Hybrid Clustering for Validation and Improvement of Subject-Classification Schemes

Frizo Janssens[1,2,3], Lin Zhang[1,4] and Wolfgang Glänzel[1,5]
[1]*K.U. Leuven, Steunpunt O&O Indicatoren, Dept. MSI, Leuven*
[2]*Attentio SA/NV, StudioTROPE building, Bloemenstraat 32, B-1000 Brussels,*
[3]*K.U. Leuven, ESAT-SCD, Leuven*
[4]*WISE Lab, Dalian University of Technology, Dalian*
[5]*Hungarian Academy of Sciences, IRPS, Budapest*
[1,2,3]*Belgium*
[4]*China*
[5]*Hungary*

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). 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 (2008), and Boyack et al. (2008) used journal cross-citation matrices, while Moya-Anegon (2007) used subject co-citation analysis to visualise the structure of science and its dynamics.

Source: Data Mining and Knowledge Discovery in Real Life Applications, Book edited by: Julio Ponce and Adem Karahoca,
ISBN 978-3-902613-53-0, pp. 438, February 2009, I-Tech, Vienna, Austria

Earlier, a completely different approach was introduced by Callon et al., (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. Kostoff et al., 2001, 2005; Glenisson et al., 2005a,b).

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, 2001; 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 et al. (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 et al., 1991; combination of coupling and co-word analysis: Small (1998); hybrid coupling-lexical approach: Janssens et al., 2007b, 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 like 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 Cognitive mapping vs. subject classification

The objective of the present study is two-fold. 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, however, a methodological one, namely to evaluate improved methods of hybrid clustering techniques.

The 22-field subject classification scheme of the Essential Science Indicators (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". 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 8,305. Most of the subsequent analyses were performed in Java and MATLAB. We also made use of the MATLAB Tensor Toolbox (Bader, 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, 2004) and encoded in the Vector Space Model using the TF-IDF weighting scheme (Baeza-Yates, 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 a 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) (Deerwester, 1990; Berry, 1995), 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. Text-based similarities were calculated as the cosine of the angle between the vector representations of two papers (Salton, 1986).

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. 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 between two journals (or clusters, respectively), a normalised similarity can be calculated by dividing it by the square root of the product of the total number of citations to or from the first journal (cluster), and the total number of citations to or from the second. Intra-cluster 'self-citations' are counted only once.

For visualisation of the networks we use the similarities just described as edge weights between two clusters or fields (see Figure 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. That's why we opted to use "second order" journal cross-citation similarities for clustering. The journal cross-citation numbers are usually stored in a square, symmetric matrix. With "second-order similarities" we mean that the cross-citation values between a journal and all other journals (i.e., row or column of the matrix with cross-citation numbers) are used as input for another step of pairwise similarity calculation. The second-order similarities are found by calculating the cosine of the angle between pairs of vectors containing all symmetric journal cross-citation values between the two respective journals and all other journals. Hence, the ultimate similarity of two journals is based on their respective similarities with all other journals.

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

3.3 Clustering

In order to subdivide the journal set into clusters we used the agglomerative hierarchical cluster algorithm with Ward's method (Jain, 1988). It is a hard clustering algorithm, which means that each individual journal is assigned to exactly one 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 and the Rand index are used to compare the obtained results with an intellectual classification scheme.

3.3.2 Dendrogram

A preliminary judgment is offered by a dendrogram, which provides a visualisation of the distances between (sub-) clusters (see Figure 4 for an example). It shows the iterative grouping or splitting of clusters in a hierarchical tree. 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. Because of the difficulty to define the optimal cut-off point on a dendrogram (Jain, 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 for a document ranges from –1 to +1 and measures how similar it is to documents in its own cluster vs. documents in other clusters (Rousseeuw, 1987). 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 Figures 1 & 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 & Girvan, 2004; Newman, 2006). 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.

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 and Rand index

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. The Rand index is another external validation measure to quantify the correspondence between a clustering outcome and a ground-truth categorisation (Jain, 1988). In contrast to the Jaccard coefficient, the Rand index does take into account negative matches as well. Both measures result in a value between 0 and 1, with 1 indicating identical partitions. In Figure 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 et al., 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 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 links between journals to be considered 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 weighted linear combination of the corresponding distance 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 et al., 2005; Janssens et al., 2007a, 2007b, 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, 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).

