Journal of Informetrics 4 (2010) 185–193

Contents lists available at ScienceDirect

Journal of Informetrics

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

Subject clustering analysis based on ISI category classification

Lin Zhang[a,b], Xinhai Liu[c,d], Frizo Janssens[b,c], Liming Liang[a,e], Wolfgang Glänzel[b,f,*]

[a] Dalian University of Technology, WISE Lab, Dalian, China
[b] K.U. Leuven, Centre for R&D Monitoring, Department of MSI, Leuven, Belgium
[c] K.U. Leuven, ESAT-SCD, Leuven, Belgium
[d] Wuhan University of Science and Technology, CISE & ERCMAMT, Wuhan, China
[e] Henan Normal University, Institute of Science Technology & Society, Xinxiang, China
[f] Hungarian Academy of Sciences, IRPS, Budapest, Hungary

ARTICLE INFO

Article history:
Received 14 August 2009
Received in revised form 4 November 2009
Accepted 9 November 2009

Keywords:
Journal classification
Cross-citation network
Multi-level Aggregation Method

ABSTRACT

The study focuses on the analysis of the information flow among the ISI subject categories and aims at finding an appropriate field structure of the Web of Science using the subject clustering algorithm developed in previous studies. The clustering journals and ISI subject categories provide two subject classification schemes through different perspectives and levels. The two clustering results have been compared and their accordance and divergence have been analyzed. Several indicators have been used to compare the communication characteristics among different ISI subject categories. The neighbour map of each category clearly reflects the affinities between the "core" category and its satellites around.

© 2009 Elsevier Ltd. All rights reserved.

1. Introduction

A series of previous studies focused on the analyses of journal clustering based on a complete journal–journal cross-citation matrix (Janssens, Zhang, & Glänzel, 2009; Zhang, Glänzel, & Liang, 2009; Zhang, Janssens, Liang, & Glänzel, 2009). ISI, now part of Thomson Scientific, has assigned each journal included to one or more subject categories. Based on this classification scheme, the journal–journal matrix can be aggregated to a category–category matrix, which is much more densely populated than that on the journal level. The present study will focus on the analysis of the information flow among the ISI subject categories. This will be done for two important reasons. This exercise aims at finding an appropriate field structure of the Web of Science using the subject clustering algorithm developed in previous studies. Furthermore, since ISI subject categories are based on journal assignment the question arises of what changes if journal cross-citation is replaced by subject cross-citation. If changes are not essential, the elaborate clustering of more than 8000 journals could be substituted by a somewhat easier analysis of roughly 250 ISI categories and the journal level could, as it were, be skipped. However, we stress that cross-citations are calculated from individual paper-to-paper links whatever aggregation levels are chosen. The other reason is to analyze whether multiple journal assignment to subject categories interferes with, distorts or even determines the resulting cluster structure. Before we introduce the methodological rudiments, we briefly summarise the historical background and the outcomes of previous or related studies.

Along with the development of computerised scientometrics, mapping of science plays an important role in the construction and analysis of science structure. For instance, a variety of techniques for analyzing journal–journal citation relationships

* Corresponding author at: K.U. Leuven, Centre for R&D Monitoring, Department of MSI, Leuven, Belgium.
E-mail addresses: linzhang1117@gmail.com (L. Zhang), xinhai.liu@esat.kuleuven.be (X. Liu), Frizo.Janssens@esat.kuleuven.be (F. Janssens), liangliming1949@sina.com (L. Liang), Wolfgang.Glanzel@econ.kuleuven.be (W. Glänzel).

1751-1577/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2009.11.005

have been reported in the literature (Doreian & Fararo, 1985; Leydesdorff, 2006; Tijssen, De Leeuw, & Van Raan, 1987). An alternative method of co-citation clustering has been investigated in constructing a World Atlas of Sciences for ISI (Garfield, Malin, & Small, 1975; Leydesdorff, 1987; Small, 1999). Boyack, Klavans, and Börner (2005) applied eight alternative measures of journal similarity to a dataset of 7121 journals covering over one million documents in the combined Science Citation and Social Sciences Citation Indexes, to show a global map of science using the force-directed graph layout tool VxOrd. Chen (2008) proposes an approach to classify scientific networks in terms of aggregated journal-journal citation relations of the ISI Journal Citation Reports using the affinity propagation method. As mentioned in the outset, Zhang, Glänzel, et al. (2009) and Zhang, Janssens, et al. (2009) have also investigated different methods for the analysis and classifications of scientific journals. Besides using journals as the units of analysis, some recent investigations focus on the science structure based on the subject categories. Glänzel and Schubert (2003) designed a new classification scheme of science fields and subfields for scientometric evaluation purposes. Moya-Anegon et al. (2004) proposed a new technique that uses thematic classification as entities of co-citation, and presented an ego-centred network of 222 ISI categories including science and social sciences. Leydesdorff and Rafols (2009) classified the ISI 172 science categories into 14 groups based on factor analysis, and compared the interdisciplinarity of each category using betweenness centrality. Compared to other researchers, we applied a new clustering technique to classify the ISI science and social sciences categories into 7 groups based on the category–category cross-citation similarities, and further compared the results with the 7 hybrid clustering solution of 8305 journals in a previous study (Zhang, Janssens, et al., 2009). Furthermore, several indicators have been used to analyze the communication characteristics of different categories.

