Information Processing and Management 45 (2009) 683–702

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Hybrid clustering for validation and improvement of subject-classification schemes

Frizo Janssens [a,b,c,*], Lin Zhang [a,d], Bart De Moor [c], Wolfgang Glänzel [a,e]

[a] K.U. Leuven, Centre for R&D Monitoring (ECOOM), Dept. MSI, Leuven, Belgium
[b] Attentio SA/NV, StudioTROPE Building, Bloemenstraat 32, B-1000 Brussels, Belgium
[c] K.U. Leuven, ESAT-SCD, Leuven, Belgium
[d] WISE Lab, Dalian University of Technology, Dalian, China
[e] Hungarian Academy of Sciences, IRPS, Budapest, Hungary

ARTICLE INFO

Article history:
Received 27 August 2008
Received in revised form 8 June 2009
Accepted 14 June 2009
Available online 21 July 2009

Keywords:
Subject classification
Journal cross-citation
Mapping of science
Hybrid clustering

ABSTRACT

A hybrid text/citation-based method is used to cluster journals covered by the Web of Science database in the period 2002–2006. The objective is to use this clustering to validate and, if possible, to improve existing journal-based subject-classification schemes. Cross-citation links are determined on an item-by-paper procedure for individual papers assigned to the corresponding journal. Text mining for the textual component is based on the same principle; textual characteristics of individual papers are attributed to the journals in which they have been published. In a first step, the 22-field subject-classification scheme of the Essential Science Indicators (ESI) is evaluated and visualised. In a second step, the hybrid clustering method is applied to classify the about 8300 journals meeting the selection criteria concerning continuity, size and impact. The hybrid method proves superior to its two components when applied separately. The choice of 22 clusters also allows a direct field-to-cluster comparison, and we substantiate that the science areas resulting from cluster analysis form a more coherent structure than the "intellectual" reference scheme, the ESI subject scheme. Moreover, the textual component of the hybrid method allows labelling the clusters using cognitive characteristics, while the citation component allows visualising the cross-citation graph and determining representative journals suggested by the PageRank algorithm. Finally, the analysis of journal 'migration' allows the improvement of existing classification schemes on the basis of the concordance between fields and clusters.

1. Introduction

The history of cognitive mapping of science is as long as the history of computerised scientometrics itself. While the first visualisations of the structure of science were considered part of information services, i.e., an extension of scientific review literature (Garfield, 1975, 1988), bibliometricians soon recognised the potential value of structural science studies for science policy and research evaluation as well. At present, the identification of emerging and converging fields and the improvement of subject delineation are in the foreground.

The main bibliometric techniques are characterised by three major approaches, particularly the analysis of citation links (cross-citations, bibliographic coupling, co-citations), the lexical approach (text mining), and their combination. The widely used method of co-citation clustering was introduced independently by Small (1973, 1978) and Marshakova (1973).

* Corresponding author. Address: K.U. Leuven, Centre for R&D Monitoring (ECOOM), Dept. MSI, Leuven, Belgium.
E-mail address: frizo.janssens@esat.kuleuven.be (F. Janssens).

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.06.003

Although the principle of bibliographic coupling had already been discovered earlier by Fano (1956) and Kessler (1963), coupling-based techniques have been used for mapping the structure of science only decades after co-citation analysis had become a standard tool in visualising the structure of science (e.g., Glänzel & Czerwon, 1996; Small, 1998). Cross-citation-based cluster analysis for science mapping has to be distinguished from the previous two methods; while the former two types can be – and usually are – based on links connecting individual documents, the latter approach requires aggregation of documents to units like journals, subject categories, etc., among which cross-citation links are established. The obvious advantages of this method (e.g., the possibility to analyse directed information flows among these units or the assignment/aggregation of units to larger structures) are contrasted by some limitations and shortcomings such as possible biases caused by the use of predefined units. Thus, for instance, Leydesdorff (2006), Leydesdorff and Rafols (2009), and Boyack, Börner, and Klavans (2009) used journal cross-citation matrices, while Moya-Anegon et al. (2007) used subject co-citation analysis to visualise the structure of science and its dynamics.

Earlier, a completely different approach was introduced by Callon, Courtial, Turner, and Bauin (1983) and Callon, Law, and Rip (1986). Their mapping and visualisation tool Leximappe was based on a lexical approach, particularly, co-word analysis. The notion of lexical approach, which was originally based on extracting keywords from records in indexing databases, was later on deepened and extended by using advanced text-mining techniques in full texts (cf. Glenisson, Glänzel, Janssens, & De Moor, 2005; Glenisson, Glänzel, & Persson, 2005; Kostoff, Buchtel, Andrews, & Pfeil, 2005; Kostoff, Toothman, Eberhart, & Humenik, 2001).

Whatever method is used to study the structure of science, cluster algorithms have beyond doubt become the most popular technique in science mapping. The sudden, large interest the application of these techniques has found in the community is contrasted by objections and criticism from the viewpoint of information use in the framework of research evaluation (e.g., Noyons, 1999; Jarneving, 2005). For instance, clustering based on co-citation and bibliographic coupling has to cope with several severe methodological problems. This has been reported, among others by Hicks (1987) in the context of co-citation analysis and by Janssens, Glänzel, and De Moor (2008) with regard to bibliographic coupling. One promising solution is to combine these techniques with other methods such as text mining (e.g., combined co-citation and word analysis: Braam, Moed, & Van Raan, 1991a; combination of coupling and co-word analysis: Small (1998); hybrid coupling-lexical approach: Janssens, Glänzel, & De Moor, 2007; Janssens et al., 2008). Most applications were designed to map and visualise the cognitive structure of science and its change in time, and, from a policy-relevant perspective, to detect new, emerging disciplines. Improvement of subject-classification schemes was in most cases not intended. Jarneving (2005) proposed a combination of bibliometric structure–analytical techniques with statistical methods to generate and visualise subject coherent and meaningful clusters. His conclusions drawn from the comparison with 'intellectual' classification were rather sceptical. Despite several limitations, which will be discussed further in the course of the present study, cognitive maps proved useful tools in visualising the structure of science and can be used to adjust existing subject-classification schemes even on the large scale as we will demonstrate in the following.

The main objective of this study is to compare (hybrid) cluster techniques for cognitive mapping with traditional 'intellectual' subject-classifications schemes. The most popular subject-classification schemes created by Thomson Scientific (Philadelphia, PA, USA) are based on journal assignment. Therefore, journal cross-citation analysis puts itself forward as underlying method and we will cluster the document space using journals as predefined units of aggregation. In contrast to the method applied by Leydesdorff (2006), who uses the Journal Citation Reports (JCR), we calculate citations on a paper-by-paper basis and then assign individual papers indexed in the *Web of Science* (WoS) database to the journals in which they have been published. The use of the JCR would confine us to data as available in the JCR and prevent us from combining cross-citation analysis with a textual approach. What is more, proceeding from the document level allows us to control for document types and citation windows, and to combine bibliometrics-based techniques with other methods such as text mining. This results in a higher precision since irrelevant document types and 'low-weight journals' can be excluded. This way we can present the results of a hybrid (i.e., combined/integrated) citation–textual cluster analysis to compare those with the structure of an existing 'intellectual' subject-classification scheme created and used by Thomson Scientific. The aim of this comparison is exploring the possibility of using the results of the cluster analysis to improve the subject-classification scheme in question.

1.1. Related research

The idea of hybrid clustering is not new. For example, Weiss et al. (1996) introduced the HyPursuit system, a hierarchical network search engine which exploits content-link clustering of hypertext documents for browsing and search activities. The complete link hierarchical agglomerative clustering method was used and the similarity between documents was based on the number of terms, ancestors, and descendants in common, as well as whether they point to one another. In the presented prototype, the integrated document similarity function was limited to selecting the maximal value of the text-based and hyperlink similarities, which already offered benefits to the hypertext document clustering.

Modha and Spangler (2000) introduced the toric k-means algorithm for clustering hypertext documents using words, outlinks and in-links. The relative importance of these information sources was determined by searching the parameter space for an optimal figure-of-merit. Similarity was calculated as a weighted sum of the inner products between the individual text-based or link-based components. In the present work we also adopt a sum of document similarities, but combine the method with a hierarchical clustering algorithm instead of k-means.

He, Ding, Zha, and Simon (2001), He, Zha, Ding, and Simon (2002) discussed Web document clustering by incorporating information from hyperlink structure, co-citation patterns, and textual contents of documents. The hyperlink structure was used as the dominant factor in the combined similarity measure, and the textual content was used to modulate the strength of each hyperlink. However, textual similarity between pages was neglected if both were not connected by a hyperlink. Integration with co-citation was achieved by a linear combination of co-citation and the weighted adjacency matrix of the graph. The resulting weighted graph was the input to a spectral clustering method.

Wang and Kitsuregawa (2002) evaluated a contents-link coupled clustering algorithm for retrieved Web pages and studied the effect of out-links, in-links, specific terms, and their combination. Results suggested that both links and contents are important for Web page clustering and that better results are achieved with appropriate integration weights.

By using a Bayesian network model, Calado et al. (2006) combined link-based similarity measures with text-based classifiers to improve classification results for Web collections. In their experiments on Web pages, the link information alone outperformed the text-only classifier, but the combination could improve results.

Finally, Zhang et al. (2005) applied genetic programming techniques to discover the best fusion framework to integrate citation-based information and structural content in order to improve document classification.

Except for the latter study, all mentioned related research was applied to hyperlink documents, while the present study is applied to scientific literature.

In a first study by authors related to the current work, the pilot study of Glenisson, Glänzel, & Persson (2005), further extended and confirmed by Glenisson, Glänzel, Janssens et al. (2005), full-text analysis and traditional bibliometric methods were serially combined to improve the efficiency of the individual methods. It was clear that clusters found through application of text mining provided additional information that could be used to extend and explain structures found by bibliometric methods, and *vice versa*. However, the integration was still limited to serial combination. Real hybrid methodologies were therefore developed in subsequent studies and proved even more valuable tools to facilitate endeavours in mapping fields of science or technology.

In a previous study of Janssens, Leta, Glänzel, and De Moor (2006), the concept structure of library and information science (LIS) was obtained by full-text mining of almost 1000 articles and notes published in the period 2002–2004 in 5 representative journals with strong focus on information science. Only the 'pure' text corpus was analysed, excluding any bibliographic or bibliometric components. However, the authors have assessed the performance of clustering and classification algorithms using various data integration schemes on multiple data sets (Janssens, 2007). The best outcome was obtained by hybrid methods that exploited both text and citations. An integration method based on Fisher's inverse chi-square and another one based on linear combination of distance matrices were among the best methods and significantly outperformed corresponding text-only and link-only methods, as well as a concatenation of vectors. In (Janssens et al., 2008), we used these methods to achieve an updated, hybrid mapping of information science by using the full-text information as well as citations, and we compared the results with the text-only clustering. The same techniques from text mining, Web mining and bibliometrics were also applied to a data set containing bioinformatics research articles (Janssens et al., 2007).

While the analyses of Glenisson, Glänzel, & Persson (2005), Glenisson, Glänzel, Janssens et al. (2005) and Janssens, Leta et al. (2006), Janssens et al. (2008) were based on the full text extracted from scientific articles, the textual component of the current study is based on titles, abstracts and keywords. The present analysis also differs from the previous studies in that it is situated on the journal instead of the paper level, i.e., here we cluster journals, not papers. The underlying paper set is also at least two orders of magnitude larger (over 6 million papers). Furthermore, the link component of the integration used here is based on journal cross-citations, whereas bibliographic coupling was used in other studies (Janssens et al., 2008; Janssens et al., 2007). The actual integration is achieved by linear combination of the corresponding distance matrices, whereas (Janssens et al., 2008) and (Janssens et al., 2007) used the method based on Fisher's inverse chi-square.

In the present study, we decided to somewhat modify our earlier approach by the following reason. Bibliographic coupling used in previous studies has certainly advantages in detecting cognitive links especially as compared to co-citation analysis (cf. Glänzel & Czerwon, 1996; Hicks, 1987), however, its bias towards a polarisation between strong and weak links which is already reflected by the extremely sparse coupling matrices (Janssens, Tran Quoc, Glänzel, & De Moor, 2006) suggests to use a different approach. Thus the cross-citation approach promises to combine the advantages of references and citation or – using the parlance of Webometrics – the advantages of in-links and out-links. The linear combination of cross-citations and text mining allows for smoothly adjusting the weight of both components.