Field #	ESI Field	Field #	ESI Field
1	Agricultural Sciences	12	Mathematics
2	Biology & Biochemistry	13	Microbiology
3	Chemistry	14	Molecular Biology & Genetics
4	Clinical Medicine	15	Multidisciplinary
5	Computer Science	16	Neuroscience & Behavior
6	Economics & Business	17	Pharmacology & Toxicology
7	Engineering	18	Physics
8	Environment/Ecology	19	Plant & Animal Science
9	Geosciences	20	Psychology/Psychiatry
10	Immunology	21	Social Sciences
11	Materials Sciences	22	Space Science

Table 1. The 22 broad science fields according to 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 five-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. Figure 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). Several fields seem not to be consistent enough from both perspectives. Above all, the Silhouette values of field #2 (Biology & Biochemistry), #4 (Clinical Medicine), #7 (Engineering), #19 (Plant & Animal Science) and #21 (Social Sciences) substantiate that at least five of the 22 fields are not sufficiently consistent.

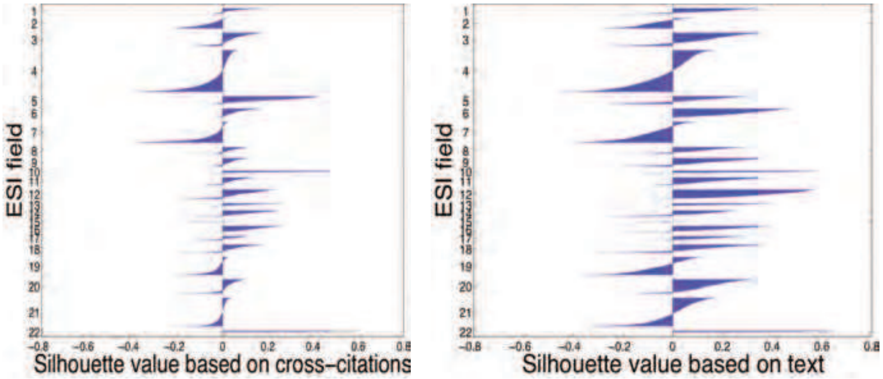

Fig. 1. Silhouette plot for 22 ESI fields based on journal cross-citations (left) and based on text (right)

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

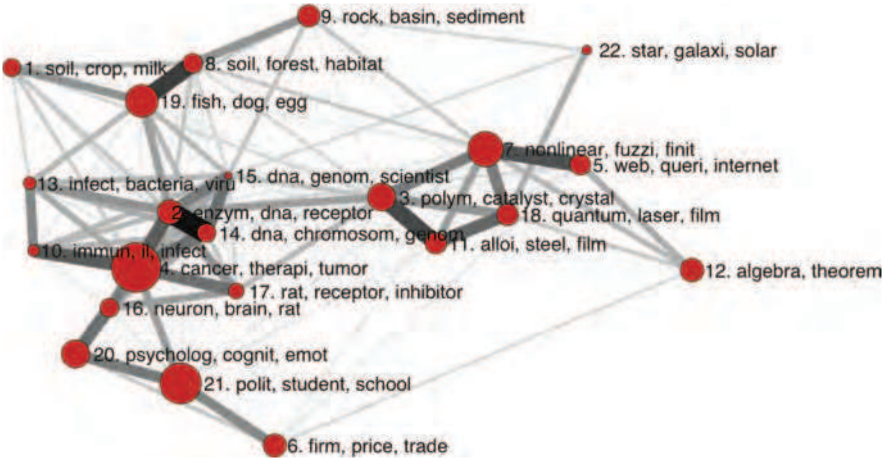

Fig. 2. Network of the 22 ESI fields based on cross-citation links

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 & Animal Science (#19) and Environment/Ecology (#8), as well as Biology & Biochemistry (#2), Molecular Biology & 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 Figure 2. For the visualisation we used Pajek (Batagelj & Mrvar, 2002). The network map confirms the strong links we have found based on the best terms between fields #2 & #14, #3 & #11, #5 & #7, and #8 & #19, respectively.