2. Data sources and processing

The data have been collected from the *Web of Science* of Thomson-Reuters. Altogether 9487 journals which were assigned to the 246 categories of sciences, social sciences and arts and humanities in the entire period of 2002–2006 were selected and only three document types, namely, article, 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.

3. Methods

As already mentioned at the outset, citation links are determined on the basis of paper-by-paper assignment, which provides us several advantages compared to other approaches (Zhang, Glänzel, et al., 2009). There are three procedures for the cross-citation data aggregations: document-to-document, then journal-to-journal, and finally ISI category-to-category (or large domains-to-domains). The previous work focussed on the journal level, and now our focus turns to the higher level of ISI subject categories.

The clustering method adopted in this study is the *Multi-level Aggregation Method* (*MAM*)[1] (Blondel, Guillaume, Lambiotte, & Lefebvre, 2008), which is a new clustering algorithm based on the modularity optimization. Modularity (Newman, 2006) is a benefit function used in the analysis of networks or graphs such as computer networks or social networks. It quantifies the quality of a division of a network into modules or communities. Good divisions, having high modularity values, are those in which there are dense internal connections between the nodes within modules but only sparse connections between different modules. In particular, given a network composed of n nodes or vertices connected by m links or edges and let A_{ij} be an element of the adjacency matrix of the network, which gives the number of edges between vertices i and j. And suppose we are given a candidate division of the vertices into some number of groups. The modularity of this division is defined to be the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random. In the most common version of the concept, the randomization of the edges is done so as to preserve the degree of each node. In this case, the expected number of edges between two vertices i and j following randomization is $k_i k_j/2m$ where k_i is the degree of node i, and hence the actual number of edges minus the expected number of edges between the same two vertices is $A_{ij} - k_i k_j/2m$. Summing over all pairs of vertices in the same group, the modularity, denoted Q, is then given by:

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j) \tag{1}$$

where c_i is the group to which node i belongs and $\delta(c_i,c_j)$ is the *Kronecker delta* with $\delta(c_i,c_j) = 1$, if $c_i = c_j$, namely, nodes i and j both belong to the same group, and $\delta(c_i,c_j) = 0$, otherwise.

The value of the modularity lies in the range $[-1,1]$. It is positive if the number of edges within groups exceeds the number expected on the basis of chance.

[1] Now *Multi-level Aggregation Method* (MAM) is renamed as Louvain method since it was created in the University of Louvain, Belgium.

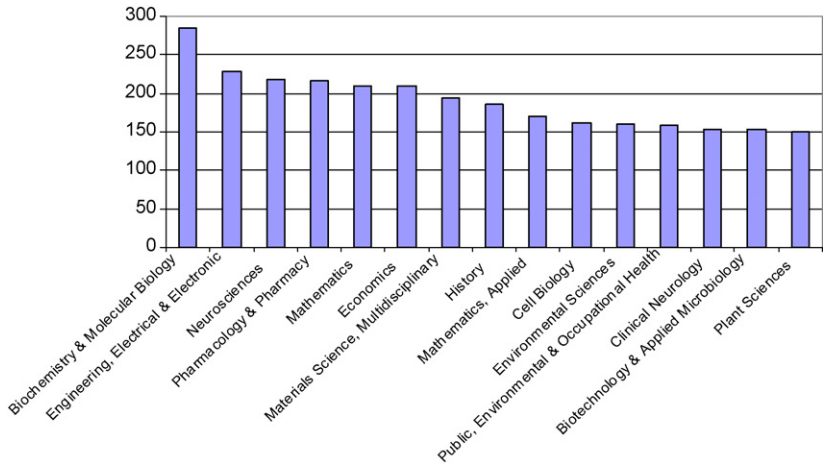


Fig. 1. 15 ISI categories which have more than 150 journals each (2002–2006).

Basically, MAM is implemented by simplifying Eqs. (1) and (2) to compute the gain in modularity $\Delta Q(i,j)$ between two nodes i and j in each step.

$$\Delta Q(i,j) = \begin{cases} \dfrac{A_{ij}}{2m} - \dfrac{K_i K_j}{(2m)^2}, & \text{if } i \text{ and } j \text{ are connected} \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

In MAM, firstly each node of the network is assigned to a single community. The two nodes i and j are merged on the basis of the maximum modularity gain $\Delta Q(i,j)$ defined in Eq. (2). The merging process is repeated until local maximum modularity is reached. Then the current communities are employed to form super nodes to repeat the above merging process of nodes. This process is applied repeatedly and sequentially for all nodes until no further improvement can be achieved.

