Hybrid clustering analysis for mapping large scientific domains

Abstract

A hybrid clustering method combining cross-citation and textual analysis is applied to cluster more than 8,000 journals covered in the Web of Science (2002-2006). Based on an agglomerative hierarchical clustering algorithm, all the journals understudy have been clustered into 7 large scientific domains, which is one of the optimal results obtained from the clustering process. The evaluation of the obtained clustering provides consistent results as considered from the cognitive perspective and the most characteristic terms, which are obtained from the textual component of the classification process, giving a clear description of each individual cluster. The cross-citation network visualises the citation relations among clusters and the asymmetric links reflect the direction of information flow among journals and clusters. A modified version of PageRank algorithm is applied to identify the representative journals in each cluster, and several indicators comprising *self-link rate*, *strong links* and *entropy* are used to compare the communication characteristics of each scientific domain.

Introduction

Hybrid clustering is a technique to integrate multiple information sources for clustering. Since a single data source normally provides somewhat one-sided information, combining several complementary information sources for clustering analysis is a promising and popular method nowadays (e.g., Modha & Spangler, 2000; He et al. 2001, 2002; Glenisson et al., 2005; Calado et al., 2006). In bibliometric studies, two basic approaches for clustering scientific units are in use, namely, the lexical approach and citation-based methods. Clustering based on only one of the two techniques might suffer from some shortcomings (e.g., Braam et al., 1991). In particular, textual similarities may obscure true document similarity by differences in vocabulary use, or introduce spurious similarities from polysemous words or words with little semantic value. On the other hand, citation based information alone might hide some similarities at the semantic level. Thus, the combination of the two methods can essentially improve the reliability and efficiency of the clustering algorithm (Janssens et al, 2008).

Along with the development of computerised scientometrics, to get 'the big picture' of a large scientific domain has allured many researches (e.g., Glänzel & Czerwon, 1996; Börner, 2003; Garfield, 1998; Moya-Anegon et al., 2004; Leydesdorff, 2004a, b). Unlike those studies, which are based on journal-citation information from the *Journal Citation Reports*, we adopted an item-by-item method to generate all the citation and text information from a large amount of journals. In a previous study, more than 8,000 journals published in *SCI*, *SSCI* & *AHCI* from *Web of Science* (2002-2006) were clustered into 22 groups based on a hybrid citation-text clustering algorithm, and were compared with the 22 classification scheme of ESI (Janssens et al., 2009). Although there is not any comparable subject classification scheme as a reference system, seven is one of the optimal number of clusters obtained from our clustering process and yields the best evaluation results. In addition, we can relate the results of the seven-cluster solution to those of the 22-cluster based study. In particular, we can analyse the effect of the splitting process of the seven larger clusters split into the 22 smaller ones according to the solution presented by Janssens et al. (2009).

Data

The data has been collected from the *Web of Science* of Thomson Scientific. Altogether 9527 journals which were covered in the entire period of 2002-2006 were selected and only four document types, namely, article, note, letter and review, were taken into consideration. More than six million papers were indexed and citations have been summed up through a variable citation window, from the publication year till 2006. All terms from titles, abstracts and keywords were extracted and used as the input of the textual analysis for the clustering process, as well as for "labelling" the obtained clusters.

Journals were checked for name changes, merging or splitting during the period under study, and finally were identified accordingly. In order to reduce noise caused by those too 'small' and/or 'poorly' cited journals, as well as to build meaningful and statistically reliable indictors, only journals that had at least 50 publications and of which the sum of their references and citations was at least 30 during 2002-2006 were taken into account. The resulting number of remaining journals was 8305. All citation data and extracted terms related to these journals were used as input for the similarity matrices used for the clustering analysis.

Methods

In contrast to other approaches (e.g. Leydesdorff 2004a, 2004b, 2006) our study is not based on Thomson Reuters' Journal Citation Reports (JCR), but on an item-by-item basis from the WoS. This approach provides us several advantages, for instance, the exclusions of not relevant documents (see *data* section), the allowance of pre-defined publication periods and citation windows and et al. Among others, the possibility of combination of citation-based methods with text mining is the main point of our hybrid clustering in this study.

The citation-text hybrid clustering was based on the agglomerative hierarchical clustering algorithm with Ward's method (Jain, 1988). This is a hard clustering algorithm, i.e., each individual journal is assigned to exactly one cluster.