Field	Best 50 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; juic; nutrient; moistur; harvest; maiz; veget; cook; leaf; soybean; nitrogen; farmer; season; vitamin; potato; weed; textur; dairi; bacteria; fresh; corn;
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; secret; bone; recombin; mitochondri; coli; brain; tumor; ligand; liver; antibodi; subunit; ion; apoptosi; yeast; intracellular; vivo; cholesterol; biologi; affin; calcium;
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; amino; chiral; nm; ir; electrochem; salt; reactor; copper; chlorid; ionic; surfact; aromat; ni; h2o; fluoresc; column; chromatographi; alkyl; cl; alcohol;
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; recurr; malign; pulmonari; receptor; oral; men; therapeut; postop; ci; hiv; vascular; mutat; ct; hepat; infant; diagnos; tumour; pregnanc; antibodi; il;
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; custom; scalabl; bit; multicast; 3d; iter; java; ip; onlin; metric; platform; polynomi; retriev; neural; circuit; heurist; algebra; robot; topolog; broadcast;
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; game; worker; inflat; job; union; foreign; brand; earn; forecast; labour; reform; export; unemploy; insur; retail; volatil; team; credit; pai; financ;
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; compress; custom; wind; friction; actuat; concret; logic; soil; geometr; laser; graph; antenna; cylind; traffic; oscil; calibr; autom; geometri; grid; reactor;
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; nitrogen; fire; ha; emiss; urban; coastal; flood; biodivers; reproduct; basin; nest; pesticid; seedl; crop; dry; microbi; watersh; graze; winter; rainfal;
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; assemblag; veloc; sedimentari; crust; melt; marin; continent; magma; or; deform; east; flux; granit; belt; fractur; shallow; earth; slope; cloud; clai;

Field	Best 50 terms
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; pathogen; innat; assai; chemokin; dendrit; allergen; viral; igg; interleukin; monocyt; apoptosi; neutrophil; epitop; allerg; immunolog; secret; inflamm; dna; vitro; th2;
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; silicon; shear; bond; microscopi; fe; ion; wear; adhes; cu; copper; nanoparticl; lamin; nanotub; aluminum; compress; roll; elast; creep; atom; al2o3;
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; metric; integ; matric; lie; exponenti; markov; curvatur; norm; eigenvalu; kernel; hilbert; cohomolog; geometr; quadrat; covari; dirichlet; semigroup; iter; parabol; theta;
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; 16; soil; biofilm; antibodi; microorgan; fungal; amino; antigen; bacillu; recombin; fungi; albican; gram; mutat; phylogenet; mice; pseudomona; ferment; rrna; genotyp;
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; enzym; genotyp; vitro; assai; vivo; il; embryon; epitheli; recombin; pcr; chromatin; mammalian; regulatori; linkag; transgen; loci; delet; haplotyp; homolog; yeast;
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; insect; fluoresc; cancer; quantum; sediment; scienc; bone; viru; australia; immun; ecolog; fish; china; atmospher; your; mind; rat; bird; ic; colour;
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; deficit; syndrom; eeg; nervou; sensori; stimuli; dementia; ms; stimulu; glutam; muscl; nucleu; astrocyt; chronic; gaba; frontal; sclerosi; auditori; cord; alcohol;
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; kinas; lung; peptid; apoptosi; ca2; serum; administ; molecul; potent; chronic; insulin; mug; mum; liposom; p450; renal; hepat; inhibitori; immune; ligand;

Field	Best 50 terms
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; fiber; hole; superconduct; motion; spectral; dielectr; collis; coher; glass; semiconductor; neutrino; perturb; detector; algebra; elast; soliton; waveguid; relativist; amplitud; alloi;
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; seedl; prei; mate; shoot; cattl; southern; trait; genera; fed; island; nov; ecolog; lake; insect; pollen; viru; river; juvenil; farm; pathogen;
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; verbal; teacher; cue; stimuli; satisfact; judgment; job; infant; development; violenc; trait; ptsd; stimulu; style; interperson; peer; prime; esteem; distress; recal;
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; worker; agenc; teach; sexual; economi; incom; academ; immigr; sociolog; moral; african; skill; mental; librari; men; sector; land; crime; china; civil;
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; cosmolog; mission; motion; observatori; burst; spectra; photometr; gravit; comet; sun; bright; infrar; grb; shock; ngc; dark; supernova; spacecraft; radial; halo;

Table 2. The best 50 TF-IDF terms describing the 22 ESI fields

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

Figure 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, Figure 3(1) presents for various cluster numbers (2 to 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 Figure 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 again 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.

