (outbreak; early; epidemiology; Hong-Kong; China; Guangzhou; Taiwan; Canada; Toronto);
- a clinical dimension (metapneumovirus; human; infection; children; young; infants; lower; respiratory tract; HIV 1; hospitalized; prevalence); and
- a dimension of the genetics of viruses (analysis; genome; phylogeny; new; protein; genes; application; virus).

Zooming in on particular C2-clusters, different aspects of cluster C3/3 could be reflected. Beginning with cluster C2/87 (located on the lower part of the map in Figure 5), this cluster is constituted by two C1 clusters and the total number of core documents was 16. Studying the titles of core documents constituting C2/87, the subject homogeneity is obvious as they are all on *Human Metapneumovirus*. Also, a preliminary explanation of the intermediate role of core document 400557 (cf. Figure 5) is that it associates this virus with SARS. SARS is for all associated with a corona virus (SARS-CoV), but also with the human metapneumo virus, though to a lesser degree (cf.Table 6).

TABLE 6: *Core document titles in C2/87.*

2826 / Human Metapneumovirus-Associated Lower Respiratory-Tract Infections Among Hospitalized Human-Immunodeficiency-Virus Type-1 (HIV-1)-Infected and HIV-1-Uninfected African Infants
130921 / Prevalence and Clinical Symptoms of Human Metapneumovirus Infection in Hospitalized-Patients
179568 / Human Metapneumovirus Infection in the Canadian Population
179739 / Comparative-Evaluation of Real-Time PCR Assays for Detection of the Human Metapneumovirus
179840 / Human Metapneumovirus Associated with Respiratory-Tract Infections in a 3-Year Study of Nasal Swabs from Infants in Italy
179849 / High Prevalence of Human Metapneumovirus Infection in Young- Children and Genetic-Heterogeneity of the Viral Isolates
209004 / Human Metapneumovirus Infections in Young and Elderly Adults

219926 / Seroprevalence of Human Metapneumovirus in Japan

268427 / Effects of Human Metapneumovirus and Respiratory Syncytial Virus-Antigen Insertion in 2 3'-Proximal Genome Positions of Bovine /Human Parainfluenza Virus Type-3 on Virus-Replication and Immunogenicity

289027 / Human Metapneumovirus Infection in the United-States - Clinical-Manifestations Associated with a Newly Emerging Respiratory-Infection in Children

378605 / Human Metapneumovirus in a Hematopoietic Stem-Cell Transplant Recipient with Fatal Lower Respiratory-Tract Disease

400557 / Human Metapneumovirus Detection in Patients with Severe Acute Respiratory Syndrome

400623 / Children with Respiratory-Disease Associated with Metapneumovirus in Hong-Kong

400624 / Human Metapneumovirus Infections in Hospitalized Children 219926 / Seroprevalence of Human Metapneumovirus in Japan

566290 / Human Metapneumovirus Infection in Thai Children

Clinical findings and studies of prevalence with regard to human metapneumo virus (with some emphasis on children) are presented by the core document titles. Looking at titles and journal subject category assignments of journals in which these core documents were published, the disciplinary structure leans towards general (internal) medicine (infectious diseases) but there is also a contribution from basic medical sciences (immunology, microbiology) (cf. Table 7).

TABLE 7: *Journal titles and assigned journal subject categories corresponding to core documents in C2/87. Numbers in brackets correspond to the frequency of articles published in a journal*

Journal Title	Journal Subject Categories
(1) Clinical Infectious Diseases	Immunology; infectious diseases; microbiology
(3) Emerging Infectious Diseases	Immunology; infectious diseases
(2) Journal of Infectious Diseases	Infectious diseases
(2) Scandinavian Journal of Infectious Diseases	Infectious diseases
(4) Journal of Clinical Microbiology	Microbiology
(1) Journal of Medical Virology	Virology
(1) Journal of Virology	Virology
(1) Pediatrics	Paediatrics
(1) Bone Marrow Transplantation	Oncology; hematology; immunology; transplantation

The next C2-cluster to be studied is C2/171 (located on upper left quadrant in Figure 5) which is formed by three C1 clusters and 13 core documents. Table 8 gives the core document titles in this cluster.