As a first step, two distance matrices were built according to the similarities between journals, respectively based on the cross-citations and the texts. In both matrices, normalised similarities were calculated as the cosine of the angle between the vector representations of two journals (Salton, 1986). The integration of the lexical information and the citation information was achieved through equally weighted linear combination of the corresponding distance matrices before the clustering algorithm was applied. To determine the optimum number of clusters, as well as for the evaluation of clustering results, we have used Silhouette and modularity values.

Besides the clustering analysis, several indicators including a modified version of Google's *PageRank* algorithm (Brin & Page, 1998), journal *strong links* and *citation entropies* (Zhang et.al, 2009) could help to understand the cluster structure and to grasp the representative journals in each cluster.

Results

Clustering data into journal groupings and making evaluations of clustering results

Determination of the optimum number of clusters is a rather difficult task as it depends on different factors including the adopted validation, the chosen similarity measures, data representation and others. Generally, the number of clusters is determined by comparing the quality of different clustering solutions based on various numbers of clusters. In this study, a preliminary judgment is offered by a dendrogram, which presents the iterative grouping or splitting of clusters in a hierarchical tree. Intuitively, we could find several appropriate cut-off points: a first one at three clusters, a second one at seven clusters, and another acceptable number at 22 (See Figure 1). The 3-cluster scheme results in three "mega-clusters", which should comprise life sciences, natural and applied sciences, and social sciences and humanities. This classification can be considered a "trivial solution" of no practical use. In terms of the 22 clusters solution, we refer to the above-mentioned study by Janssens et al., (2009). Besides, seven clusters seem to provide another interesting case for a deeper analysis. Figure 2 shows the hierarchical architecture of 7 clusters solution.

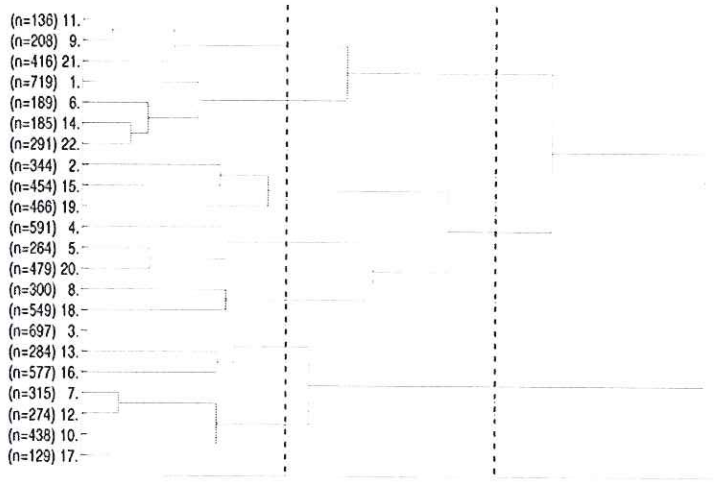

Figure 1. Cluster dendrogram for hybrid hierarchical clustering of 8305 journals, cut off at 22 clusters on the left. Two other vertical lines indicate the cut-off points for 7 and 3 clusters.

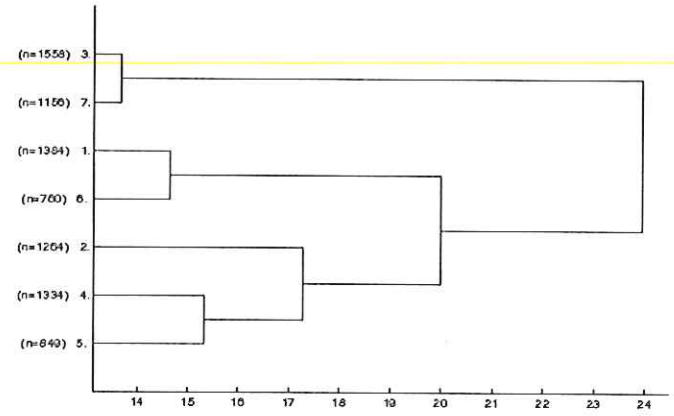

Figure 2. Cluster dendrogram for hybrid hierarchical clustering of 8305 journals, cut off at 7 clusters.