When a local modularity maximum value is reached during the optimization, it will correspond to a cluster number from the formed communities. Since generally there are several local modularity maximum values available during the optimization stage, these various cluster numbers under such modularity values can be regarded as different clustering levels (resolutions). Therefore, we can find the most approximate number of clusters from these different clustering levels because the global modularity maximum value could be found among local modularity maximum values. Thus *Multi-level Aggregation Method* provides a heuristic scheme to determine the number of clusters automatically.

Before the clustering process, a distance matrix was built according to the cross-citation similarities between categories. Salton's cosine measure was used for the normalisation of similarities (Salton & McGill, 1983, see Eq. (3)).

Similarity between categories based on cross-citations (SCC):

$$\mathrm{SCC}_{ij} = \frac{a_{ij} + a_{ji}}{\sqrt{(\mathrm{TC}_i + \mathrm{TR}_i)(\mathrm{TC}_j + \mathrm{TR}_j)}} \tag{3}$$

where i and j denote categories, TC_i the total number of citations of category i, TR_i the total number of references of category i and a_{ij} the number of citations of category i receives from category j. Here, for calculating $\mathrm{TC}_i + \mathrm{TR}_i$, the intra-category self-citations are counted only once. Our clustering algorithm will be applied based on the symmetrical similarity matrix.

4. Results

The number of subject assignment in the Web of Science (SCIE, SSCI, AHCI) is 14,608 for 9487 journals during 2002–2006, namely, roughly 1.54 categories per journal. The average number of journals for each category is 59.4. Fig. 1 presents the 15 biggest ISI subject categories, each of which has more than 150 journals.

Among the 9487 journals under study, roughly 60% journals have single assignment for categories in ISI subject classification, and others have multiple assignments. The most assignments are, respectively, from *Chemometrics and Intelligent Laboratory Systems*, *Journal of Chemometrics* and *Open Systems & Information Dynamics*, each of which has been assigned to six ISI categories. The distribution of journals over ISI subject categories is presented in Fig. 2.

When we aggregate the journal–journal cross-citations to the category–category cross-citations, the self-citations of the journals having multiple assignments will automatically generate large amounts of citations over different categories. Taken into account the big share of multiple assigned journals, as well as the big share of journal self-citations, this aggregation will definitely impact or even distort the real network among categories. In order to avoid the latent distortion, we decided to exclude all the journal self-citation data before we got the aggregated category–category citation matrix. In other words, our category-to-category cross-citation matrix is aggregated from citations only among different journals.

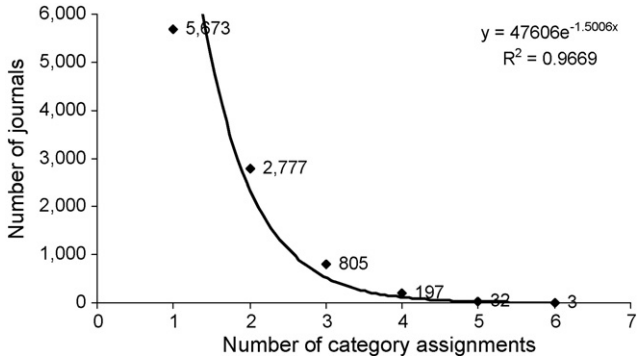


Fig. 2. Distribution of journals over ISI subject categories.

Table 1
Top 10 categories with highest entropies.

Category	Entropy	Category	Entropy
Literary reviews	7.05	Theatre	6.49
Humanities, multidisciplinary	6.95	Social sciences, interdisciplinary	6.39
Art	6.71	History and philosophy of science	6.37
Architecture	6.68	Asian studies	6.35
Film, radio, television	6.50	Literature, romance	6.29

Table 2
Top 10 science and social sciences categories with highest entropies.

Category	Entropy	Category	Entropy
History and philosophy of science	6.64	Ethics	6.21
Social sciences, interdisciplinary	6.40	Anthropology	6.20
Computer science, interdisciplinary applications	6.40	Engineering, biomedical	6.11
Public, environmental and occupational health	6.23	Medical informatics	6.08
Social issues	6.22	Sociology	6.07

4.1. Analysis of ISI subject categories based on different indicators

In order to measure in how far references/citations are spread among other journals, Zhang, Glänzel, et al. (2009) have introduced the indicator of entropy. Here we used the same indicator to measure the distribution of links among different categories.

$$\mathrm{EL}_i = -\sum_{j=1}^{n} \frac{a_{ij} + a_{ji}}{\mathrm{TC}_i + \mathrm{TR}_i} \log_2 \left[\frac{a_{ij} + a_{ji}}{\mathrm{TC}_i + \mathrm{TR}_i} \right] \tag{4}$$

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

Table 1 shows the top 10 categories with highest entropies, where 9 of them are assigned to arts and humanities. This is not surprising as it is well-known that the arts and humanities tend to communicate with a large scope of different categories. But when we analyzed the links these categories distributed, we found that they are of relatively weak affinity. This demonstrates that arts and humanities tend to have far-ranging but rather loose links, which may distort the whole picture of cross-citation network, as well as the clustering analysis among different categories. In order to avoid such distortion, we excluded 24 categories which are solely assigned to arts and humanities. Thus, in the following analysis, we will focus on the 222 categories covered in the sciences and social sciences.