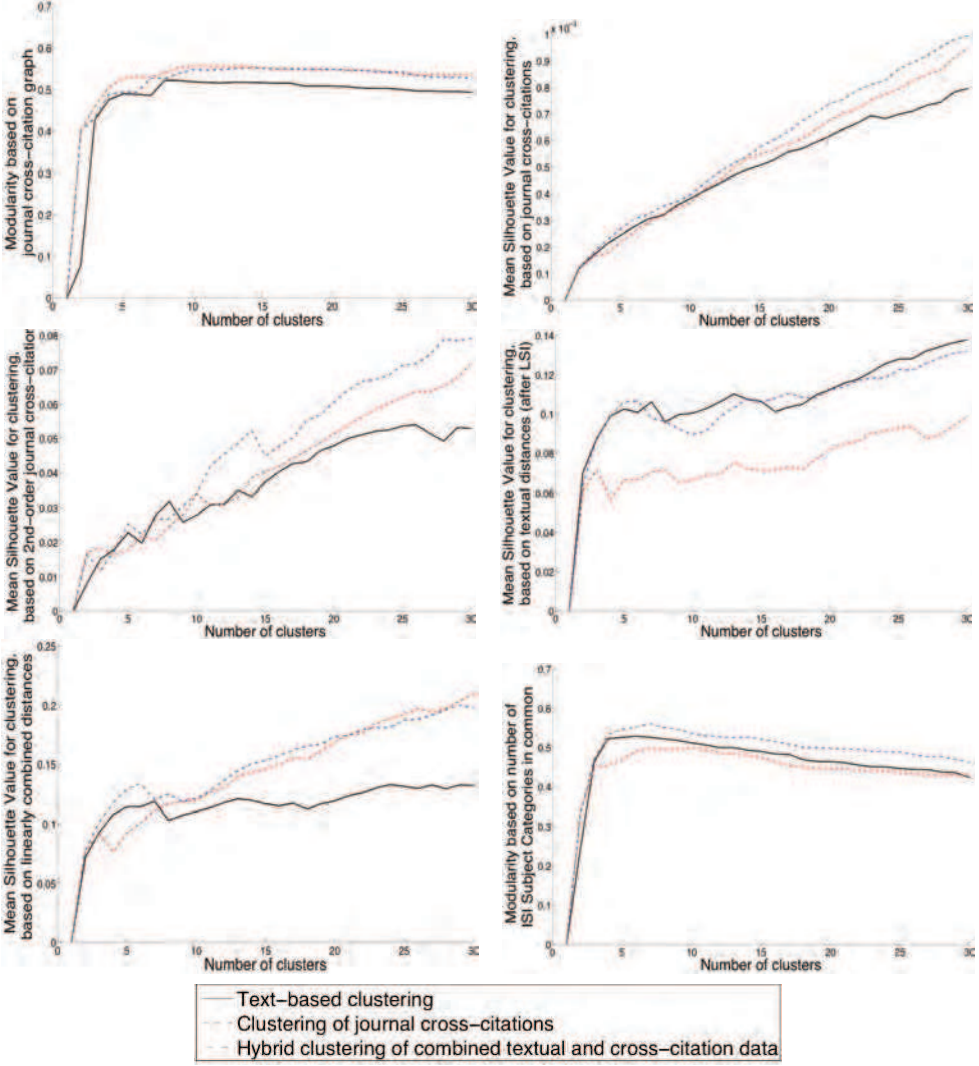

Fig. 3. Performance evaluation of text-based, citation-based and hybrid clustering based on (1) modularity calculated from the journal cross-citation graph, and based on Silhouette curves calculated from (2) journal cross-citations, (3) second-order journal cross-citations, (4) text-based distances, and (5) linearly combined distances. For an additional 'external validation' of clustering results compared to ISI Subject Categories, the lower-right figure (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).

Silhouette curves based on the complement of second-order cross-citations are shown in Figure 3(3). Again, the hybrid clustering almost always performs best.

In Figure 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.

Figure 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, Figure 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.

	Modularity based on journal cross-citation graph	Modularity based on common ISI Subject Categories	MSV based on textual distances	MSV based on 2nd order journal cross-citations	MSV based on linearly combined distances	Rand index with 22 ESI fields as reference classification
22 ESI fields	0.47533	**(0.52604)**	0.057237	0.016017	0.062807	(1)
22 citation-based clusters	**0.54676**	0.44244	0.09319	**0.057337**	**0.18938**	0.90463
22 text-based clusters	0.50451	0.45091	0.11829	0.035447	0.12987	0.90582
22 Hybrid clusters	**0.54677**	**0.48839**	**0.1206**	0.05453	**0.18951**	**0.90867**

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.

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 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 second order cross-citations. However, small the difference, the last column shows that the 22 hybrid clusters correspond best to the 22 ESI fields. It should be noted that the values in Table 3 can differ somewhat from the values in Figure 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.

4.4 Evaluation of hybrid clusters

The cluster dendrogram shows the structure in a hierarchical order (see Figure 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 Figure 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 Figure 4. A closer look at the best TF-IDF terms reveals that 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).