In consideration of the difficulty to define the optimal cult-off point on the dendrogram, we complement this method with numerical techniques. The quality of a clustering can be evaluated by calculating the modularity of the corresponding partition of the cross-journal citation graph (Newman & Girvan, 2004; Newman, 2006). As an 'external validation' of clustering results, we used modularity curves to compare the performance of different clustering solutions. In particular, the modularity was computed from a network containing all journals as nodes, but with edge weights equal to the number of ISI subject categories (in total 254 categories) commonly assigned to both journals by Thomson Scientific. Figure 3 presents the performance evaluation of the hybrid clustering based on the modularity, where we could clearly observe that seven is the optimal number for the whole curve.

Another appraisal for the performance of clustering is given by the plots with *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 sorted Silhouette values of all journals in each cluster are indicated with horizontal lines in Figure 4. According to the Silhouette plots, all clusters are quite well formed, or at least acceptable.

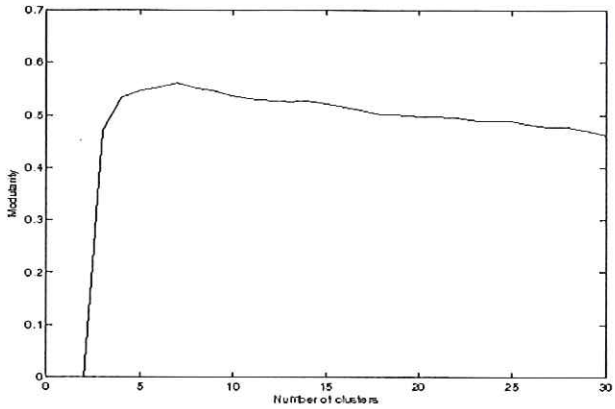

Figure 3. Performance evaluation of hybrid clustering of 8305 journals based on modularity values

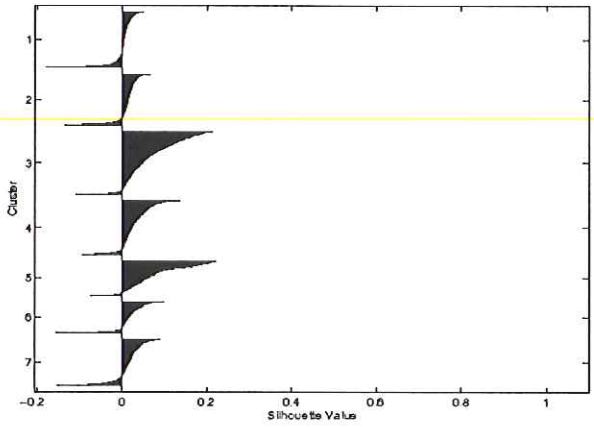

Figure 4. Visual evaluation of the hybrid clustering solution with 7 clusters by the plot with Silhouette values based on citation-text combined data

Labeling clusters using most cognitive characteristics

In order to give a description of the cognitive characteristics of individual clusters, we labelled each cluster using the best TF-IDF terms based on our textual approach. As mentioned in the outset, these terms were extracted from the title, abstract and keywords from all the documents in each cluster. The top 50 terms describing each cluster are presented in Table 1, which represent the cognitive characteristics of the clusters. Through the most related terms, the basical structure of the seven clusters is easily recognisable; three of these clusters correspond to the natural and applied sciences, two clusters represent life sciences and another two classes stand for social science and humanities. A closer look at the best characteristic terms reveals that (i) the three natural and applied sciences clusters comprise biology, agriculture and environmental sciences (#2), physics, chemistry and engineering (#4), mathematics and computer science (#5); (ii) two life-science clusters are respectively cluster #3 (biosciences and biomedical research) and #7 (clinical, experimental medicine and neurosciences); and finally, (iii) the social science clusters consist of cluster #1 (economics, business and political science) and #6 (psychology, sociology, education).

Table 1. The best 50 TF-IDF terms describing the 7 hybrid clusters

Cluster #	50 best terms
1	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	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	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	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	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	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	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;