As a contrast, we present the top 10 categories with highest entropies after the exclusion of arts and humanities (see Table 2). Social sciences categories occupy a big share; computer science, interdisciplinary applications has the highest entropy among the science categories. This result is in accordance with the research of Leydesdorff and Rafols (2009), where they got the conclusion that computer science, interdisciplinary applications, is the one with the highest interdisciplinarity among all science categories, although they used another indicator: betweenness centrality.

Opposing the entropy which measures the distribution of links within the communication network, the index of self-link mainly represents the degree of isolation (see Eq. (5)).

$$\mathrm{SLI}_i = \frac{\mathrm{SC}_i}{\sqrt{\mathrm{TC}_i \mathrm{TR}_i}} \tag{5}$$

Table 3
Top 10 science and social sciences categories with highest self-link index.

Category	SLI	Category	SLI
Astronomy and astrophysics	0.64	Economics	0.38
Mathematics	0.54	Business, finance	0.36
Law	0.54	Physics, particles and fields	0.35
Ophthalmology	0.46	Chemistry, organic	0.33
Dentistry, oral surgery and medicine	0.41	Polymer science	0.32

Table 4
Category pairs with strongest citation affinities.

Category *i*	Category *j*	Similarities
Mathematics, applied	Mathematics	0.2440
Business	Management	0.2289
Physics, multidisciplinary	Physics, particles and fields	0.2001
Business, finance	Economics	0.1948
Chemistry, multidisciplinary	Chemistry, organic	0.1808
Physics, particles and fields	Astronomy and astrophysics	0.1764
Political science	International relations	0.1759
Cell biology	Biochemistry and molecular biology	0.1758
Telecommunications	Engineering, electrical and electronic	0.1588
Physics, particles and fields	Physics, nuclear	0.1528

Table 5
The 16 science and social sciences categories with most strong links.

Category	SL	Category	SL
Biochemistry and molecular biology	13	Physics, applied	7
Materials science, multidisciplinary	9	Oncology	6
Chemistry, physical	8	Cell biology	6
Ecology	8	Chemistry, multidisciplinary	6
Engineering, electrical and electronic	8	Computer science, theory and methods	6
Geosciences, multidisciplinary	8	Economics	6
Neurosciences	8	Immunology	6
Physics, multidisciplinary	8	Psychiatry	6

where TC_i is the total number of citations of category i, TR_i the total number of references of category i and SC_i the number of self-citations of category i. The 10 most isolated categories are represented in Table 3, where the top three are, respectively, astronomy and astrophysics, mathematics and law. The striking values of SLI may indicate the high degree of specialisation, or the particular citation characteristics of these certain categories.

4.2. Analysis of strong links among subject categories

The symmetrical cross-citation matrix of ISI science and social sciences categories is densely populated: more than 90% of the cells (45,094 of $222^2 = 49{,}284$) have values above 0. In the cross-citation similarity matrix after the normalisation by Eq. (2), there are some distinctly high values protruding from the whole matrix. These extremely high similarities reflect the notable affinities between each two categories, such as mathematics, applied and mathematics, or business and management. In Table 4, we could observe the top 10 category pairs with strongest affinities (over 0.15). Most of these strongly linked categories are those of natural and applied sciences, such as physics, mathematics and chemistry, or social sciences, such as economics and political science.Unlike the cases in Table 2, where the categories tend to distribute their links to a far-ranging scope but regardless of the strength of those links, here we turned to detect the categories having relatively more strong links with others. Based on Eq. (2), we set a threshold of $SCC_{ij} \geq 0.05$, where $i \neq j$, for which we called as strong links. There are in total 239 links meeting this threshold, which occupy about 1% of all the similarities. The distribution of these strong links is highly concentrated on some of the natural and applied sciences, or life sciences. Table 5 presents the 16 categories with more than 5 strong links, where only one category: Economics belongs to social sciences. The comparison of Table 2 and Table 5 shows the considerable divergence: no one category is in common. In the cross-citation network, the categories either merely spread their information over and/or collect information from a variety of other categories but regardless of their intensity (like cases in Table 2), or tend to have strong influence from/on some particular categories but relatively weak in expanding their communication scope (like cases in Table 5). In general, social sciences categories are inclined to enlarge their link distributions, while science categories tend to have more intense links.