1.2. Cognitive mapping vs. subject classification

The objective of the present study is twofold. The first task is not merely visualising the field structure of science by presenting yet another map based on an alternative approach, but to validate and improve existing subject classifications used for research evaluation. In particular, the question arises of in how far observed 'migration' of journals among science fields can be adopted to improve classification. The second issue is a methodological one; namely, to evaluate improved methods of hybrid clustering techniques.

The 22-field subject-classification scheme of the Essential Science Indicators[1] (ESI) of Thomson Scientific, which actually forms a partition of the Web of Science universe with practically unique subject assignment, is used as the "control structure".

[1] http://www.esi-topics.com/fields/index.html.

In particular, we propose the following approach in seven steps to solve the integration of cluster analysis and cognitive mapping into subject classification.

1. Evaluation of existing subject-classification schemes and visualisation of their cross-citation graph.
2. Labelling subject fields using cognitive characteristics.
3. Studying the cognitive structure based on hybrid cluster analysis and visualisation of the cross-citation graph.
4. Evaluation of science areas resulting from cluster analysis.
5. Labelling clusters using cognitive characteristics and representative journals suggested by the PageRank algorithm.
6. Comparison of subject fields and cluster structure.
7. Migration of journals among subject fields.

2. Data sources and data processing

In order to accomplish the above objectives, more than six million papers of the type article, letter, note and review indexed in the Web of Science (WoS) in the period 2002–2006 have been taken into consideration. Citations to individual papers have been aggregated from the publication year till 2006. The complete database has been indexed and all terms extracted from titles, abstracts and keywords have been used for ‘‘labelling” the obtained clusters.

Citations received by these papers have been determined for a variable citation window beginning with the publication year, up to 2006, on the basis of an item-by-item procedure using special identification-keys made up of bibliographic data elements extracted from first-author names, journal title, publication year, volume and first page.

In a first step, journals had to be checked for name changes, merging or splitting and identified accordingly. Journals which were not covered in the entire period have been omitted. Furthermore, only journals that have published at least 50 papers in the period under study were considered. A second threshold was used afterwards to remove all journals for which the sum of references and citations was lower than 30. The resulting number of retained journals was 8305. Most of the subsequent analyses were performed in Java and MATLAB. We also made use of the MATLAB Tensor Toolbox (Bader & Kolda, 2006).

3. Methods

In this section we briefly describe the methodological background and the algorithms and procedures that have been applied. The first subsection refers to the outlines of the textual approach; this is followed by the description of the cross-citation analysis. The journal clustering techniques described in the subsequent paragraphs are applied to the textual and citation data separately and used for combined (hybrid) clustering as well. This procedure is described in the following step by step.

3.1. Text analysis

All textual content was indexed with the Jakarta Lucene platform (Hatcher & Gospodnetic, 2004) and encoded in the Vector Space Model using the TF-IDF weighting scheme reviewed by Baeza-Yates & Ribeiro-Neto (1999). Stop words were neglected during indexing and the Porter stemmer was applied to all remaining terms from titles, abstracts, and keyword fields. The resulting term-by-document matrix contained nine and a half million term dimensions (9,473,061), but by ignoring all tokens that occurred in one sole document, only 669,860 term dimensions were retained. Those ignored terms with document frequency equal to one are useless for clustering purposes. The dimensionality was further reduced from 669,860 term dimensions to 200 factors by Latent Semantic Indexing (LSI) (Berry, Dumais, & O'brien, 1995; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), which is based on the Singular Value Decomposition (SVD). The reduction of the number of features in a vector space by application of LSI improves the performance of retrieval, clustering, and classification algorithms. Typical values for the number of factors found in literature range from 100 (Deerwester et al., 1990) (for less than 2000 abstracts and 7000 terms) to about 300 (Berry, Dumais, & O'brien, 1995) (for about 12,000 documents and 40,000 terms). On a dual Opteron 250 with 16 Gb RAM, the processor time needed to calculate the LSI was about 105 minutes. Text-based similarities were calculated as the cosine of the angle between the vector representations of two papers (Salton & Mcgill, 1986). For simplicity and efficiency, the method used to summarise the subject of a field or cluster is based on selecting the terms with the highest mean TF-IDF weights over all journal papers in the field or cluster, where the IDF factor is calculated on the complete term-by-paper matrix (more than six million papers). Other, more advanced term selection methods exist. For example, Treeratpituk and Callan (2006) automatically select and assign a few concise labels to hierarchical clusters by combining statistical features from the cluster, parent cluster, and a corpus of general English into a descriptive score. Geraci, Maggini, Pellegrini, and Sebastiani (2008) label clusters by combining intra-cluster and inter-cluster term extraction based on a variant of the information gain measure, and by looking within the titles of Web pages for the substring that best matches the selected top-scoring words. Each cluster gets assigned a set of descriptive and discriminative words and clusters sharing the same signature are merged.

3.2. Citation analysis

Since the present study analyses the structure of science on the level of journals, all local citations between papers are aggregated to form a journal cross-citation graph. For cluster analysis we ignored the direction of citations by symmetrising the journal cross-citation matrix; intra-cluster 'self-citations' are counted only once. At the level of journal clusters, the journal cross-citations can be further aggregated into inter-cluster citations.

From the raw number of cross-citations C_{ij} between two journals (or clusters) i and j, a normalised similarity can be calculated as:

$$\frac{C_{ij}}{\sqrt{\sum_k C_{ik} \cdot \sum_k C_{jk}}} \tag{1}$$

For visualisation of the networks we use the similarities just described as edge weights between two clusters or fields (see Fig. 2 for an example). For clustering, however, we calculated the similarity of two journals somewhat differently because we didn't want to ignore, for instance, that both journals could be highly cited by a third one. The similarities S_{ij} used for clustering were found by calculating the cosine of the angle between the pair of vectors containing all symmetric journal cross-citation values between the two respective journals (i and j) and all other journals (i.e., row or column of the matrix C):

$$S_{ij} = \frac{\sum_k C_{ik} \cdot C_{jk}}{\sqrt{\sum_k C_{ik}^2} \cdot \sqrt{\sum_k C_{jk}^2}} \tag{2}$$

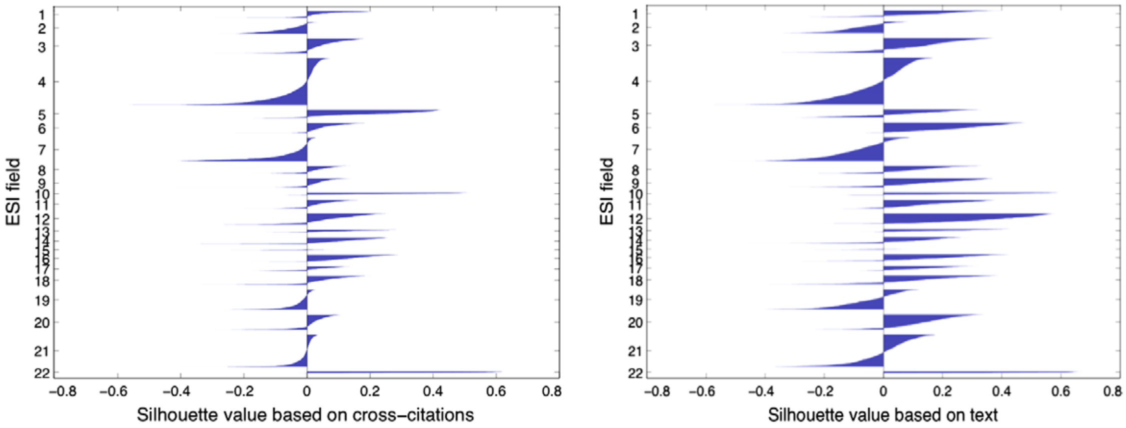

Fig. 1. Silhouette plot for 22 ESI fields based on journal cross-citations (left) and based on text (right).

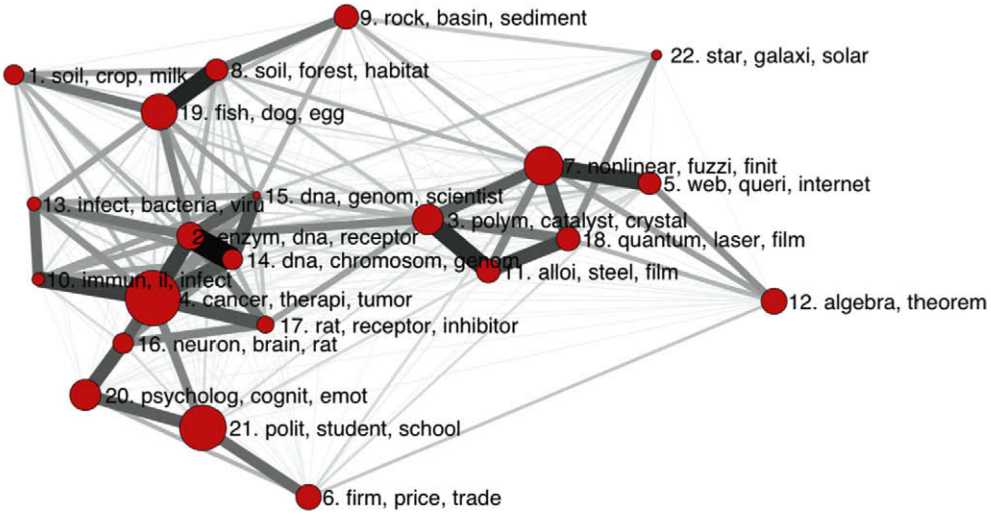

Fig. 2. Network of the 22 ESI fields based on cross-citation links.

Hence, the ultimate similarity of two journals is based on their respective similarities with all other journals. In the remainder of this article we will refer to these similarities as "cosine similarities of cross-citation vectors".

The journal cross-citation graph is also analysed to identify important high-impact journals. We use the PageRank algorithm (Brin & Page, 1998) to determine representative journals in each cluster. Besides, the graph can also be used to evaluate the quality of a clustering outcome.

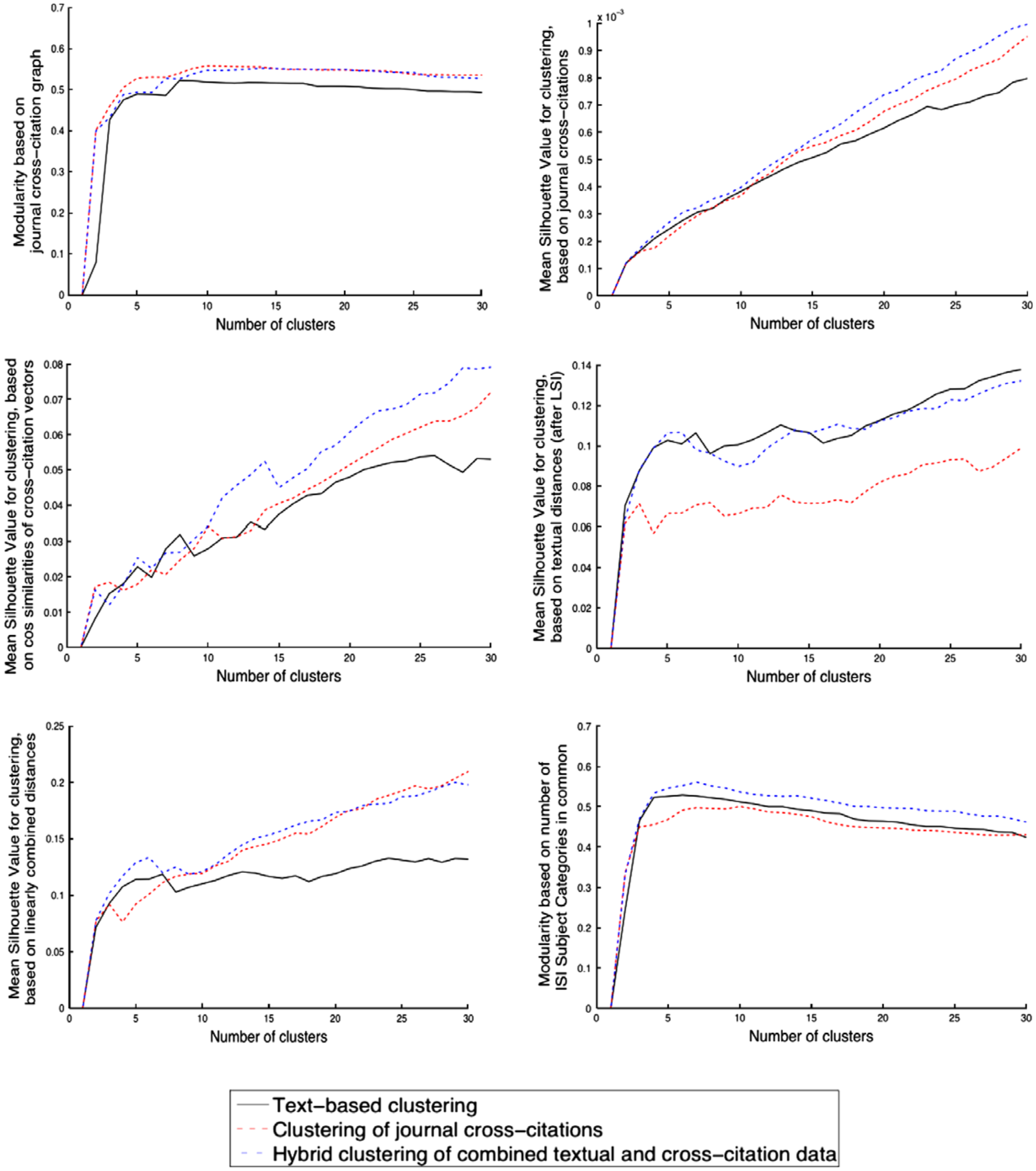

Fig. 3. Performance evaluation of text-based, citation-based and hybrid clustering based on (1) modularity calculated from the journal cross-citation graph, and based on Silhouette curves calculated from (2) journal cross-citations, (3) cosine similarities of cross-citation vectors, (4) text-based distances, and (5) linearly combined distances. For an additional 'external validation' of clustering results compared to ISI Subject Categories, the lower-right Fig. (6) uses modularity computed from a network containing all journals as nodes, but with edge weights equal to the number of ISI Subject Categories commonly assigned to the corresponding journals by ISI/Thomson Scientific (out of the total of 254 categories).