Studying the cognitive structure based on hybrid cluster analysis

Studying the hierarchical and network structure of hybrid clusters

In the hierarchical architecture presented in Figure 1, each of the seven clusters splits into several sub-clusters if the cut-off line is given at 22 clusters. One of the social science clusters (#6, psychology, sociology, education) breaks up in three sub-clusters: a cluster comprising sociology and education, and two clusters both represent psychology but with different focuses, namely cognitive and medical issues. Four sub-clusters make up another social science cluster (#1, economics, business and political science), but these sub-classifications are not that straightforward. One sub-cluster (#6 in the 22 clusters scheme) covers economics, business and political science, while other three sub-classifications have rather fuzzy structures with heterogeneous profiles related to 'social sciences and humanities'. The sub-schemes of the three natural and applied science clusters are more clearly recognizable. Cluster #2 (biology, agriculture and environmental sciences) splits up to an environmental science cluster, a geosciences cluster and a cluster reflecting biology and agriculture. The other two science clusters are even more easily divisible. The cluster of 'physics, chemistry and engineering' (#4) just separates into three clusters respectively covering physics, chemistry and engineering. Similarly, the division of cluster #5 (mathematics and computer science) indeed presents two sub-classifications: mathematics, and computer science.

The interpretation of the sub-categories of the life sciences is somewhat complicated. The biosciences and biomedical cluster (#3) corresponds to three sub-clusters, one is for bioscience, other two mainly reflect microbiology & veterinary science, and hematology & oncology. On the other hand, cluster #7 (clinical, experimental medicine and neurosciences) splits into four smaller clusters, respectively, a clinical and social medicine cluster, one cluster standing for neuro- and behavioral sciences, another cluster focusing on non-internal medicine, and finally one classification mainly with cardiovascular and respiratory medicine issues.

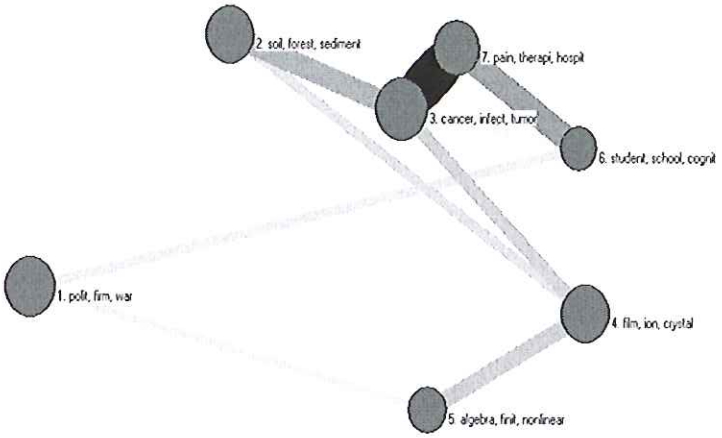

Figure 5. Network structure of 7 hybrid clusters represented by the three most relevant terms

The cross-citation network structure among different clusters based on the 7 clusters solution is visualized by *Pajek* (Batagelj & Mrvar, 2002) in Figure 5. For measuring the strength of citation links among two clusters *i* and *j*, a normalised similarity is calculated from the symmetrised raw number of cross-citaton C*ij* as:

$$\frac{C_{ij}}{\sqrt{\sum_k C_{ik} \cdot \sum_k C_{jk}}} \quad (1)$$

where the intra-cluster 'self-citations' are counted only once. The thickness of lines is proportional to the citation link strength between each two clusters and the size of the nodes is set proportional to the number of journals in a given cluster. All clusters are represented by the three most relevant TF-IDF terms. Among the whole network, the strongest link between cluster #3 (biosciences and biomedical research) and #7 (clinical, experimental medicine and neurosciences) is quite striking, which is also in accordance with the dendrogram architecture. But compared to the dendrogram, we indeed observe some different structures in the cross-citation network: one of the social science clusters (#6, psychology, sociology, education) dissociates from the other social science cluster #1 (economics, business and political science). Instead, it has a strong link with cluster #7 (clinical, experimental medicine and neurosciences). This phenomenon is unexpected, however, quite plausible; there is always strong affinity between psychology science and neurosciences, representing the 'social' and 'clinical' aspects of psychology, behavioural and neuroscience. Among the three natural and applied sciences clusters, two clusters ("#4 physics, chemistry and engineering" and "#5 mathematics and computer science") are getting together, whereas the third one, cluster #2 (biology, agriculture and environmental sciences) has a closer link with a life science cluster (#3, biosciences and biomedical research).