The categories shown in Table 5 could be considered as "central nodes" among the whole communication network. These "central" actors would form some coherent sub-clusters in the network, and act as "cores" in these clusters. It is worthwhile to have a look at those sub-clusters, where there are dense information communications. Here, we adopted an ego-centred neighbour map to present each of these sub-clusters. The category under study is positioned in the core, and its strongly

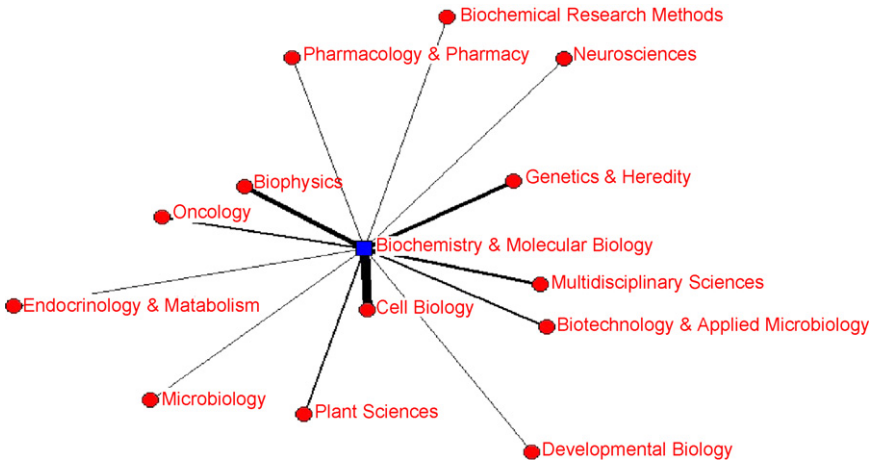


Fig. 3. Ego-centred neighbour map of biochemistry and molecular biology (visualisation by Netdraw).

related categories orbit around it. The affinity of the relationships is reflected here by both of the strength and the length of their links. As an example, Fig. 3 presents the neighbour map of biochemistry and molecular biology, where all its 13 strong links generate from the centre to the satellites. cell biology, biophysics and genetics and heredity are obviously the most closely related categories for biochemistry and molecular biology.

Using the same method, we could get the ego-centred neighbour maps for all the ISI subject categories. For those nodes with less or even no strong links, we may set some lower thresholds to present their communication neighbourhood. For reasons of research affinity, here we would give an example of the category information science and library science (see Fig. 4). For information science and library science, there are actually only two links meeting the threshold of strong ties, namely for computer science, information systems and management, respectively. In order to present the neighbour map, we lowered the threshold to 0.01. The map tells that the most strongly interlinked nodes for information science and library science are computer science, information systems; management; communication; business; medical informatics; computer science, software engineering; computer science, interdisciplinary applications; operations research and management science; computer science, theory and methods; health care sciences and services; ergonomics and computer science, cybernetics, respectively. This result is somewhat different from the conclusion drawn by Moya-Anegon et al. (2004); they actually found the most closely related categories of information science and library science are computer sciences and information systems, communication, history and philosophy of science, management, computer sciences and interdisciplinary applications, planning and development, business and social sciences-interdisciplinary, respectively.

4.3. ISI subject category cross-citation clustering analysis

The ISI category of multidisciplinary sciences, which includes journals such as *Science, Nature, PNAS USA* and *Endeavor* may not be a good case for cluster analysis. If for example an article about a specific discipline such as Allergy is published in

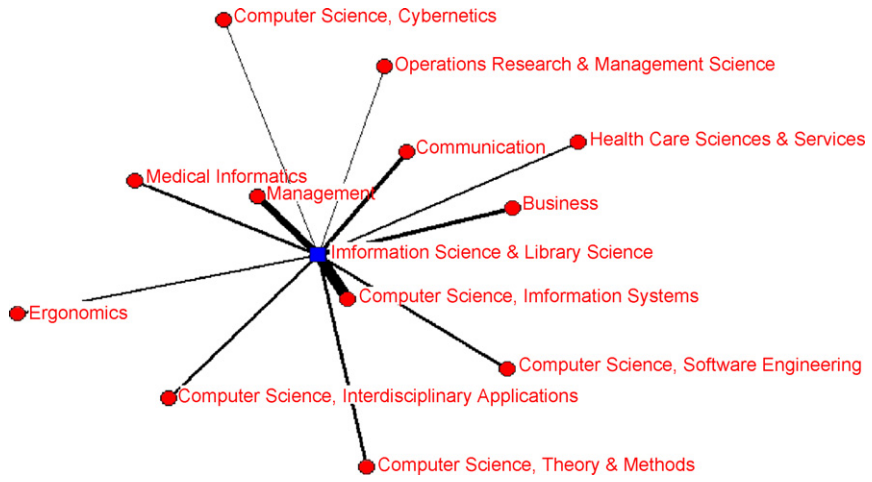


Fig. 4. Ego-centred neighbour map of information science and library science (visualisation by Netdraw).