TABLE 8: *Core document titles in C2/171.*

2604 / Quantitative-Analysis and Prognostic Implication of SARs Coronavirus RNA in the Plasma and Serum of Patients with Severe Acute Respiratory Syndrome
179548 / Evaluation of Reverse Transcription-PCR Assays for Rapid Diagnosis of Severe Acute Respiratory Syndrome-Associated with a Novel Coronavirus
219805 / Early Events of SARs Coronavirus Infection in Vero Cells
27182 / Establishment of a Fluorescent Polymerase-Chain-Reaction Method for the Detection of the SARs-Associated Coronavirus and Its Clinical-Application
28545 / Design and Application of 60Mer Oligonucleotide Microarray in SARs Coronavirus Detection
28546 / Molecular Phylogeny of Coronaviruses Including Human SARs-Cov
28547 / Phylogeny of SARs-Cov as Inferred from Complete Genome Comparison
333937 / Activation of Ap-1 Signal-Transduction Pathway by SARs Coronavirus Nucleocapsid Protein
345210 / Genomic Characterization of the Severe-Acute-Respiratory- Syndrome Coronavirus of Amoy Gardens Outbreak in Hong-Kong
603695 / Coronavirus in Severe Acute Respiratory Syndrome (SARs)
603696 / Severe Acute Respiratory Syndrome - Identification of the Etiologic Agent
62784 / Mutation Analysis of 20 SARs Virus Genome Sequences - Evidence for Negative Selection in Replicase Orf1B and Spike Gene
275300 / The Crystal-Structures of Severe Acute Respiratory Syndrome Virus Main Protease and Its Complex with an Inhibitor

As can be seen from Table 8, this cluster focuses exclusively on coronavirus and SARS. The emphasis is on the analysis of the genetical characterisation and on methods for the detection and description of the viruses (e.g. laboratory methods, isolation and cultivation). The clinical focus seen in cluster C2/87 is thus replaced with more basic research on the virus causing SARS. This is also reflected by the composition of the set of

publishing journals and journal subject categories of this cluster, where the contribution from chemistry and biochemistry is salient (cf. Table 9).

TABLE 9: *Journal titles and assigned journal subject categories corresponding to core documents in cluster C2/171. Numbers in brackets correspond to the frequency of papers published in a journal.*

Journal Title	Journal Subject Categories
(3) Chinese Science Bulletin	Multidisciplinary sciences
(2) Trends in Molecular Medicine	Biochemistry & molecular biology; cell biology; medicine, research & experimental
(1) ACTA Pharmacologica Sinica	Chemistry,multidisciplinary; pharmacology & pharmacy
(1) Biochemical and Biophysical Research Communications	Biochemistry & molecular biology; biophysics
(1) Chinese Medical Journal	Medicine, general & internal
(1) Clinical Chemistry	Medical laboratory technology
(1) Journal of Clinical Microbiology	Microbiology
(1) Journal of Medical Virology	Virology
(1) Lancet	Medicine, general & internal
(1) Proceedings of the National Academy of Sciences of the United States of America	Multidisciplinary sciences

Finally, we have already, to some extent, examined the largest C2 cluster (C2/170) in C3/3 (located in the upper right quadrant in Figure 5). This cluster was formed by three C1 clusters and 32 core documents. In this cluster, several case studies as well as clinical aspects on diagnosis and prevention are reported, and the overall focus is again on clinical aspects of SARS (cf.Tables 3 and 4). This is in line with the distribution of journal subject categories and journal titles, reflecting that several medical disciplines and sub-disciplines, with an emphasis on general medicine, are involved (cf. Table 10).

TABLE 10: *Journal titles and assigned journal subject categories corresponding to core documents in cluster C2/170. Numbers in brackets correspond to the frequency of papers published in a journal.*

Journal Title	Journal Subject Categories
(5) Chinese Medical Journal	Medicine, general & internal
(3) Emerging Infectious Diseases	Immunology; infectious diseases
(2) American Journal of Kidney Diseases	Urology & nephrology
(2) Annals of Internal Medicine	Medicine, general & internal
(2) Canadian Medical Association Journal	Medicine, general & internal
(2) Lancet	Medicine, general & internal
(2) Paediatrics	Paediatrics
(1) Archives of Disease in Childhood	Paediatrics
(1) Archives of Otolaryngology-Head & Neck Surgery	Otorhinolaryngology; surgery
(1) British Dental Journal	Dentistry, oral surgery & medicine
(1) British Medical Journal	Medicine, general & internal
(1) Chinese Science Bulletin	Multidisciplinary sciences
(1) Critical Care Medicine	Critical care medicine
(1) FEBS Letters	Biochemistry & molecular biology; biophysics; cell biology
(1) Gastroenterology	Gastroenterology & hepatology
(1) Journal of Medical Microbiology	Microbiology
(1) Laryngoscope	Medicine, research & experimental; otorhinolaryngology
(1) Nephrology Dialysis Transplantation	Transplantation; urology & nephrology
(1) Pathology	Pathology
(1) Presse Medicale	Medicine, general & internal
(1) Science	Multidisciplinary sciences

It is to be noted that the different MDS maps (Figures 5 and 6) match surprisingly well, though the first map groups papers according to shared references and the second map title words according to their co-occurrence frequency in titles in core documents. Hence, C2/87 seems to correspond to the lower part of the "title word" map, C2/171 to the left-middle-upper part and C2/170 to the right-middle-upper part.