For the detection of relationship among journals, topics or subjects, preferably symmetric links are used since inequality of directions or even the total lack of one direction in immaterial in this context. Therefore we have ignored the direction of citations by symmetrising the cluster cross-citation matrix in the above network structure. But actually, the information exchange between each two clusters is asymmetric. In order to measure how information flows among different clusters, namely, whether there is clear direction of communication, we first applied an indicator measuring the strength of citation links between two individual clusters in the asymmetric matrix, see Formula 2.

$$CL_{ij} = \frac{a_{ij}}{\sqrt{TC_i \cdot TR_j}} \quad (2)$$

where i and j denote clusters, TC_i the total number of citations of cluster i, TR_j the total number of references of cluster j, and a_{ij} the number of citations of cluster i receives from cluster j.

Then we calculated $CL_{ij} - CL_{ji}$ between each two clusters i and j. According to our result, there are four pairs of clusters having relatively big values of $CL_{ij} - CL_{ji}$ (at least 0.01, present in Figure 6). It's interesting that the distribution of these asymmetric links is quite concentrated: only among four clusters (#2, #3, #4 and #7), where cluster #3 behaves as a core. In particular, Figure 6 visualizes the following phenomenon: the top two asymmetric links both originate from cluster #3 (biosciences and biomedical research), respectively to cluster #2 (biology, agriculture and environmental sciences), and to cluster #4 (physics, chemistry and engineering), namely, information flows from cluster #3 to the other two clusters. Cluster #7 (clinical, experimental medicine and neurosciences) originates information to cluster #3, while cluster #4 tends to be the acceptor of information (both from cluster #2 and #3). It seems that the information tends to flow from the life sciences towards the natural and applied sciences.

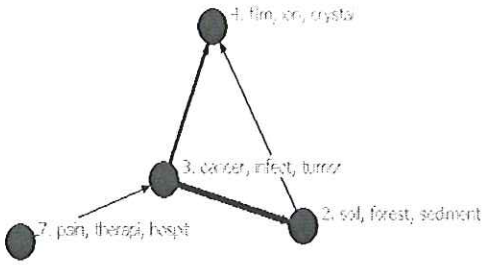

Figure 6. Direction of information flow among different clusters. Visualisation by Netdraw (Borgatti, 2002)

Detecting "important" journals of each cluster among the cross-citation network

Pinski et al. (1976) proposed an indicator of 'journal influence' by taking into account not simply the number of citations from one journal to the other, but also the influence of the citing journal. Similarly to the *influence weight*, Google's PageRank algorithm (Brin & Page, 1998) computes the status of a Web page based on a combination of the number of hyperlinks that point to the page and the status of the pages that the hyperlinks originate from. Based on a modified PageRank algorithm (see Formula 3), we ranked all the journals in each cluster according to their values of PageRanking. Journals present in Table 1 are the top ten nodes with highest PageRanking in each cluster, which are those having most frequent contacts with prominent journals of the same cluster. To a certain degree, these journals can represent their clusters in an adequate manner.

$$PR_i = \frac{(1-\alpha)}{n} + \alpha \sum_j PR_j \frac{a_{ij}/P_i}{\sum_k \frac{a_{kj}}{P_k}} \quad (3)$$

where PR_i is the PageRank of journal i, α is a scalar between 0 and 1 (α=0.9 in our implementation), n is the number of journals in the cluster, a_{ij} 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.

Table 2. The ten most important journals of each cluster according to a modified version of Google's PageRank algorithm (see Formula 3)