3.3. Clustering

In order to subdivide the journal set into clusters we used the agglomerative hierarchical cluster algorithm with Ward's method (Jain & Dubes, 1988). It is a hard clustering algorithm, which means that each individual journal is assigned to exactly one cluster. Starting from singleton clusters (each object represents one cluster), agglomerative hierarchical clustering proceeds by iteratively grouping those objects or clusters that are least distant from each other according to a linkage criterion. In Ward's method, at each iteration step those objects are grouped such that the increase in total within-cluster sum of squares over all clusters is minimized. This iterative merging continues until all objects are in one big cluster.

3.3.1. Number of clusters

Determination of the optimal number of clusters in a data set is a difficult issue and depends on the adopted validation and chosen similarity measures, as well as on data representation. In general, the number of clusters is determined by comparing the quality of different clustering solutions based on various numbers of clusters. Cluster quality can be assessed by internal or external validation measures. Internal validation solely considers the statistical properties of the data and clusters, whereas external validation compares the clustering result to a known gold standard partition. Halkidi, Batistakis and Vazirgiannis (2001) gave an overview of quality assessment of clustering results and cluster validation measures. The strategy that we adopted to determine the number of clusters is a combination of distance-based and graph-based methods. This compound strategy encompasses observation of a dendrogram, text- and citation-based mean Silhouette curves, and modularity curves. Besides, the Jaccard similarity coefficient is used to compare the obtained results with an intellectual classification scheme. Other methods exist to estimate the number of clusters in a collection. For example, the stability-based method of Ben-Hur, Elisseeff, and Guyon (2002), which allow for visually and quantitatively detecting the most stable number of clusters from a stability diagram. The method can be used with any clustering algorithm and can also detect lack of structure in data. The main idea is that perceived structure should remain stable if only a subsample of objects is available, or if noise objects are added to the data set.

3.3.2. Dendrogram

A preliminary judgment is offered by a dendrogram, which provides a visualisation of the distances between (sub-) clusters (see Fig. 4 for an example). It shows the iterative grouping or splitting of clusters in a hierarchical tree. The horizontal lines connect clusters and the line length represents the distance between two connected clusters. A candidate number of clusters can be determined visually by looking for a cut-off point where an imaginary vertical line would cut the tree such that resulting clusters are well separated, i.e., all horizontal lines that are crossed by the vertical line are relatively long. The tree can be cut off at different levels, tuning the granularity of categorizations, without the need for reclustering. Because of the difficulty to define the optimal cut-off point on a dendrogram (Jain & Dubes, 1988), we complement this method with other techniques.

3.3.3. Silhouette curves

A second appraise for the number of clusters is given by the curve with mean *Silhouette values*. The Silhouette value $S(i)$ for an object i ranges from -1 to +1 and measures how similar it is to objects in its own cluster vs. objects in other clusters (Rousseeuw, 1987). $S(i)$ is defined as follows:

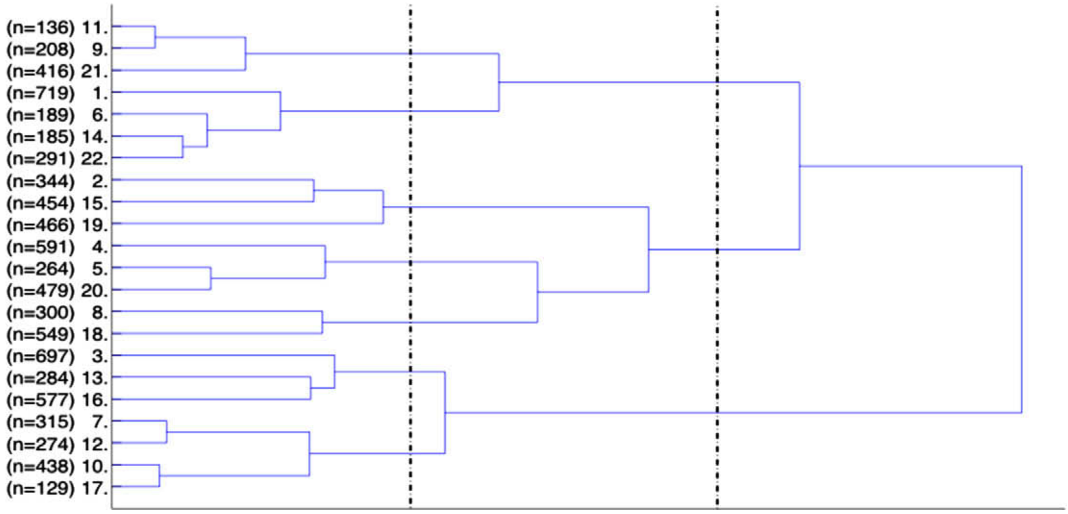

Fig. 4. Cluster dendrogram for hybrid hierarchical clustering of 8305 journals, cut off at 22 clusters on the left-hand side. Two other vertical lines indicate the cut-off points for 7 and 3 clusters.

$$S(i) = \frac{\min_j(B(i, C_j)) - W(i)}{\max\{\min_j(B(i, C_j)), W(i)\}}, \tag{3}$$

where $W(i)$ is the average distance from object i to all other objects within its cluster, and $B(i, C_j)$ is the average distance from object i to all objects in another cluster C_j.

The average Silhouette value for all clustered objects (e.g., journals) is an intrinsic measurement of the overall quality of a clustering solution with a specific number of clusters. Since Silhouette values are based on distances, depending on the chosen distance measure and reference data different Silhouette values can be calculated. For instance, we use the complement of cosine similarity applied to text and citation data.

The quality of a specific partition can be visualised in a *Silhouette plot*. In a Silhouette plot (see Figs. 1 and 5), the sorted Silhouette values of all members of each cluster (or field) are indicated with horizontal lines. The more the Silhouette profile of a cluster (field) is to the right of the vertical line at the value 0, the more coherent the cluster (field) is, whereas negative values indicate that the corresponding objects should rather belong to another cluster (field).

3.3.4. Modularity curves

The quality of a clustering can also be evaluated by calculating the modularity of the corresponding partition of the cross-journal citation graph (Newman, 2006; Newman & Girvan, 2004). Up to a multiplicative constant, modularity measures the number of intra-cluster citations minus the expected number in an equivalent network with the same clusters but with citations given at random. Intuitively, in a good clustering there are more citations within (and fewer citations between) clusters than could be expected from random citing. The expected number of citations between two journals is based on their respective degrees and on the total number of citations in the network.

Newman and Girvan (2004) defined modularity as follows. A $k \times k$ symmetric matrix e is defined whose element e_{ij} is the fraction of all edges in the network that link vertices (journals) in community or cluster i to journals in cluster j. The trace of this matrix $trace(e) = \sum_i e_{ii}$ gives the fraction of edges in the network that connect vertices in the same cluster. The row (or column) sums $a_i = \sum_j e_{ij}$ are further defined, which represent the fraction of edges that connect to vertices in cluster i. Modularity Q is then defined as follows:

$$Q = \sum_i (e_{ii} - a_i^2) = trace(e) - \|e^2\| \tag{4}$$

where $||x||$ indicates the sum of the elements of the matrix x.

For an additional 'external validation' of clustering results, we also use modularity curves computed from a network containing all journals as nodes, but with edge weights equal to the number of ISI Subject Categories commonly assigned to both journals by Thomson Scientific (out of the total of 254).

3.3.5. Jaccard similarity coefficient

The Jaccard index is the ratio of the cardinality of the intersection of two sets and the cardinality of their union. The Jaccard similarity coefficient is an extension of the Jaccard index and can be used as a measure for external cluster validation. It is used to compare a clustering result $C = \{C_1, C_2, \ldots, C_k\}$ with an external partitioning $P = \{P_1, P_2, \ldots, P_l\}$, where k and l represent the number of clusters and partitions, respectively. For C and P we define $n \times n$ matrices M^C and M^P, where n is the total number of documents. Each Boolean value M_{ij}^C indicates whether documents i and j belong to the same cluster in C, while M^P indexes all documents that are in the same partition in P. Let N_{00} represent the number of pairs of documents that do not belong to the same cluster in C, nor in P, i.e., the number of elements (i, j) for which $M_{ij}^C = M_{ij}^P = 0$. Likewise, $N_{\ 01}$ counts the number of elements for which $M_{ij}^C = 0$ and $M_{ij}^P = 1$. N_{10} and N_{11} are defined analogously. The Jaccard coefficient is then defined as follows:

$$J(C, P) = \frac{N_{11}}{N_{11} + N_{01} + N_{10}} \tag{5}$$

The resulting value between 0 and 1 quantifies the correlation between the two binary matrices, while disregarding negative agreements (N_{00}). In Fig. 8, we use the Jaccard index to compare each cluster with every field from the intellectual ESI classification, in order to detect the best-matching fields for each cluster.

3.3.6. Hybrid clustering

As mentioned at the outset, in general four major approaches are used for clustering sets of scientific papers, particularly, the lexical approach and three citation-based methods, namely cross-citation, bibliographic coupling, and co-citation analysis. Each of the methods alone suffers from severe shortcomings. For example, typical problems with bibliographic coupling and co-citations are sparse matrices, the lack of consensual referencing in some areas (Braam, Moed, & Van Raan, 1991b; Jarneving, 2007), document types with insufficient number of references (e.g., letters) that have to be excluded (bibliographic coupling), the incompleteness due to missing citations to recent years (co-citation analysis), the missing 'critical mass' for emerging field detection (co-citation analysis, cf. Hicks, 1987), and the bias towards high-impact journals (co-citation analysis). If strict citation-based criteria are applied, then the resulting citations-by-document matrix is extremely sparse. In this case, rejection of relationship between two entities (e.g., journals or documents) tends to be unreliable. On

Table 1
The 22 broad science fields according to the *Essential Science Indicators* (ESI).

Field #	ESI field	Field #	ESI field
1	Agricultural sciences	12	Mathematics
2	Biology and biochemistry	13	Microbiology
3	Chemistry	14	Molecular biology and genetics
4	Clinical medicine	15	Multidisciplinary
5	Computer science	16	Neuroscience and behavior
6	Economics and business	17	Pharmacology and toxicology
7	Engineering	18	Physics
8	Environment/ecology	19	Plant and animal science
9	Geosciences	20	Psychology/psychiatry
10	Immunology	21	Social sciences
11	Materials sciences	22	Space science

the other hand, any lexical (text-based) approach is usually based on rather rich vocabularies and peculiarities of natural language. The result is, according to our observations, a rather ‘smooth’ or gradual transition between what is related and what is not. Therefore, the relationship is somewhat fuzzy and not always reliable. Hence, both the textual and citation-based approaches provide different perceptions of similarities among the same data. Textual information might indicate similarities that are not visible to bibliometric techniques, but true document similarity can also be obscured by differences in vocabulary use, or spurious similarities might be introduced as a result of textual pre-processing, or because of polysemous words or words with little semantic value. The combination of the two worlds helps to improve the reliability of relationship and therefore of the clustering algorithm as well.