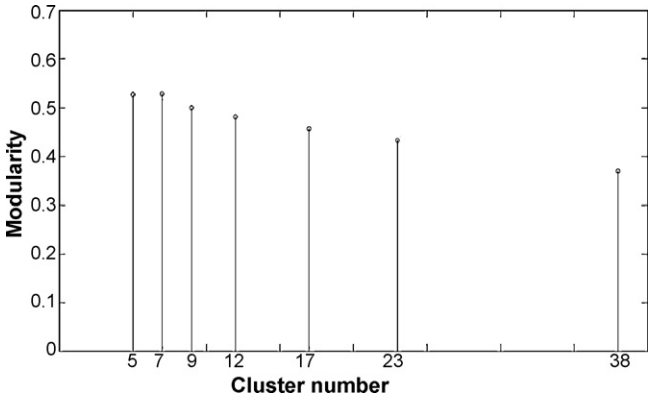


Fig. 5. Modularity values based on different local optimal numbers of clusters.

one of these journals, it is not reflected in the map of its domain, but is labelled as 'multidisciplinary'. In order to avoid the latent distortion, we excluded the multidisciplinary sciences category before applying the clustering algorithm. Therefore, there are 221 ISI subject categories left for the clustering analysis.

According to the *Multi-level Aggregation Method* clustering algorithm based on the cross-citation similarities, 7 is one of the local optimal values to be the number of clusters (see Fig. 5).

Table 6 presents the balanced clustering result of the 7 clusters solution. The categories in each cluster reveal the following structure: (i) the four natural and applied sciences clusters comprise biology, environmental science and geography (#1, 36 categories), agriculture and food Science (#2, 16 categories), computer science and mathematics (#3, 23 categories), chemistry, physics and engineering (#4, 40 categories); (ii) two social sciences clusters consist of cluster #5 (sociology, economics and political science, 41 categories) and cluster #6 (psychology and education, 22 categories); and finally, (iii) one life science and medical science cluster (#7, 43 categories).

The division between different clusters is basically clear and recognisable, which reflects the affinities and divergence among the 221 subject categories. There are a few obscure classifications in this structure, for instance, some agricultural categories are classified in cluster #1 (biology, environmental science and geography), where we admit that there is close relationship among agriculture and biology, environmental science and geography; some engineering categories are spread over several clusters, but it is mainly due to the application characteristics of the engineering science.

Zhang, Janssens, et al. (2009) applied a hybrid clustering method combining cross-citation and textual analysis to cluster more than 8000 journals covered in the Web of Science (2002–2006) into 7 groups. In that classification, there are (i) three natural and applied sciences clusters comprise biology, agriculture and environmental sciences; physics, chemistry and engineering; mathematics and computer science; (ii) two life-science clusters are, respectively, biosciences and biomedical research; clinical, experimental medicine and neurosciences; and finally, (iii) two social sciences clusters consist of economics, business and political science; psychology, sociology, education.

The two classification schemes, respectively, based on subject categories and journals are basically coincident, only with the exception that there are one life and medical science cluster and two separate clusters related to agriculture, biology and environmental in the category cross-citation clustering, while in the journal hybrid clustering system, the life and medical science cluster splits up into two clusters, and the two clusters related to agriculture, biology and environmental integrate into one. The divergence between the two structures may due to firstly the differences of the clustering similarities and algorithms, and secondly, to the fact that the journal classification is automatically generated from the clustering program based on the journal cross-citation and text similarities, but the category clustering is based on the journal assignments to ISI subject categories, which are on the basis of a number of criteria including the journal's title, its citation patterns, etc. The differences between the two classification schemes may also reveal some possible improvement for the journal assignments to the ISI categories.

The cross-citation network structure among different clusters of ISI categories is visualised by *Pajek* (Batagelj & Mrvar, 2002) in Fig. 6. 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:

$$S_{ij} = \frac{C_{ij}}{\sqrt{\sum_k C_{ik} \sum_k C_{jk}}} \tag{6}$$

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 categories in a given cluster. The life-science cluster has strong link with the agriculture science cluster, and is also close to one of the social sciences clusters: psychology and education. Chemistry, physics and engineering connect the computer science and mathematics and the life sciences. This network structure is similar to that of the 7 hybrid clusters of journals (Zhang, Janssens, et al., 2009).

Table 6
Cross-citation clusters for 221 ISI science and social sciences categories.