Cluster 1	Cluster 2	Cluster 3	Cluster 4
1. Q J ECON	1. ANNU REV PLANT BIOL	1. ANNU REV IMMUNOL	1. REV MOD PHYS
2. J ECON LIT	2. ANNU REV ECOL EVOL S	2. NAT REV IMMUNOL	2. MAT SCI ENG R
3. J FINANC	3. PLANT CELL	3. NAT IMMUNOL	3. CHEM REV
4. J POLIT ECON	4. ANNU REV PHYTOPATHOL	4. NAT REV CANCER	4. PROG POLYM SCI
5. J FINANC ECON	5. CURR OPIN PLANT BIOL	5. CELL	5. PHYS REP
6. AM POLIT SCI REV	6. ANNU REV ENTOMOL	6. CA-CANCER J CLIN	6. ANNU REV FLUID MECH
7. REV ECON STUD	7. SYSTEMATIC BIOL	7. ANNU REV BIOCHEM	7. ANN REV MATER RES
8. ECONOMETRICA	8. ANNU REV ASTRON ASTR	8. NAT REV MOL CELL BIO	8. PROG ENERG COMBUST
9. REV FINANC STUD	9. OCEANOGR MAR BIOL	9. IMMUNITY	9. SURF SCI REP
10. ADMIN SCI QUART	10. MICROBIOL MOL BIOL R	10. NAT GENET	10. ANNU REV NUCL PART S
Cluster 5	**Cluster 6**	**Cluster 7**	
1. ACM COMPUT SURV	1. ANNU REV PSYCHOL	1. ANNU REV NEUROSCI	
2. J AM MATH SOC	2. PSYCHOL BULL	2. NAT REV NEUROSCI	
3. ACTA MATH-DJURSHOLM	3. PSYCHOL REV	3. NEURON	
4. ANN MATH	4. REV EDUC RES	4. NAT NEUROSCI	
5. J ACM	5. BEHAV BRAIN SCI	5. CIRCULATION	
6. INVENT MATH	6. AM EDUC RES J	6. MILBANK Q	
7. J ROY STAT SOC B	7. J PERS SOC PSYCHOL	7. JAMA-J AM MED ASSOC	
8. COMMUN PUR APPL MATH	8. COGNITIVE PSYCHOL	8. PROG NEUROBIOL	
9. B AM MATH SOC	9. ARCH GEN PSYCHIAT	9. CURR OPIN NEUROBIOL	
10. FOUND COMPUT MATH	10. J EXP PSYCHOL GEN	10. TRENDS NEUROSCI	

Comparison of clusters based on different indicators

According to our methods, if journals are clustered into one group, it means that they have relatively high similarities with regard to the combined information of cross-citation and text. The question arises whether there are other "common characteristics" of journals in the same cluster and what the differences are between different clusters considering these communication characteristics.

As a first indicator, journal "*Self-Link Rate*" (SLR) is normally considered a measure of "isolation". Diametrically opposing the indicator of isolation, *Entropy* describes how far citations are spread among other journals in the communication network, see Formula 4.

$$EL_i = -\sum_{j=1}^{n} \frac{a_{ij} + a_{ji}}{TC_i + TR_i} \cdot \log_2 \left(\frac{a_{ij} + a_{ji}}{TC_i + TR_i} \right) \quad (4)$$

where i and j denote journals, TC_i the total number of citations of journal i, TR_i the total number of references of journal i and a_{ij} the number of citations of journal i receives from journal j.

Based on these definitions, we calculated the *Entropy* and *SLR* for all the individual journals, and present the average values of each cluster in Table 3. The two life science clusters, #3 (biosciences and biomedical research) and #7 (clinical, experimental medicine and neurosciences) stand out from other clusters due to their high average *Entropies*. On the other hand, the social science cluster #1 (economics, business and political science) is found at the end of the rank. However, the case of the average *SLR* is just on the contrary: cluster #1 ranks highest according to *SLR*, while the previous conquerors, cluster #3 and #7 are among the lowest ones. This phenomenon is interesting but not against our expectation: high self-link rate is the obstacle for journals spreading information over and/or collecting information from a variety of other entities. In terms of the categories, life science journals, especially journals related to biosciences and biomedical research (cluster #3) incline to extend their communication links into a far-ranging scope, while journals with regard to social sciences (economics, business and political science) are relatively isolated.

Table 3. Comparison of clusters based on average Self-link rate and Entropy

Cluster #	Average Entropy	Average SLR	Total number of journals
3	7.32	0.08	1558
7	6.88	0.11	1156
6	5.98	0.17	760
2	5.94	0.16	1264
4	5.92	0.15	1334
5	5.64	0.13	849
1	5.15	0.20	1384

Besides *entropy*, another indicator reflecting the centrality of journals within each cluster was also calculated. This indicator is based on the high frequency of strong links with other journals. In particular, the strength of links was calculated by the raw number of cross-citations between two journals divided by the square root of the product of the total number of citations to or from the first journal, and the total number of citations to or from the second one. Based on the above definition, when the journal link strength is over 0.025, it's called *strong links*; and when an individual journal has at least 5 *strong links*, it could be considered as a "central journal". While *entropy* merely measures the diversity of contacts in a social network regardless of their intensity, the second approach reflects a strong influence from/on other representatives.