Therefore, the present study combines cross-citation analysis with text mining. The former can be applied to directed links as well as to the symmetrised transaction matrix. Symmetrisation also compensates for the incompleteness caused by the lack of citations to recent years and allows for considering links between journals strong and subject-relevant even if these are asymmetric or even unidirectional. In order to reduce noise caused by ‘small’ journals and extremely weak citation links, thresholds have been applied to both citation links and number of papers (see previous section).

The text mining analysis supplements the citation analysis. In particular, the textual information is integrated with the bibliometric information before the clustering algorithm is applied. In the present study, the actual integration is achieved by a sum of the corresponding similarity matrices. The methodology and advantages of hybrid clustering have been substantiated in more detail in earlier studies devoted to the analysis of different research fields (see Glenisson, Glänzel, Janssens et al. (2005); Janssens, 2007; Janssens et al., 2007, 2008). In addition, the lexical approach allows to ‘label’ clusters using automatically detected salient terms.

In Section 4.3, Silhouette and modularity curves will be used to compare results of text-based, citation-based and hybrid clustering, and we will substantiate that the hybrid method in general outperforms the other two.

3.4. Multidimensional scaling

Multidimensional scaling (MDS) can be used to represent high-dimensional vectors (for example, the centroids of journal clusters) in a lower dimensional space by explicitly requiring that the pairwise distances between the points approximate the original high-dimensional distances as precisely as possible (Mardia, Kent, & Bibby, 1979). If the dimensionality is reduced to two or three dimensions, these mutual distances can directly be visualised. It should, however, be stressed that interpretations concerning such a low-dimensional approximation of very high-dimensional distances must be handled with care.

4. Results

4.1. Evaluation of existing ‘intellectual’ subject-classification schemes

The multidisciplinary databases *Science Citation Index Expanded* (SCIE) and *Social Sciences Citation Index* (SSCI) of Thomson-Reuters (formerly Institute for Scientific Information, ISI, Philadelphia, PA, USA) traditionally did not provide a direct subject assignment for indexed papers. The annual Science Citations Index Guides, the Journal Citation Reports (JCR) and more recently the Website of Thomson Scientific, however, contain regularly updated lists of (S)SCI journals assigned to one or more subject matters (ISI Subject Categories) each. For lack of an appropriate subject-heading system, more or less modified versions of this Subject Category scheme were often used in bibliometric studies too, namely as an indirect subject assignment to individual papers based on the journals in which they had been published. Such assignment systems based on journal classification have been developed among others by Narin and Pinski (see, for instance, Narin, 1976; Pinski & Narin, 1976). This was followed by classification schemes developed by other institutes as well. Nowadays two ISI systems are widely used, in particular, the ISI Subject Categories, which are available in the JCR and through journal assignment in the Web of Science as well, and the Essential Science Indicators (ESI).

While the first system assigns multiple categories to each journal and is too fine grained (254 categories) for comparison with cluster analysis, the ESI scheme is forming a partition (with practically unique journal assignment) and the 22 fields are large enough. Therefore the ESI classification seems to be a good choice for our analysis.

Subject fields will be considered like automatically generated clusters. One precondition for easy comparison with results from hard clustering is that the reference classification system must form a partition of the WoS universe, while most schemes allow multiple assignments (e.g., the above-mentioned ISI Subject Categories). The only commonly known subject scheme for ISI products that meets the criterion is the ESI classification system. This subject-classification scheme is in principle based on unique assignment; only about 0.6% of all journals were assigned to more than one field over a 5-year period. For the present exercise, assignment has to be de-duplicated in the case of journals which merged or split up during the period of 5 years, declaredly a somewhat arbitrary procedure. Nonetheless, the assignment remains correct and results in no more than a slightly narrower scope for several journals. The field structure of the ESI scheme is presented in Table 1.

The question arises whether field classification according to the ESI scheme could still be improved. In particular, we will analyse whether journal assignments to fields can be considered optimum. Fig. 1 presents the evaluation of the 22 ESI fields based on the cross-citation- (left) and text-based (right) Silhouette values (see Section 3.3.3). Since the ESI fields form a partition, this approach allows to evaluate their consistency as if the fields were results of a clustering procedure. Multi-, inter- and cross-disciplinarity of journals can certainly affect the results. Several fields seem not to be coherent enough from both perspectives (i.e., the cross-citation and textual approach). Above all, the Silhouette values of field #2 (Biology and Biochemistry), #4 (Clinical Medicine), #7 (Engineering), #19 (Plant and Animal Science) and #21 (Social Sciences) substantiate that at least five of the 22 fields are not sufficiently coherent.

Table 2
The best 30 TF-IDF terms describing the 22 ESI fields.

Field	Best 30 terms
1	Soil; crop; milk; fruit; seed; cultivar; wheat; dry; rice; ha; chees; diet; fat; ferment; nutrit; meat; farm; grain; starch; fertil; irrig; agricultur; dietari; intak; wine; flour; antioxid; sensori; fatti; sugar
2	Enzym; DNA; receptor; rat; peptid; metabol; lipid; genom; insulin; muscl; transcript; ca2; amino; glucos; mutat; RNA; molecul; diabet; kinas; inhibitor; hormon; mice; mRNA; neuron; fluoresc; mutant; cancer; assai; serum; vitro
3	Polym; catalyst; crystal; ion; bond; molecul; solvent; atom; ligand; hydrogen; film; polymer; adsorpt; aqueou; poli; nmr; methyl; spectroscopi; thermal; chemistri; bi; electrod; spectra; cu; catalyt; cation; mol; copolym; anion; angstrom
4	Cancer; therapi; tumor; infect; surgeri; pain; hospit; arteri; syndrom; diabet; injuri; bone; lesion; chronic; symptom; surgic; renal; breast; carcinoma; serum; transplant; lung; mortal; muscl; liver; coronari; cardiac; physician; rat; hypertens
5	Web; queri; internet; graph; schedul; wireless; semant; logic; node; busi; video; processor; traffic; execut; fuzzi; server; machin; packet; finit; fault; ltd; grid; hardwar; messag; cach; mesh; xml; multimedia; qo; bandwidth
6	Firm; price; trade; economi; busi; capit; invest; wage; tax; financi; organiz; incom; bank; compani; sector; corpor; employ; stock; monetari; custom; labor; privat; strateg; welfar; incent; asset; profit; employe; polit; household
7	Nonlinear; fuzzi; finit; machin; robot; sensor; motion; veloc; nois; crack; thermal; ltd; circuit; vehicl; neural; fuel; voltag; vibrat; elast; beam; shear; turbul; schedul; fault; deform; film; plane; stochast; iter; steel
8	Soil; forest; habitat; river; sediment; ecolog; lake; pollut; land; ecosystem; climat; season; veget; fish; seed; landscap; biomass; nutrient; predat; agricultur; sludg; toxic; groundwat; bird; stream; wast; sea; island; wastewat; wetland
9	Rock; basin; sediment; sea; fault; ocean; miner; seismic; climat; isotop; earthquak; ic; tecton; ma; soil; southern; volcan; atmospher; mantl; geolog; wind; northern; reservoir; metamorph; precipit; river; cretac; lake; faci; eastern
10	Immun; il; infect; antigen; antibodi; mice; vaccin; receptor; cytokin; hiv; cd4; lymphocyt; ifn; autoimmun; dc; cd8; macrophag; viru; inflammatori; peptid; hla; mhc; tnf; nk; ig; molecul; tumor; lp; serum; tcr
11	Alloi; steel; film; coat; corros; glass; crack; microstructur; ceram; powder; fiber; grain; thermal; sinter; polym; crystal; deform; fabric; weld; fibr; fatigu; concret; fractur; si; specimen; cast; tensil; melt; cement; ni
12	Algebra; theorem; finit; asymptot; infin; manifold; let; polynomi; graph; nonlinear; invari; omega; inequ; singular; lambda; convex; proof; compact; ellipt; conjectur; bar; epsilon; infinit; sigma; phi; symmetr; stochast; hyperbol; banach; topolog
13	Infect; bacteria; viru; bacteri; pathogen; DNA; genom; pcr; parasit; coli; enzym; mutant; yeast; microbi; viral; hiv; RNA; vaccin; immun; encod; virul; antibiot; transcript; sp; assai; escherichia; virus; plasmid; clone; candida
14	DNA; chromosom; genom; transcript; mutat; receptor; kinas; mous; mice; RNA; allel; mutant; apoptosi; cancer; mRNA; rat; phenotyp; muscl; polymorph; embryo; tumor; drosophila; phosphoryl; ca2; neuron; actin; clone; encod; prolifer; mitochondri
15	DNA; genom; scientist; receptor; brain; soil; climat; earth; molecul; neuron; RNA; chromosom; mice; mutat; africa; transcript; biologi; ocean; infect; fossil; india; sea; evolutionari; rock; fuel; logic; southern; island; enzym; marin
16	Neuron; brain; rat; receptor; cortex; motor; cognit; cortic; cerebr; mice; neural; stroke; sleep; nerv; lesion; synapt; seizur; epilepsi; axon; schizophrenia; hippocamp; spinal; symptom; pain; alzheim; hippocampu; dopamin; injuri; parkinson; neurolog
17	Rat; receptor; inhibitor; toxic; therapeut; cancer; metabol; vitro; mice; liver; pharmacokinet; oral; therapi; pharmaceut; enzym; antagonist; assai; vivo; pharmacolog; DNA; tablet; inflammatori; tumor; metabolit; lipid; brain; agonist; diabet; cytotox; antioxid
18	quantum; laser; film; beam; spin; atom; scatter; crystal; ion; nonlinear; excit; photon; lattic; nois; thermal; oscil; dope; symmetri; veloc; emiss; finit; decai; spectra; wavelength; si; diffract; neutron; nm; plane; acoust
19	fish; dog; egg; forest; genu; breed; habitat; seed; infect; diet; sp; season; larva; reproduct; leaf; bird; nest; hors; cow; soil; predat; sea; cat; taxa; flower; fruit; veget; parasit; pig; milk
20	Psycholog; cognit; emot; student; mental; adolesc; anxieti; symptom; school; item; child; psychiatr; gender; sexual; attitud; cope; mother; interview; schizophrenia; suicid; skill; questionnair; belief; abus; therapi; men; word; psychotherapi; aggress; mood
21	Polit; student; school; teacher; gender; urban; nurs; court; reform; war; legal; discours; profession; parti; disabl; interview; capit; rural; attitud; child; ethnic; privat; welfar; democraci; democrat; ethic; employ; justic; feder; violenc
22	Star; galaxi; solar; orbit; radio; telescop; emiss; stellar; veloc; disk; galact; earth; planet; flux; atmospher; satellit; wind; mar; cosmic; binari; cloud; flare; dust; spectral; luminos; redshift; jet; accret; dwarf; planetari

4.2. Labelling subject fields using cognitive characteristics and visualisation of the cross-citation network

Simultaneously to the above validation, the textual approach also provides the best TF-IDF terms – out of a vocabulary of 669,860 terms – describing the individual fields. These terms are presented in Table 2. Although these terms already provide an acceptable characterisation of the topics covered by the 22 fields, considerable overlaps are apparent between pairs of fields, respectively: Engineering (#7) and Computer Science (#5), Chemistry (#3) and Materials Science (#11), Plant and Animal Science (#19) and Environment/Ecology (#8), as well as Biology and Biochemistry (#2), Molecular Biology and Genetics (#14) and Clinical Medicine (#4). In addition, the terms characterising the social sciences (#21) reflect a pronounced heterogeneity of the field. The structural map of the 22 ESI fields based on cross-citation links is presented in Fig. 2. For the visualisation we used Pajek (Batagelj & Mrvar, 2003). The network map confirms the strong links we have found based on the best terms between fields #2 and #14, #3 and #11, #5 and #7, and #8 and #19, respectively.