Cluster #	ISI subject categories
1	Agricultural engineering; agriculture, soil science; biodiversity conservation; biology; biology, miscellaneous; construction and building technology; ecology; engineering, civil; engineering, environmental; engineering, geological; engineering, marine; engineering, ocean; engineering, petroleum; entomology; environmental sciences; evolutionary biology; fisheries; forestry; geochemistry and geophysics; geography, physical; geology; geosciences, multidisciplinary; imaging science and photographic technology; limnology; marine and freshwater biology; materials science, paper and wood; materials science, textiles; meteorology and atmospheric sciences; mineralogy; oceanography; ornithology; paleontology; remote sensing; toxicology; water resources; zoology
2	Agriculture, dairy and animal science; agriculture, multidisciplinary; agronomy; biotechnology and applied microbiology; chemistry, applied; food science and technology; horticulture; infectious diseases; microbiology; mycology; nutrition and dietetics; parasitiology; plant sciences; tropical medicine; veterinary sciences; virology
3	Automation and control systems; computer science, artificial intelligence; computer science, cybernetics; computer science, hardware and architecture; computer science, information systems; computer science, interdisciplinary applications; computer science, software engineering; computer science, theory and methods; engineering, electrical and electronic; engineering, industrial; engineering, manufacturing; ergonomics; information science and library science; mathematical and computational biology; mathematics; mathematics, applied; medical informatics; operations research and management science; robotics; statistics and probability; telecommunications; transportation; transportation science and technology
4	Acoustics; astronomy and astrophysics; biochemical research methods; chemistry, analytical; chemistry, inorganic and nuclear; chemistry, medicinal; chemistry, multidisciplinary; chemistry, organic; chemistry, physical; crystallography; electrochemistry; energy and fuels; engineering, aerospace; engineering, chemical; engineering, mechanical; engineering, multidisciplinary; instruments and instrumentation; materials science, ceramics; materials science, characterization, testing; materials science, coatings and films; materials science, composites; materials science, multidisciplinary; mathematics, interdisciplinary applications; mechanics; metallurgy and metallurgical engineering; mining and mineral processing; nanoscience and nanotechnology; nuclear science and technology; optics; physics, applied; physics, atomic, molecular and chemical; physics, condensed matter; physics, fluids and plasmas; physics, mathematical; physics, multidisciplinary; physics, nuclear; physics, particles and fields; polymer science; spectroscopy; thermodynamics
5	Agricultural economics and policy; anthropology; area studies; business; business, finance; communication; criminology and penology; demography; economics; education, scientific disciplines; environmental studies; ethics; ethnic studies; family studies; geography; health care sciences and services; health policy and services; history; history and philosophy of science; history of social sciences; industrial relations and labor; international relations; law; management; medical ethics; medicine, legal; medieval and renaissance studies; nursing; planning and development; political science; psychology, applied; public administration; public, environmental and occupational health; social issues; social sciences, biomedical; social sciences, interdisciplinary; social sciences, mathematical methods; social work; sociology; urban studies; women's studies
6	Applied linguistics; behavioral sciences; clinical neurology; education and educational research; education, special; geriatrics and gerontology; gerontology; neuroimaging; neurosciences; psychiatry; psychology; psychology, biological; psychology, clinical; psychology, developmental; psychology, educational; psychology, experimental; psychology, mathematical; psychology, multidisciplinary; psychology, psychoanalysis; psychology, social; rehabilitation; substance abuse
7	Allergy; anatomy and morphology; andrology; anesthesiology; biochemistry and molecular biology; biophysics; cardiac and cardiovascular system; cell biology; critical care medicine; dentistry, oral surgery and medicine; dermatology and venereal diseases; developmental biology; emergency medicine; endocrinology and metabolism; engineering, biomedical; gastroenterology and hepatology; genetics and heredity; hematology; immunology; integrative and complementary medicine; materials science, biomaterials; medical laboratory technology; medicine, general and internal; medicine, research and experimental; microscopy; obstetrics and gynecology; oncology; ophthalmology; orthopedics; otorhinolaryngology; pathology; pediatrics; peripheral vascular disease; pharmacology and pharmacy; physiology; radiology, nuclear medicine and medical imaging; reproductive biology; respiratory system; rheumatology; sport sciences; surgery; transplantation; urology and nephrology

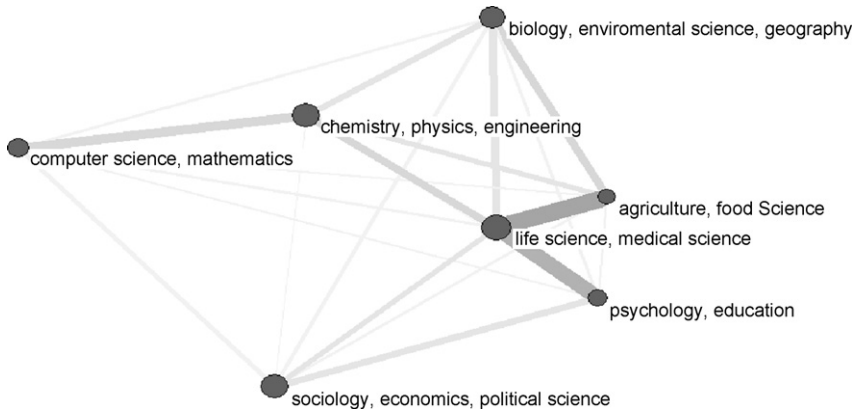


Fig. 6. Cross-citation network among 7 clusters of ISI categories (visualisation based on *Pajek*).