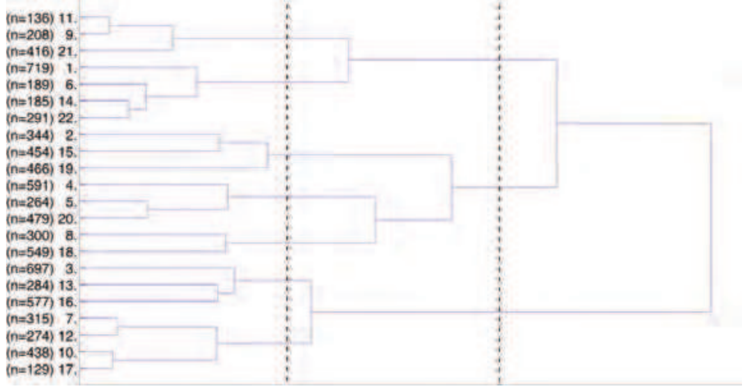

Fig. 4. Cluster dendrogram for hybrid hierarchical clustering of 8305 journals, cut off at 22 clusters on the left-hand side. Two other vertical lines indicate the cut-off points for 7 and 3 clusters.

The hybrid, i.e. the combined citation-textual based clustering yields acceptable results (see Figure 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.

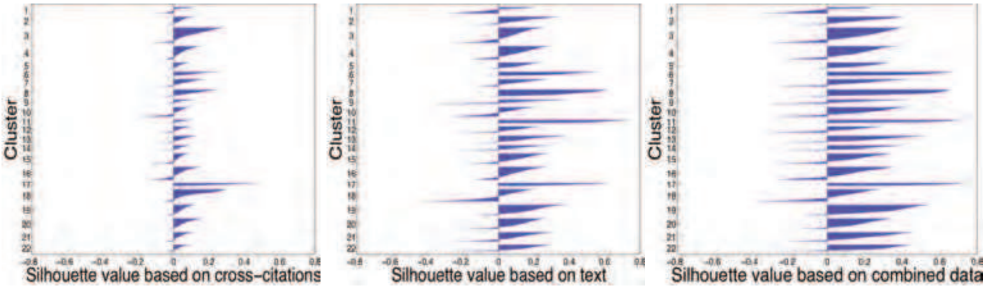

Fig. 5. Evaluation of the hybrid clustering solution with 22 clusters by citation based Silhouette plot (left), text based Silhouette plot (centre) and the plot with Silhouette values based on combined data (right).

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 Figure 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 50 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.

Cluster (# journals)	Best 50 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; narr; profession; sexual; parti; legal; incom; english; job; music; anxieti; invest; german; welfar; academ; belief; write; sector; violenc; religi; teach
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; polynomi; hydrogen; fiber; fault; machin; season; emiss; crack; fuzzi; shear; habitat; nois; steel; dry; plane; fe; catalyst; elast; sp; glass
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; inhibitor; renal; enzym; breast; surgic; lung; therapeut; mortal; vaccin; genom; transcript; nurs; assai; transplant; inflammatori; peptid; insulin; cardiac; carcinoma; oral

Table 4. Best 50 TF-IDF terms describing the 3 top-level clusters

According to the terms, 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 & environment are covered. Cluster #3, finally, covers biosciences, biomedical research, clinical & experimental medicine and neurosciences.

Cluster (# journals)	Best 50 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; essai; union; foreign; democraci; job; land; wage; civil; china; labour; book; narr; worker; democrat; german; school; liber; internet; text; religi
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; predat; marin; reproduct; nest; larva; bacteria; taxa; winter; cultivar; ha; nitrogen; ecosystem; seedl; eastern; ic; atmospher; flower; breed; wheat; bird
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; chromosom; bone; kinas; breast; vitro; chronic; muscl; mrna; therapeut; transplant; syndrom; insulin; dog; inflammatori; hepat; lesion; rna; pcr; diet; molecul
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; shear; powder; spin; elast; fabric; adsorpt; si; nonlinear; excit; sensor; fuel; fe; poli; polymer; diffract; emiss; aqueou; ni; nmr; corros
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; web; compact; epsilon; neural; integ; circuit; symmetr; ellipt; bar; fault; node; matric; geometr; markov; sigma; exponenti; queri; custom; wireless; video
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; profession; questionnair; suicid; violenc; classroom; verbal; youth; academ; peer; therapi; men; development; semant; stimuli; discours; linguist; phonolog; deficit; infant; offend
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; implant; cognit; ct; ey; cerebr; smoke; pregnanc; fractur; tumor; mri; cardiovascular; elderli; ci; motor; spinal; sleep; oral; questionnair; myocardi; vascular

Table 5. Best 50 TF-IDF terms describing the 7 top-level clusters