4.3. Cluster analysis: text-based, citation-based and hybrid

Fig. 3 compares the performance of text-based, cross-citation and hybrid clustering by several evaluation methods, for various numbers of clusters. For each of the three clustering types, Fig. 3(1) presents for various cluster numbers (2–30) the modularity calculated from the journal cross-citation graph. Since this evaluation is based on cross-citation data, it is not a surprise that the text-only clustering provides worse results than cross-citation clustering, which performs best here. However, very interesting to note is that the hybrid clustering (integrated text and cross-citation information) provides results highly comparable to those from cross-citation clustering, especially for 7 or more than 12 clusters. The modularity scores for cross-citation clustering indicate that any number of clusters larger than 9 is acceptable. On the other hand, the modularity curve for text-only clustering contains a maximum for eight clusters.

In Fig. 3(2), Silhouette curves based on (the complement of) cross-citation values show the somewhat counter-intuitive but beneficial result that hybrid clustering always performs better than cross-citation clustering, although the evaluation only considers citations here. This demonstrates the power of hybrid clustering: the combined heterogeneous citation–textual approach is superior to both methods applied separately. Nevertheless, this figure does not provide a clear clue with respect to the number of clusters to choose.

Silhouette curves based on the complement of cosine similarities of cross-citation vectors are shown in Fig. 3(3). Again, the hybrid clustering almost always performs best.

In Fig. 3(4), the Silhouette values are computed only from textual distances. Naturally, the citation-based clustering performs worst here, while the integrated clustering scores almost as good as the text-only clustering and for some cluster numbers even better.

Fig. 3(5) shows Silhouette curves based on linearly combined text-based and citation-based distances (with equal weight). Here, combined data and mere citations give comparable results, which might be an indication that there is a preponderance of citation over text data in the combined Silhouette values.

Finally, Fig. 3(6) provides an *external validation* of clustering results by expert knowledge available in the ISI Subject Categories assigned to journals by ISI/Thomson Scientific. The modularity curves are computed from a network containing all journals as nodes, but with edge weights equal to the number of ISI Subject Categories in common (out of the total of 254 categories). Again very interesting to see is that hybrid clustering outperforms text-only and citation-based clustering. The optimal number of clusters according to this type of evaluation is 7.

In Table 3 we compare the quality of the partition of 22 ESI fields with the quality of the 22 clusters resulting from citation-based, text-based and hybrid clustering. The only evaluation measure for which the 22 human-made ESI fields score best is modularity based on ISI Subject Categories. As already explained before, this evaluation type computes modularity from a network containing all journals as nodes and with edge weights equal to the number of ISI Subject Categories commonly assigned to the corresponding journals by ISI/Thomson Scientific (out of the total of 254 categories). Since there is a direct correspondence between the 22 ESI fields and these 254 Subject Categories (a field is an aggregation of multiple

Table 3
Evaluation of 22 ESI fields and 22 citation-based, text-based and hybrid clusters by modularities and mean Silhouette values (MSV). Highest values in each column are shown in bold.

Fields and clusters	Modularity based on journal cross-citation graph	Modularity based on common ISI subject categories	MSV based on textual distances	MSV based on cosine similarities of cross-citation vectors	MSV based on linearly combined distances
22 ESI fields	0.47533	**(0.52604)**	0.057237	0.016017	0.062807
22 Citation-based clusters	**0.54676**	0.44244	0.09319	**0.057337**	**0.18938**
22 Text-based clusters	0.50451	0.45091	0.11829	0.035447	0.12987
22 Hybrid clusters	**0.54677**	**0.48839**	**0.1206**	0.05453	**0.18951**

subject categories), it is not at all surprising (not to mention unfair) that the ESI fields outperform the clusters for this type of evaluation. For all other data-driven evaluation types it is clear that automatic clustering does better than human expert classification. Hybrid clustering always performs at least as good as text-based or citation-based clustering, except for evaluation by cosine similarities of cross-citation vectors. A standard t-test for the difference in means revealed that all differences between mean Silhouette values are statistically significant at the 1% significance level (p-value < 10–8). The modularity (based on journal cross-citation graph) of hybrid and citation-based clustering is 15% higher than that of the 22 ESI fields. The modularity of text-based clustering is 6.1% higher than that of the 22 ESI fields. Modularity of hybrid clustering is 8.4% higher than that of text-based clustering.

The modularity based on common ISI Subject Categories for hybrid clustering is also 8.3% and 10.4% higher than for text-based and citation-based clustering, respectively. It should be noted that the values in Table 3 can differ somewhat from the values in Fig. 3 because, for the sake of a fair comparison with ESI fields, in the table only 7729 journals were considered for which a field assignment was available.

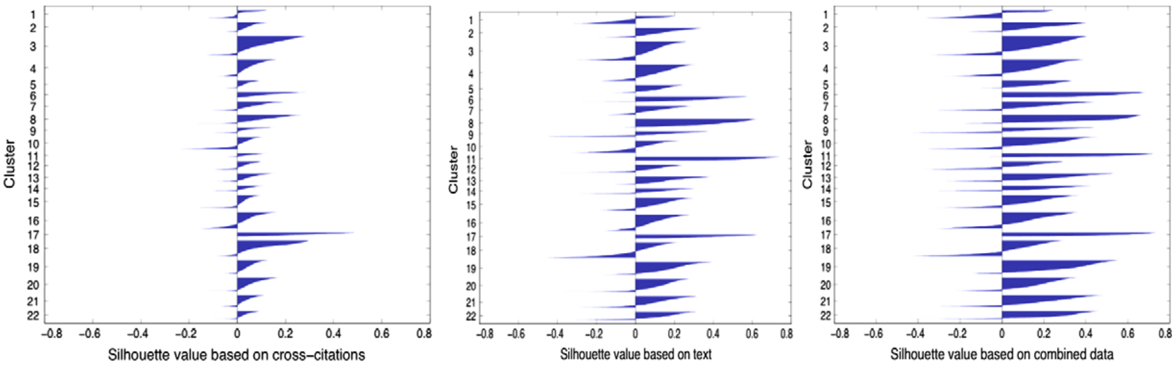

Fig. 5. Evaluation of the hybrid clustering solution with 22 clusters by citation-based Silhouette plot (left), text-based Silhouette plot (centre) and the plot with Silhouette values based on combined data (right).

Table 4
Best 30 TF-IDF terms describing the 3 top-level clusters.

Cluster (# journals)	Best 30 terms
1 (n = 2144)	Polit; student; school; firm; cognit; psycholog; war; gender; price; emot; mental; capit; teacher; trade; economi; reform; adolesc; child; busi; discours; attitud; urban; skill; court; organiz; moral; text; employ; privat; interview
2 (n = 3447)	Soil; finit; film; nonlinear; thermal; ion; crystal; algebra; polym; ltd; forest; atom; veloc; sediment; laser; quantum; motion; graph; theorem; seed; alloi; asymptot; deform; sea; fish; bond; coat; grain; sensor; beam
3 (n = 2714)	Cancer; infect; therapi; tumor; receptor; rat; DNA; pain; diabet; mice; bone; brain; muscl; hospit; syndrom; chronic; injuri; mutat; surgeri; serum; lesion; arteri; neuron; immun; liver; hiv; il; symptom; antibodi; metabol

Table 5
Best 30 TF-IDF terms describing the 7 top-level clusters.

Cluster (# journals)	Best 30 terms
1 (n = 1384)	Polit; firm; war; price; trade; economi; capit; busi; reform; urban; court; parti; gender; privat; invest; organiz; sector; corpor; employ; moral; labor; legal; incom; financi; discours; tax; music; compani; contemporari; welfar
2 (n = 1264)	Soil; forest; sediment; fish; seed; habitat; sea; season; river; lake; sp; basin; rock; genu; veget; crop; leaf; climat; southern; ecolog; egg; land; ocean; fruit; dry; island; biomass; northern; miner; nutrient
3 (n = 1558)	Cancer; infect; tumor; receptor; DNA; rat; therapi; mice; mutat; immun; il; antibodi; liver; serum; genom; enzym; transcript; hiv; diabet; assai; inhibitor; viru; antigen; vaccin; peptid; apoptosi; metabol; carcinoma; lung; renal
4 (n = 1334)	Film; ion; crystal; polym; thermal; atom; alloi; laser; bond; coat; quantum; beam; steel; hydrogen; catalyst; crack; glass; fiber; molecul; nm; spectroscopi; spectra; veloc; ltd; finit; cu; vibrat; solvent; deform; electrod
5 (n = 849)	Algebra; finit; nonlinear; graph; theorem; asymptot; polynomi; fuzzi; infin; manifold; let; invari; stochast; schedul; inequ; convex; robot; singular; proof; logic; omega; machin; iter; topolog; nois; traffic; infinit; metric; motion; lambda
6 (n = 760)	Student; school; cognit; psycholog; teacher; mental; adolesc; emot; child; symptom; anxieti; gender; psychiatr; skill; attitud; abus; teach; item; word; interview; disabl; mother; schizophrenia; sexual; alcohol; speech; instruct; belief; cope; english
7 (n = 1156)	Pain; therapi; hospit; injuri; arteri; nurs; brain; surgeri; neuron; symptom; physician; syndrom; muscl; bone; diabet; rat; lesion; coronari; chronic; stroke; cancer; mortal; cardiac; surgic; receptor; infect; nerv; hypertens; men; infant

4.4. Evaluation of hybrid clusters

The cluster dendrogram shows the structure in a hierarchical order (see Fig. 4). We visually find a first clear cut-off point at three clusters, a second one around seven, and 22 clusters also seemed to be an acceptable/appropriate number. This value coincides with the number of fields according to the ESI classification scheme. The Silhouette plots in Fig. 5 and the mean Silhouette values in Table 3 substantiate that the 22 hybrid clusters are furthermore acceptable for both the citation and the text-mining approach. The same conclusion can be drawn from computed modularity scores.

The number of three clusters results in an almost trivial classification. Intuitively, these three high-level clusters should comprise natural and applied sciences, medical sciences, and social sciences and humanities. The solutions with 3 and 22 clusters will be analysed in more detail in Section 4.5. The solution comprising of seven clusters results in a non-trivial classification. The best TF-IDF terms (see Table 5) show that three of these clusters represent the natural/applied sciences, whereas two classes each stand for the life sciences and the social sciences and humanities. This situation is also reflected by the cluster dendrogram in Fig. 4. A closer look at the best TF-IDF terms reveals that the social-sciences cluster (#1 of the 3-cluster solution) is split into the cluster #1 (economics, business and political science) and #6 (psychology, sociology, education), the life-science cluster (#3 in the 3-cluster scheme) is split into clusters #3 (biosciences and biomedical research) and #7 (clinical, experimental medicine and neurosciences) and, finally, the sciences cluster #2 of the 3-cluster scheme is distributed over three clusters in the 7-cluster solution, particularly, the cluster comprising biology, agriculture and environmental sciences (#2), physics, chemistry and engineering (#4) as well as mathematics and computer science (#5).

Table 6
The 30 best TF-IDF terms describing the 22 hybrid citation–lexical clusters.