5. Conclusions and discussions

The *Multi-level Aggregation Method* generates a balanced clustering result of the ISI subject categories. The components of these clusters clearly distinguish from each other, and each cluster represents one of the scientific domains, respectively. Several indicators have been used to compare the communication characteristics among different ISI subject categories. In general, social sciences categories are inclined to enlarge their link distributions, while science categories tend to have deep contacts. The neighbour map of each category clearly reflects the affinities between the "core" category and its satellites around.

There are indeed structural differences between the elaborate clustering of more than 8000 journals and the clustering of the ISI subject categories. The former clustering results are generated automatically based on the journal-to-journal similarities, and are then labelled using the best TF-IDF terms from all documents under study in these individual journals; while in the clustering based on ISI subject categories, we first assign all the individual journals into different categories according to the ISI assignment and then aggregate all the journal-to-journal citation data to category-to-category citation data. The clustering is thus analyzed at the category level, and the labelling for each cluster is based on the names of ISI categories included. Therefore, the two clustering results provide two subject classification schemes through different perspectives and levels. The two classifications are structurally comparable but differences indeed exist. The divergence between the two structures may be due to the interferences from the multiple journal assignment to ISI subject categories, and on the other hand, may also reflect some possible improvement of the journal assignment scheme in ISI.

The above-mentioned main idiosyncrasy of the ISI category classification is that it may assign one single journal to different categories in view of the subject matter of the documents they publish. If category *i* and category *j* share relatively a lot of common journals, it may also reflect a strong affinity between the two categories. Thus, the subject clustering based on the co-assignment similarities may be another nice experiment to analyze the subject classification system, and it may be also interesting to compare these different clustering results. However, this will be part of our future research.

Acknowledgement

The research was supported by Centre for R&D Monitoring of the Flemish Government, the National Natural Science Foundation of China (grant no. 70673019), China Scholarship Council and ERCMAMT (Engineering Research Centre of Metallurgical Automation and Measurement Technology, Ministry of Education, Hubei, China).

References

Batagelj, V., & Mrvar, A. (2002). Pajek–analysis and visualization of large networks. *Graph Drawing*, *2265*, 477–478 [ISSN: 0302-9743].

Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics*, *64*, 351–374.

Chen, C. M. (2008). Classification of scientific networks using aggregated journal–journal citation relations in the journal citation reports. *Journal of the American Society for Information Science and Technology*, *59*(14), 2296–2304.

Doreian, P., & Fararo, T. J. (1985). Structural equivalence in a journal network. *Journal of the American Society for Information Science*, *36*, 28–37.

Garfield, E., Malin, M. V., & Small, H. (1975). A system for automatic classification of scientific literature. *Journal of the Indian Institute of Science*, *57*, 61–74.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics*, *56*(3), 357–367.

Janssens, F., Zhang, L., & Glänzel, W. (2009). Hybrid clustering for validation and improvement of subject-classification schemes. *Information Processing & Management*, *45*(6), 683–702.

Leydesdorff, L. (1987). Various methods for the mapping of science. *Scientometrics*, *11*, 291–320.

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 and Technology*, *57*(5), 601–613.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology*, *60*(2), 348–362.

Moya-Anegon, F., Vargas-Quesada, B., Herrero-Solana, V., Chinchilla-Rodriguez, Z., Corera-Alvarez, E., & Munoz-Fernandez, F. J. (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. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Science of the United States of America*, *103*(23), 8577–8582.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Small, H. (1999). Visualizing science by citation mapping. *Journal of the American Society for Information Science*, *50*, 799–813.

Tijssen, R., De Leeuw, J., & Van Raan, A. F. J. (1987). Quasi-correspondence analysis on square scientometric transaction matrices. *Scientometrics*, *11*, 347–361.

Zhang, L., Glänzel, W., & Liang, L. M. (2009). Tracing the role of individual journals in a cross-citation network based on different indicators. *Scientometrics*, *81*(3), 821–838.

Zhang, L., Janssens, F., Liang, L. M., & Glänzel, W. (2009). Hybrid clustering analysis for mapping large scientific domains. In B. Larsen, & J. Leta (Eds.), *Proceedings of ISSI 2009–the 12th international conference on scientometrics and informetrics* Rio de Janeiro, (pp. 178–188).