Table 4. Comparison of clusters based on the share of "central journals"

Cluster #	Share of "central journals"	Total number of journals
6	0.23	760
5	0.18	849
2	0.15	1264
4	0.14	1334
7	0.12	1156
1	0.10	1384
3	0.08	1558

The shares of 'central journals' within each cluster are present in Table 4. There are over one fifth journals acting as central nodes in cluster #6 (psychology, sociology, education), which means that lots of coherent sub-clusters exist in this domain, with these 'central journals' as their 'cores'. As an opponent, only 8% 'central nodes' are found in cluster #3 (biosciences and biomedical research). Combined with the analysis of 'average entropy' where cluster #3 acts as the leader, we could conclude that journals of biosciences and biomedical sciences tend to have far-ranging links, but these relationships are rather loose. The 'mathematics and computer science' cluster (#5) serves as an example with the opposite characteristics, low entropy but high centrality. It's also interesting to note that there is clear diversity between the two social science clusters (#6 and #1): basically, the 'psychology, sociology, education' cluster behaves well in both *entropy* and *centrality*, but journals of 'economics, business and political science' are short of both 'decentralisation' and 'intensity' of contacts in the communication network.

Conclusions

The hybrid clustering combining cross-citation and textual information generates an excellent and well-balanced map of scientific journals based on seven large domains; two of life sciences, three covering natural and applied sciences and another two related to social sciences. The number of journals in each cluster ranges between about 850 and 1,550. The

evaluation provides good result and the most characteristic terms give a clear description of each individual cluster. The analysis of the dendrogram shows how the 7 clusters break down into 22 sub-clusters, and also helps to understand the intra-structure of each individual clusters. The cross-citation network visualised the relationship among different clusters, where we could observe the distinct affinity between the two life science clusters. Although the life sciences have most apparent links with other subjects, some of the communication links are rather unidirectional. In accordance with our expectation, there are different communication characteristics within different scientific domains. Journals related to biosciences and biomedical sciences tend to have far-ranging but rather loose links. On the other hand, the high percentage (over 20%) of 'central nodes' among the 'psychology, sociology, education' journals was somewhat unexpected.

References

Batagelj, V. & Mrvar, A. (2002). Pajek – analysis and visualization of large networks. *Graph Drawing*, 2265, 477–478, ISSN.

Borgatti, S.P. (2002), *Netdraw network visualization*. Analytic Technologies: Harvard, MA.

Börner, K., Chen, C. & Boyack, K.W. (2003). Visualizing knowledge domains. Annual Review of Information Science and Technology, 37, 179-255.

Braam, R., Moed, H., & Van Raan, A. (1991). Mapping of science by combined co-citation and word analysis. 2. Dynamical aspects. *Journal of the American Society for Information Science*, 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. *Journal of the American Society for Information Science*, 57, 2, 208–221.

Garfield, E. (1998). *Mapping the world of science*. The 150 Anniversary Meeting of the AAAS, Philadelphia, PA, http://www.garfield.library.upenn.edu/papers/mapsciworld.html

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

He, X.; Ding, C.H.Q.; Zha, H. & Simon, H.D. (2001). Automatic topic identification using webpage clustering. *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.

Jain, A. & Dubes, R. (1988). *Algorithms for Clustering Data*. Prentice Hall, New Jersey

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

Janssens, F., Lin, Z. & Glänzel, W. Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, 2009, to be published.

Leydesdorff, L. (2004a). Clusters and maps of science journals based on bi-connected graphs in the Journal Citation Reports. *Journal of Documentation*, 60, 4, 371-427.

Leydesdorff, L. (2004b). Top-down decomposition of the Journal Citation Report of the Social Science Citation Index: Graph- and factor-analytical approaches. *Scientometrics*, 60, 2, 159-180.

Leydesdorff, L. (2006). Can scientific journals be classified in terms of aggregated journal-journal citation relations using the Journal Citation Reports? *Journal of the American Society for Information Science & Technology*, 57, 5, 601-613.

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

Moya-Anegon F, Vargas-Quesada B, Herrero-Solana V, et al. (2004). A new technique for building maps of large scientific domains based on the cocitation of classes and categories. *Scientometrics*, 61, 1, 129-145.

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

Newman, M.E.J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America,* 103, 23, 8577-8582.

Pinski, G. & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: theory, with application to the literature of physics. *Information Processing and 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*, McGraw-Hill, Inc, New York.

Zhang, L., Glänzel, W. & Liang, L.M. Tracing the role of individual journals in a cross-citation network based on different indicators, *Scientometrics*, forthcoming.