Cluster	Best 30 terms
1	Polit; war; court; music; moral; essai; legal; philosophi; narr; text; literari; book; contemporari; french; religi; write; german; discours; ethic; civil; reform; christian; philosoph; justic; fiction; coloni; nineteenth; archaeolog; religion; aesthet
2	Rock; basin; sediment; fault; sea; climat; soil; ic; miner; ocean; seismic; atmospher; river; wind; isotop; veloc; earth; star; tecton; earthquak; solar; precipit; ma; volcan; mantl; southern; lake; satellit; geolog; cloud
3	Receptor; rat; DNA; genom; enzym; transcript; mutat; mice; metabol; peptid; diabet; cancer; insulin; chromosom; kinas; inhibitor; lipid; ca2; muscl; mRNA; RNA; neuron; molecul; vitro; apoptosi; mous; liver; tumor; glucos; assai
4	Polym; ion; catalyst; crystal; bond; molecul; film; solvent; atom; hydrogen; ligand; nmr; polymer; poli; aqueou; adsorpt; thermal; methyl; spectroscopi; spectra; copolym; cation; fiber; cu; nm; bi; coat; mol; blend; nanoparticl
5	Crack; finit; concret; veloc; elast; turbul; vibrat; shear; thermal; nonlinear; beam; deform; fuel; motion; ltd; steel; cylind; combust; convect; flame; fatigu; compress; fractur; vehicl; jet; reynold; plane; wind; stiff; pipe
6	Price; firm; trade; tax; economi; capit; incom; wage; invest; bank; financi; monetari; stock; welfar; labor; inflat; sector; privat; incent; household; earn; game; asset; employ; insur; forecast; reform; foreign; unemploy; volatil
7	Brain; neuron; rat; stroke; lesion; receptor; pain; cerebr; ct; mri; motor; cognit; injuri; spinal; nerv; mr; seizur; epilepsi; cortex; neurolog; tumor; arteri; dementia; sleep; cortic; syndrom; muscl; therapi; symptom; parkinson
8	Algebra; finit; theorem; manifold; infin; nonlinear; polynomi; let; graph; asymptot; singular; omega; invari; inequ; lambda; ellipt; convex; compact; conjectur; epsilon; hyperbol; infinit; proof; bar; symmetr; phi; sigma; topolog; banach; lie
9	Word; cognit; speech; semant; english; linguist; phonolog; stimulu; stimuli; lexic; cue; sentenc; speaker; verb; prime; perceptu; student; acoust; text; item; discours; verbal; auditori; emot; recal; syntact; brain; hear; skill; deficit
10	Nurs; hospit; physician; therapi; cancer; pain; mortal; pregnanc; infant; symptom; ethic; smoke; diabet; infect; birth; ci; men; interview; hiv; injuri; profession; chronic; syndrom; questionnair; worker; school; adolesc; student; surgeri; mental
11	Student; school; teacher; teach; classroom; instruct; skill; academ; curriculum; literaci; disabl; learner; profession; colleg; cognit; peer; child; faculti; gender; reform; write; psycholog; pupil; graduat; attitud; undergradu; text; emot; interview; belief
12	Bone; ey; muscl; sport; athlet; pain; implant; surgeri; fractur; injuri; knee; dental; hip; surgic; nerv; anterior; retin; postop; oral; tendon; corneal; teeth; flap; periodont; graft; lesion; ocular; posterior; therapi; radiograph
13	Infect; dog; hiv; vaccin; viru; hors; cow; milk; parasit; cat; pig; cattl; antibodi; diet; immun; pcr; calv; breed; viral; herd; sheep; pathogen; serum; antigen; therapi; farm; malaria; antibiot; veterinari; hospit
14	Firm; organiz; busi; librari; web; internet; custom; compani; employe; job; onlin; brand; team; strateg; journal; career; corpor; satisfact; student; price; trust; advertis; academ; profession; librarian; attitud; organis; leadership; cognit; enterpris
15	Soil; seed; crop; cultivar; leaf; fruit; bacteria; wheat; dry; rice; enzym; pathogen; shoot; nitrogen; microbi; ferment; bacteri; ha; pollut; sediment; milk; nutrient; DNA; fertil; germin; biomass; seedl; season; agricultur; coli
16	Cancer; tumor; therapi; il; carcinoma; transplant; breast; immun; infect; lung; liver; renal; antibodi; receptor; antigen; mice; malign; serum; chronic; prostat; lesion; surgeri; tumor; chemotherapi; mutat; inflammatori; DNA; bone; recurr; surgic
17	Coronari; arteri; cardiac; ventricular; myocardi; hypertens; cardiovascular; aortic; atrial; infarct; diabet; therapi; valv; vascular; stent; endotheli; surgeri; pulmonari; mortal; cholesterol; systol; bypass; syndrom; lv; graft; diastol; renal; vein; rat; echocardiographi
18	Fuzzi; schedul; robot; logic; machin; graph; nonlinear; traffic; web; asymptot; circuit; neural; nois; finit; stochast; fault; queri; custom; wireless; video; node; semant; heurist; antenna; motion; markov; polynomi; bayesian; iter; processor
19	Forest; habitat; fish; genu; egg; predat; season; sp; nest; sea; ecolog; larva; reproduct; lake; bird; prei; island; taxa; seed; river; veget; soil; breed; southern; ecosystem; nov; mate; genera; diet; biomass
20	Film; alloi; laser; quantum; crystal; ion; steel; thermal; atom; beam; coat; glass; si; grain; microstructur; corros; silicon; dope; spin; ceram; powder; nm; scatter; fabric; neutron; diffract; dielectr; photon; cu; electrod
21	Psycholog; adolesc; mental; emot; cognit; symptom; child; anxieti; psychiatr; abus; student; school; alcohol; schizophrenia; sexual; mother; gender; attitud; suicid; interview; cope; violenc; therapi; questionnair; youth; disabl; offend; men; belief; item
22	Polit; urban; parti; gender; reform; economi; capit; democrat; democraci; employ; sector; war; land; sociolog; geographi; union; labor; rural; elect; welfar; ethnic; labor; discours; immigr; privat; actor; trade; civil; poverti; firm

The hybrid, i.e., the combined citation–textual based clustering yields acceptable results (see Fig. 5), and is distinctly superior to both methods applied separately. Nonetheless, we must not conceal that we can also find clusters of lesser quality, notably cluster #1, in the hybrid classification.

4.5. Cognitive characteristics of clusters

As already mentioned in the previous section, another nice point to cut off the dendrogram is at three clusters (cf. the right-most vertical line in Fig. 4). Although this refers to a rather trivial case, it might be worthwhile to have a look at term representation of this structure before we deal with ‘labelling’ the 22 clusters that we have obtained from the hybrid algorithm. This will also help us to understand the hierarchical architecture of the subject structure of science. Table 4 lists the best 30 terms for each of the three top-level clusters which definitely confirm the presence of the expected clusters. Indeed, cluster #1 comprises the social-sciences, cluster #2 the natural and applied sciences and cluster #3 the medical sciences. The distribution of journals over clusters is surprisingly well-balanced.

According to the terms in Table 4, economics, business and psychology are the dominant issues in the first cluster which represents the social sciences. The most characteristic terms of the second cluster represent the full spectrum of the sciences including mathematics, geosciences and engineering. Also some subfields of agriculture and environment are covered. Cluster #3, finally, covers biosciences, biomedical research, clinical and experimental medicine and neurosciences.

The 30 best TF-IDF terms describing the 22 hybrid citation–lexical clusters are listed in Table 6. Cluster #1 of the 3-cluster scheme is split up in seven clusters, particularly, in #1, #6, #9, #11, #14, #21 and #22. However, this sub-classification of the social sciences is less straightforward. Cluster #6 represents economics and business and political science, cluster #9 stands for psychology and linguistics, cluster #21 covers psychology and psychiatry, #11 comprises sociology and education, and cluster #1 is rather focussed on the humanities. Cluster #14 and #22 seem to have more heterogeneous profiles among these ‘social and humanity clusters’. Although cluster #14 largely covers information and library science, the terms reflect a large overlap with other clusters. The same applies to cluster #22, which has obviously an even fuzzier structure. On the other hand, #9 and #21 are both covering psychology but focussing on different aspects, namely cognitive (#9) and medical (#21) issues.

Similarly to the structural analysis in Section 4.2, we use now a network with citation links among clusters to study the relationship of clusters based on hybrid cross-citation/textual information (see Fig. 6). The observations made on the basis of most characteristic terms are confirmed by the link structure. The social-sciences and humanities clusters form two groups that are each strongly interlinked; one consists of clusters #1, #6, #14 and #22 with focus on humanities, economics, business, political and library science, the other one comprises #9, #11 and #21 with sociology, education and psychology. This is in line with the hierarchical structure shown in Fig. 4. These two groups correspond to the two social-sciences clusters in the 7-cluster solution (cf. Section 4.4).

In the natural and applied sciences, we have found eight clusters, particularly, #2, #4, #5, #8, #15, #18, #19 and #20. On the basis of the most important TF-IDF terms (see Table 6) we can assign clusters #2, #15 and #19 to geosciences, environ-

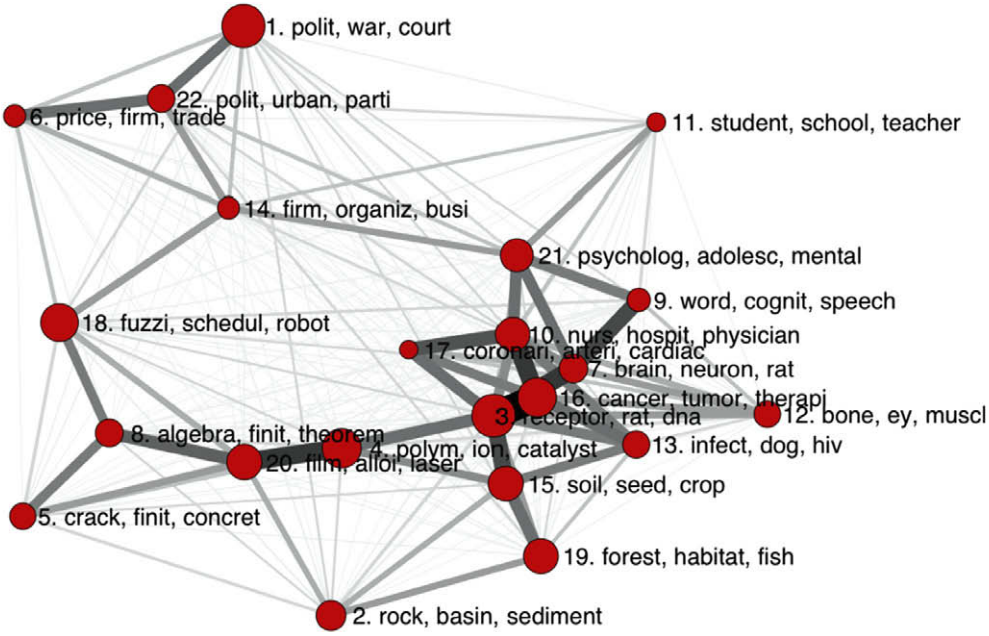

Fig. 6. Network structure of hybrid clusters represented by the three most important terms.

mental science, biology and agriculture, which, in turn, form a larger group corresponding to the first of the three ‘‘mega-clusters" in the 7-cluster solution. The graphic network presentation in Fig. 6 confirms this interpretation. These three clusters form a group at the bottom. The other sciences clusters are more clearly recognisable, and distinctly separate fields. Thus clusters #4 represents chemistry, #20 physics, #5 engineering, #8 mathematics, and cluster #18 computer science. These science clusters form two groups, #4, #20 and #5 form one group of chemistry, physics and engineering, while #8 and #18 form the third group comprising mathematics and computer science. The network presentation and the hierarchical architecture in the dendrogram confirm the term characterisation.

The interpretation of the most characteristic terms of the life-science clusters is somewhat more complicated. Here we have a biomedical and a clinical group. These two groups are in line with the hierarchical structure of the dendrogram in Fig. 4 but less clearly distinguished in the graphical network presentation (Fig. 6). Nonetheless, the terms provide an excellent description for at least some of the medical clusters: cluster #7 stands for the neuro- and behavioral sciences, #3 for bioscience, #10 for the clinical and social medicine, #13 microbiology and veterinary science, #12 non-internal medicine, #16 hematology and oncology and #17 cardiovascular and respiratory medicine. According to the dendrogram clusters 3, 13, 16 and clusters 7, 10, 12, 17 form one larger cluster each. On the basis of the best terms, we can characterise these groups as the bioscience–biomedical and the clinical and neuroscience group, respectively.

In order to gain a better understanding of the cluster structure, we have ranked the journals of each of the 22 clusters according to a modified version of Google's PageRank algorithm (Brin & Page, 1998) in which the number of citations is taken into account, normalised by the number of published papers. The following equation was used,

$$PR_i = \frac{(1-\alpha)}{n} + \alpha \sum_j PR_j \frac{a_{ji}/P_i}{\sum_k \frac{a_{jk}}{P_k}} \tag{6}$$

where PR_i is the PageRank of journal i, a is a scalar between 0 and 1 (a = 0.9 in our implementation), n is the number of journals in the cluster, a_{ji} the number of citations from journal j to journal i, and P_i is the number of papers published by journal i, all in the period under study. Both sums iterate over the journals in the same cluster that contains journal i. Journal self-citations were removed prior to application of the algorithm.

The five journals with highest PageRank are presented in Table 7. The PageRank of a journal can be understood here as the probability that a random reader will be reading that journal, when he randomly, continuously, and with equal probability looks up cited references to other journals (different from the current one), but once in a while randomly picks another journal from the library (cluster). Journals from arts and humanities (according to the ISI Subject Categories) were removed prior to application of the PageRank algorithm because of the low reliability of citation indicators in these disciplines. Zhang, Glänzel, and Liang (2009) have shown that high entropy of journal cross-citations, relatively low impact and high share of journal self-citation makes it difficult to build reliable citation indicators for the humanities. This has to do with the subject-specific peculiarities in scholarly communication.

In general, the journals ranking highest represent their cluster in an adequate manner (cf. Table 7). Results of the PageRanking thus provide a realistic and representative picture of the hybrid clustering.