It seems clear that the C2-clusters are subjected consistent and a common denominator (SARS) can be identified. The interdisciplinary character on all levels (C1 - C3) is obvious. The merging of the three C2-clusters on the C3 level thus connects research on two different viruses and

the pathology of diseases caused by them, genesis of agents and corresponding clinical research and observations. The merging of C2/87 with the other two clusters could, however, be questioned. Though a least common denominator (SARS) exists, both the measured distances and the assessed cognitive distances between C2-clusters showed that there is a need for the separation of cluster C2/87 from the other C2-clusters. Nevertheless, the associations of the clusters could be of interest as they all gather around a common problem, though from different perspectives.

5. Discussion

As one of many plausible modes of science mapping, the practicability of bibliographic coupling in combination with the complete link cluster method for the mapping of core documents in the research front has been demonstrated. From a statistical point of view, the clusters on three levels of agglomeration were all below the mean cluster coherence. Hence, the demonstration of this method applied on the primary research theme "outbreak of SARS" has not applied data clustered towards the upper tail of the distribution of internal average coupling strengths (AvgCS(*C*)). There is therefore cause to hypothesise that in general, with regard to cluster homogeneity, clusters with at least the same quality would show up for other research themes, given the approximately same disciplinary settings.

On basis of the results from this example of core document mapping, one is inclined to delimit the combination of bibliographic coupling of core documents with the complete link cluster method to the second level of agglomeration as the final cluster on the C-3 level 'sprawled' in terms of subject coherence. Hence, the first example of clustering (C1-level) showed up with what one may label as a 'current cognitive core' reporting novel medical news and there was little doubt about the usefulness of the information contained. As reflected by MDS, informative links between documents in different clusters remained between clusters on the C1-level, motivating the subsequent iteration of clustering. It was demonstrated that the C2-cluster chosen for analysis was subject coherent, and, from a statistical point of view, coherent in terms of the degree of interconnectedness. Hence, one may from one perspective assume that the upper limit for iterative clustering, applying the complete link method as described, is the C2-level. However, from a more pragmatic perspective, in a 'real-life situation', it is hard to imagine that any professional provider of scientific information would not continue to crunch data until no pertinent information could be extracted. Thus, as demonstrated, interesting and

possibly useful links between the C2-clusters may be obtained, connecting diverse research foci through some common aspect.

Conclusively, it was shown that pertinent information may be extracted and mapped using the applied method combining bibliographic coupling with the complete link method. Such a task is seldom easy to pursue as in most mapping enterprises, the most apparent difficulty is the filtering out of insignificant information (noise), and the lack of distinctly discernable thresholds, as well as a theoretical foundation for such. Thus, the empirical findings reported by Glänzel and Czerwon (1995, 1996) concerning core documents should be regarded an important progress and hopefully more future applications of core-document mapping will be presented to the research community.

Notes

(1) The term 'research front' has several interpretations, some on the basis of the cocitation analytical model, but here we regard 'research front(s)' as a metaphorical expression for the more current published research in a field.

(2) The exact number of disciplines will not be given by assigned journal subject categories as these are journal classifications, covering the scope of journals, not the scope of the individual paper.

References

GLÄNZEL, W. & CZERWON, H. J. 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*. River Forest, Illinois, June 7-10, 1995, p. 167-176.

GLÄNZEL, W. & CZERWON, H. J. A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. In: *Scientometrics*, 1996, 37(2), 1996, p. 195-221.

JARNEVING, B. The cognitive structure of current cardiovascular research. In: *Scientometrics*, 50 (3), 2001, p. 365-389.

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

KESSLER, M.M. Bibliographic coupling between scientific papers. In: *American Documentation*, 14(1), 1963a, p. 10-25.

KESSLER, M.M. Bibliographic coupling extended in time: Ten case histories. In: *Information Storage and Retrieval*, 1, 1963b, p. 169-187.

KESSLER, M.M. Comparison of the results of bibliographic coupling and analytic subject indexing. In: *American Documentation*, 16(3), 1965, p. 223-233.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. In: *Psychometrica*, 29 (1), 1964, p. 1-27.

KRUSKAL, J. B.; WISH, M. B. Multidimensional scaling. In: *Quantitative Applications in the Social Sciences*, vol. 11. London: Sage Publications, inc, 1978.

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

PETERS, H. P. F.; BRAAM, R. R. şi van Raan, A. F. J. Cognitive resemblance and citation relations in chemical engineering publications. In: *Journal of the American Society for Information Science*. 46(1), 1995, p. 9-21.

SEN, S. K. & GAN. S. K. A mathematical extension of the idea of bibliographic coupling and its applications. In: *Annals of Library Science and Documentation*, 30(2) 1983, p. 78-82.

SHARABCHIEV, Y.T. Comparative analysis of 2 methods of cluster analysis of bibliographic citation. (In Russian). In: *Naucno-techniceskaja informacija* / 2, 1998.

VLADUTZ, G. & COOK, J. Bibliographic coupling and subject relatedness. In: *Proceedings of the ASIS Annual Meeting*, 47, 1994, p. 204-207.