4.6. Comparison of subject and cluster structure

In this subsection we compare the structure resulting from the hybrid clustering with the ESI subject classification. This comparison is based on the *centroids* of the clusters and fields. The centroid of a cluster or field is defined as the linear combination of all documents in it and is thus a vector in the same vector space. For each cluster and for each field, the centroid was calculated and the MDS of pairwise distances between all centroids is shown in Fig. 7.

In Fig. 8, we use the Jaccard index to determine the concordance between our clustering solution and the ESI Scheme by comparing each cluster with every field, in order to detect the best-matching fields for each cluster. The darker a cell in the matrix, the higher the Jaccard index, and hence the more pronounced the overlap between the corresponding cluster and ESI field. For example, cluster #4 (Chemistry) definitely corresponds to ESI field #3 (Chemistry). The same applies to field and cluster #6 (Economics and business). Clearly, ESI field #21 has the least concordance as this field is spread over seven clusters. It is defined as one single field in social sciences. It is not a surprise that the strongest match is found with our somewhat 'fuzzy' multidisciplinary social cluster. On the other hand, clusters #13 and #14 are quite similarly spread over four ESI fields each. The analysis does not result in a one-to-one correspondence between clusters and fields. For example, ESI field 20 (Psychology/Psychiatry) is the best-matching field for both clusters 9 and 11, whereas ESI field 8 (Environment/Ecology) is not the best-matching field for any cluster. Such deviations from one-to-one correspondence are interesting starting points to look for possible improvements of the intellectual classification scheme.

4.7. Migration of journals among subject fields and clusters

If clustering algorithms are adjusted or changed, one can observe the following phenomenon. Some units of analysis are leaving clusters they formerly belonged to and end up in different clusters. This phenomenon is called 'migration'. We can distinguish between 'good migration' and 'bad migration'. 'Good migration' is observed if the goodness of the unit's classification improves, otherwise we speak about 'bad migration'. We can also apply this notion of migration to the comparison of clustering results with any reference classification. In the following we will use the ESI scheme as reference classification.

Table 7
The five most important journals of each cluster according to a modified version of Google's PageRank algorithm (see Formula 6).

Cluster 1	Cluster 2	Cluster 3	Cluster 4
1. Yale Law J	1. Annu Rev Astron Astr	1. Annu Rev Biochem	1. Chem Rev
2. Harvard Law Rev	2. Astrophys J Suppl S	2. Cell	2. Prog Polym Sci
3. Stanford Law Rev	3. Earth-Sci Rev	3. Nat Rev Mol Cell Bio	3. Accounts Chem Res
4. Am Hist Rev	4. Rev Mineral Geochem	4. Annu Rev Cell Dev Bi	4. Annu Rev Phys Chem
5. Columbia Law Rev	5. Annu Rev Earth Pl Sc	5. Annu Rev Genet	5. Adv Drug Deliver Rev
Cluster 5	Cluster 6	Cluster 7	Cluster 8
1. Prog Energ Combust	1. Q J Econ	1. Annu Rev Neurosci	1. J Am Math Soc
2. Annu Rev Fluid Mech	2. J Financ	2. Nat Rev Neurosci	2. Ann Math
3. Prog Aerosp Sci	3. J Econ Lit	3. Neuron	3. Acta Math-Djursholm
4. P Combust Inst	4. J Polit Econ	4. Nat Neurosci	4. Invent Math
5. Combust Flame	5. J Financ Econ	5. Prog Neurobiol	5. Commun Pur Appl Math
Cluster 9	Cluster 10	Cluster 11	Cluster 12
1. Psychol Rev	1. Milbank Q	1. Rev Educ Res	1. Crit Rev Oral BIOL M
2. Behav Brain Sci	2. Annu Rev Publ Health	2. Am Educ Res J	2. Prog Retin Eye Res
3. Cognitive Psychol	3. Jama-J Am Med assoc	3. Educ Eval Policy An	3. Am J Sport Med
4. J Exp Psychol Gen	4. Health Serv Res	4. Educ Psychol-US	4. Periodontol 2000
5. Cognition	5. J Health Soc Behav	5. J Learn Sci	5. Sports Med
Cluster 13	Cluster 14	Cluster 15	Cluster 16
1. J Acq Immun Def Synd	1. Admin Sci Quart	1. Annu Rev Plant Biol	1. Annu Rev Immunol
2. Aids	2. Acad Manage J	2. Plant Cell	2. Nat Rev Immunol
3. Clin Microbiol Rev	3. Organ Sci	3. Curr Opin Plant Biol	3. Nat Immunol
4. J Infect Dis	4. Acad Manage Rev	4. Annu Rev Phytopathol	4. Ca-cancer J Clin
5. Clin Diagn Virol	5. Mis Quart	5. Microbiol Mol Biol R	5. Immunity
Cluster 17	Cluster 18	Cluster 19	Cluster 20
1. Circulation	1. ACM Comput Surv	1. Annu Rev Ecol Evol S	1. Rev Mod Phys
2. Circ Res	2. J ACM	2. Systematic Biol	2. Mat Sci Eng R
3. J Am Coll Cardiol	3. J R Stat Soc B	3. Annu Rev Entomol	3 Annu Rev Nucl Part S
4. Arterioscl Throm Vas	4. Vldb J	4. Oceanogr Mar Biol	4. Phys Rep
5. Cardiovasc Res	5. IEEE T Robotic Autom	5. Trends Ecol Evol	5. Prog Mater Sci
Cluster 21	Cluster 22		
1. Psychol Bull	1. Am Polit Sci Rev		
2. Annu Rev Psychol	2. World Polit		
3. J Pers Soc Psychol	3. Am J Polit Sci		
4. Arch Gen Psychiat	4. Am Sociol Rev		
5. Pers Soc Psychol Rev	5. Annu Rev Sociol		

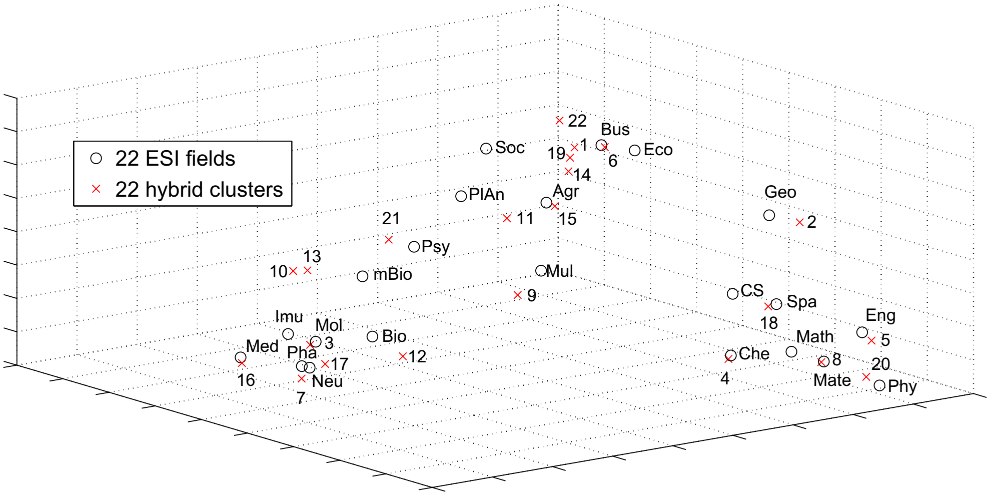

Fig. 7. Three-dimensional MDS map visualising distances between the centres (centroids) of the 22 ESI fields and the 22 clusters containing 8305 WoS journals.

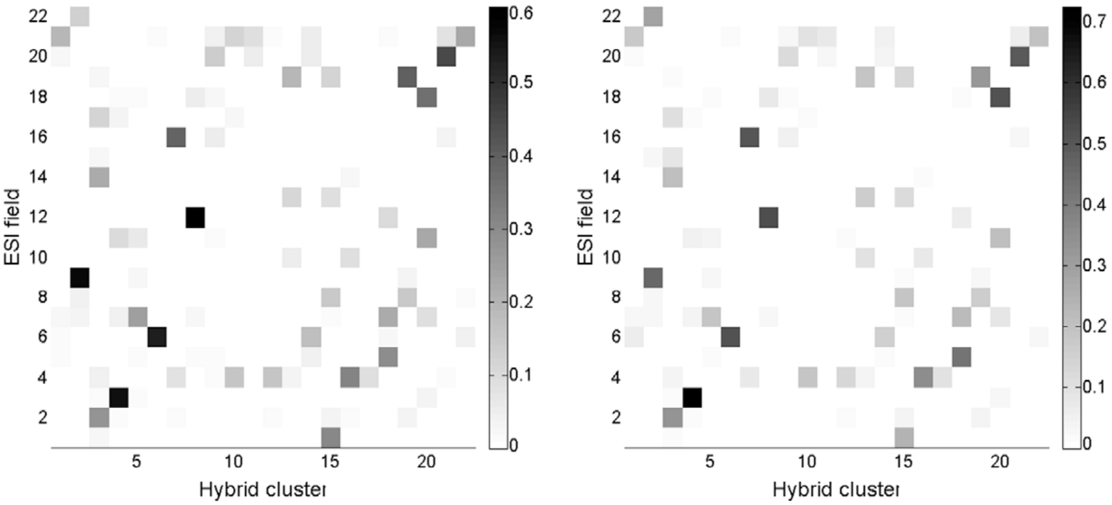

Fig. 8. Concordance between our clustering solution and the ESI Scheme visualised by grayscale cells representing the Jaccard index for each cluster and field pair. The darkest cells represent the best-matching pairs of fields and clusters. In the left figure, the Jaccard index is computed from the number of journals a cluster and a field have in common, while the right figure takes the size of each journal into account by counting the numbers of overlapping papers.

In the previous section we visualised the concordance between the clustering and the ESI classification. To determine for each ESI field the cluster that best matches the field, we used the Jaccard index on basis of the number of overlapping journals (cf. upper part of Fig. 8). Out of 8305 journals under study, there were more than one third, namely, 3204 journals that were not assigned to the cluster which best matches their ESI field. As already mentioned above, we call these journals 'migrated journals'. The largest 'exodus' comprising 226 migrating journals occurred from the ESI "Engineering" field to cluster #18 (Computer science), whereas the best-matching cluster for the Engineering field is actually cluster #5 (Engineering). The top 10 strongest patterns of migration are listed in Table 8, which indicate possible improvements of journal assignments.

To measure the quality of migrations, we calculated the differences in Silhouette values before and after migration (based on textual and citation distances), for each migrated journal. Most migrated journals improved their Silhouette values. In the following, we will give some examples of good migrations and bad migrations.

'Good migrations' are observed if journals improved their Silhouette values after migration. Based on their titles and scopes (not shown), apparently they should indeed be assigned to the cluster to which they have moved. We observed numerous good migrations and the following cases will serve just as examples.

The *Journal of Analytical Chemistry* and *Chemia Analityczna* migrated from ESI field #7 (Engineering) to cluster 4 (chemistry). The best-matching ESI cluster were field #3 (Chemistry) in this case (cf. Fig. 8). Similarly, *Land Economics*, *Developing Economies* and *Economic Development and Cultural Change* migrated from field #21 (Social Sciences, general) to the more specific cluster 6 (economics and business). Here, the corresponding ESI field were #6 (Economics and business). In the life sciences, we found the following good migration. The journals *Neuropathology, Revista de Neurologia, Current Opinion in Neurology, Revue Neurologique, Lancet Neurology, European Journal of Neurology, Neurologist, Nervenheilkunde, Visual Neuroscience, Seminars in Neurology, Epilepsy and Behavior* and *Journal of Neuroimaging* migrated from field #4 (Clinical Medicine) to cluster #7 (neuroscience and behavior) which rather corresponds to ESI field #16 (Neuroscience and behavior). Finally, we mention a migration between engineering and mathematics. The journals *Quarterly of Applied Mathematics, Bit Numerical Mathematics, Siam Journal on Discrete Mathematics* and *Discrete Applied Mathematics*, which were assigned to the ESI field

Table 8
Top 10 strongest migration patterns.

Migration pattern	Number of migrated journals
From ESI field 7 (engineering) to cluster 18	226
From ESI field 14 (molecular biology and genetics) to cluster 3	159
From ESI field 21 (social sciences, general) to cluster 10	145
From ESI field 11 (materials science) to cluster 20	139
From ESI field 4 (clinical medicine) to cluster 7	132
From ESI field 19 (plant and animal science) to cluster 15	108
From ESI field 21 (social sciences, general) to cluster 21	98
From ESI field 7 (engineering) to cluster 20	95
From ESI field 4 (clinical medicine) to cluster 3	86
From ESI field 8 (environment/ecology) to cluster 15	86

Engineering (field #7), were found in our 'Mathematics' cluster (#8) which in turn corresponds to WSI field #12 (Mathematics). In the case of bad migration, the Silhouette values decreased after migration, that is, their Silhouette values in the ESI scheme were better than in the hybrid clustering. The reasons for this phenomenon are not always clear. According to their titles and scopes this migration is not always convincing. For instance, *Journal of Astrophysics and Astronomy, New Astronomy, Astrophysical Journal* and *Astronomy and Astrophysics* migrated from the ESI field 22 (Space Science) to cluster 2 (geosciences) corresponding to ESI field #9, where we have to admit that journals in astronomy and astrophysics are in general spread over the geosciences and physics clusters. *Viral Immunology* migrated from field #10 (Immunology) to cluster #13 (microbiology and veterinary science) and *Canadian Journal of Microbiology* migrated from field #13 (Microbiology) to cluster #15 (agricultural and environmental sciences). Both clusters are rather spread over several ESI fields each (see Fig. 8).

The distinction between good and bad makes a target-oriented adjustment of the existing classification scheme possible. Good migration can be used to reassign journals within the old scheme on the basis of the concordance with the results of clustering.

5. Conclusions

The hybrid clustering using textual information and cross-citations provided good results and proved superior to its two components when applied separately. The goodness of the resulting classification was even better than that of the "intellectual" reference scheme, the ESI subject scheme. Both classification systems form partitions of the Web of Science so that the direct comparison of clusters and fields was possible. According to our expectations, not all clusters have a unique counterpart in the ESI scheme and *vice versa* although the number of clusters coincided with the number of ESI fields. Although the Silhouette and modularity values substantiate a more coherent structure of the hybrid clustering as compared with the ESI subject scheme, not all clusters are of high quality. Problems have been found, for instance, in clusters #1 and #12 where interdisciplinarity and strong links with other clusters distort the intra-cluster coherence. However, intellectual classification schemes usually do have a category "multidisciplinary sciences" as well. Although the result of a hard clustering algorithm often does contain a cluster with objects (journals) not strongly related to any other cluster, forming a "multidisciplinary sciences" cluster is not an inherent goal of the algorithm, and actually is not really meaningful either in the light of our outset goal to improve the classification of the sciences. Consequently, real multidisciplinary journals are scattered around different clusters.

Based on the external validation of clustering results by expert knowledge present in ISI subject categories, seven clusters seem to yield best results. Although there is no adequate subject-classification scheme with 7 categories to be used as reference system, a more detailed analysis of this solution will be part of future research. Additional ideas for future research are a further improvement of the hybrid clustering algorithm by iterative cleaning of clusters as a post-processing step; allowing multiple assignments by fuzzy clustering; evaluating other algorithms like spectral clustering; and, finally, dynamic analysis by dynamic hybrid clustering.

The continuous rise of computing power might one day allow a large-scale mapping of the scientific universe explorable at various levels of detail. What's more, application of advanced natural language processing and machine summarisation at the scale of large bibliographic corpora might offer some insight into semantics beyond mere statistical processing.

Acknowledgements

The authors would like to thank Dr. Ricardo Baeza-Yates and an anonymous reviewer for fruitful comments, and acknowledge support from the Research Council KUL: GOA AMBioRICS, CoE EF/05/007 SymBioSys, PROMETA, several PhD/postdoc and Fellow Grants; Flemish Government: Steunpunt O&O Indicatoren; FWO: PhD/postdoc Grants, Projects G.0241.04 (Functional Genomics), G.0499.04 (Statistics), G.0232.05 (Cardiovascular), G.0318.05 (subfunctionalization), G.0553.06 (VitamineD), G.0302.07 (SVM/Kernel), research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, GBOU-McKnow-E (Knowledge management algorithms), GBOU-ANA (biosensors), TAD-BioScope-IT, Silicos; SBO-BioFrame, SBO-MoKa, TBM-Endometriosis, TBM-IOTA3, O&O-Dsquare; Belgian Federal Science Policy Office: IUAP P6/25 (BioMaGNet, Bioinformatics and Modeling: from Genomes to Networks, 2007–2011) ; EU-RTD: ERNSI: European Research Network on System Identification; FP6-NoE Biopattern; FP6-IP e-Tumours, FP6-MC-EST Bioptrain, FP6-STREP Strokemap.

References

Bader, B. W., & Kolda, T. G. (2006). Algorithm 862: MATLAB tensor classes for fast algorithm prototyping. *ACM Transactions on Mathematical Software, 32*(4), 635–653.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Cambridge: Addison-Wesley.

Batagelj, V., & Mrvar, A. (2003). Pajek – analysis and visualization of large networks. In M. Jünger & P. Mutzel (Eds.), *Graph drawing software* (pp. 77–103). Berlin: Springer.

Ben-Hur, A., Elisseeff, A., & Guyon. I. (2002). A stability based method for discovering structure in clustered data. In *Pacific symposium on biocomputing* (pp. 6–17).

Berry, M., Dumais, S. T., & O'brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM Review, 37*(4), 573–595.

Boyack, K. W., Börner, K., & Klavans, R. (2009). Mapping the structure and evolution of chemistry research. *Scientometrics, 79*(1), 45–60.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991a). Mapping of science by combined co-citation and word analysis, part I: Structural aspects. *JASIS, 42*(4), 233–251.

Braam, R. R., Moed, H. F., & Van Raan, A. F. J. (1991b). Mapping of science by combined co-citation and word analysis, part II: Dynamical aspects. *JASIS, 42*(4), 252–266.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Calado, P., Cristo, M., Goncalves, M. A., de Moura, E. S., Ribeiro-Neto, B., & Ziviani, N. (2006). Link-based similarity measures for the classification of web documents. *JASIST, 57*(2), 208–221.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks – An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Callon, M., Law, J., & Rip, A. (Eds.). (1986). *Mapping the dynamics of science and technology: Sociology of science in the real world*. London: Macmillan Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *JASIS, 41*(6), 391–407.

Fano, R. M. (1956). Information theory and the retrieval of recorded information. In J. H. Shera, A. Kent, & J. W. Perry (Eds.), *Documentation in action* (pp. 238–244). New York: Reinhold Publ. Co.

Garfield, E. (1975). ISIS atlas of science may help students in choice of career in science. *Current Contents, 29*, 5–8.

Garfield, E. (1988). The encyclopedic ISI atlas of science launches three new sections – Biochemistry, immunology, and animal & plant sciences. *Current Contents, 7*, 3–8.

Geraci, F., Maggini, M., Pellegrini, M., & Sebastiani, F. (2008). Cluster generation and labeling for web snippets: A fast, accurate hierarchical solution. *Internet Mathematics, 3*(4), 413–444.

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.

Glenisson, P., Glänzel, W., Janssens, F., & De Moor, B. (2005). Combining full text and bibliometric information in mapping scientific disciplines. *Information Processing & Management, 41*(6), 1548–1572.

Glenisson, P., Glänzel, W., & Persson, O. (2005a). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics, 63*(1), 163–180.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems, 17*(2–3), 107–145.

Hatcher, E., & Gospodnetic, O. (2004). *Lucene in action*. New York: Manning Publications Co.

He, X., Ding, C. H. Q., Zha, H., & Simon, H. D. (2001). Automatic topic identification using webpage clustering. In *Proceedings of the 2001 IEEE international conference on data mining* (pp. 195–202). Washington, DC, USA: IEEE Computer Society.

He, X., Zha, H., Ding, C. H. Q., & Simon, H. D. (2002). Web document clustering using hyperlink structures. *Computational Statistics & Data Analysis, 41*(1), 19–45.

Hicks, D. (1987). Limitations of co-citation analysis as a tool for science policy. *Social Studies of Science, 17*(2), 295–316.

Jain, A., & Dubes, R. (1988). *Algorithms for clustering data*. New Jersey: Prentice Hall.

Janssens, F. (2007). *Clustering of scientific fields by integrating text mining and bibliometrics*. Ph.D. Thesis, Faculty of Engineering, Katholieke Universiteit Leuven, Belgium. <http://www.hdl.handle.net/1979/847>.

Janssens, F., Glänzel, W., & De Moor, B. (2007). Dynamic hybrid clustering of bioinformatics by incorporating text mining and citation analysis. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining, San Jose, CA, USA, August 2007* (pp. 360–369). New York: ACM.

Janssens, F., Glänzel, W., & De Moor, B. (2008). A hybrid mapping of information science. *Scientometrics, 75*(3), 607–631.

Janssens, F., Leta, J., Glänzel, W., & De Moor, B. (2006). Towards mapping library and information science. *Information Processing & Management, 42*(6), 1614–1642 [Special Issue on Informatics].

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 *Proceedings of the 1st International Conference on Multidisciplinary Information Sciences and Technologies (InSciT2006)*. Mérida, Spain, October 2006 (pp. 615–619).

Jarneving, B. (2005). *The combined application of bibliographic coupling and the complete link cluster method in bibliometric science mapping*. Ph.D. Thesis, University College of Borås/Göteborg University, Sweden.

Jarneving, B. (2007). Bibliographic coupling and its application to research-front and other core documents. *Journal of Informatics, 1*(4), 287–307.

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

Kostoff, R. N., Buchtel, H. A., Andrews, J., & Pfeil, K. M. (2005). The hidden structure of neuropsychology: Text mining of the journal cortex, 1991–2001. *Cortex, 41*(2), 103–115.

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change, 68*(3), 223–253.

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal–journal citation relations using the journal citation reports? *JASIST, 57*(5), 601–613.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *JASIST, 60*(2), 348–362.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis*. London, UK: Harcourt Brace & Co., Academic Press.

Marshakova, I. V. (1973). System of connections between documents based on references (as the science citation index). *Nauchno-Tekhnicheskaya Informatsiya Seriya, 2*(6), 3–8.

Modha, D. S., & Spangler, W. S. (2000). Clustering hypertext with applications to web searching. In *Proceedings of the 11th ACM international conference on hypertext and hypermedia* (pp. 143–152). New York, NY, USA: ACM Press.

Moya-Anegon, F. de., Vargas-Quesada, B., Chinchilla Rodriguez, Z., Corera-Alvarez, E., Munoz Fernandez, F. J., & Herrero-Solana, V. (2007). Visualizing the marrow science. *JASIST, 58*(14), 2167–2179.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, D.C: Computer Horizons, Inc.

Newman, M. E. J. (2006). Modularity and community structure in networks. *PNAS US, 103*, 23.

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E, 69*, 2.

Noyons, E. C. M. (1999). *Bibliometric mapping as a science and research management tool*. Leiden, The Netherlands: DSWO Press, Leiden University.

Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications. *Information Processing & Management, 12*(5), 297–312.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*(1), 53–65.

Salton, G., & Mcgill, M. J. (1986). *Introduction to modern information retrieval*. New York: McGraw-Hill, Inc.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIS, 24*(4), 265–269.

Small, H. (1978). Cited documents as concept symbols. *Social Studies of Science, 8*, 240–327.

Small, H. (1998). A general framework for general large-scale maps of science in two or three dimensions: The SciViz system. *Scientometrics, 41*(1–2), 125–133.

Treeratpituk, P., & Callan, J. (2006). Automatically labeling hierarchical clusters. *Proceedings of the 2006 international conference on digital government research* (Vol. 151, pp. 167–176). San Diego, CA, USA: ACM.

Wang, Y., & Kitsuregawa, M. (2002). Evaluating contents-link coupled web page clustering for web search results. In *Proceedings of the 11th international conference on information and knowledge management* (pp. 499–506). New York, NY, USA: ACM Press.

Weiss, R., Vélez, B., Sheldon, M. A., Namprempre, C., Szilagyi, P., Duda, A., et al (1996). HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering. In *Proceedings of the 7th ACM conference on Hypertext* (pp. 180–193). Bethesda, Maryland, USA: ACM Press.

Zhang, B., Chen, Y., Fan, W., Fox, E. A., Goncalves, M. A., Cristo, M., et al (2005). Intelligent fusion of structural and citation-based evidence for text classification. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 667–668). New York, NY, USA: ACM Press.

Zhang, L., Glänzel, W., & Liang, L. M. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, in press, doi:10.1007/s11192-008-2245-y.