Scientometrics
DOI 10.1007/s11192-016-1890-9

The study of subject-classification based on journal coupling and expert subject-classification system

Jing Zhang[1] · Xiaomin Liu[1] · Lili Wu[1]

Received: 27 August 2015

Abstract As the framework of scientific research, subject-classification plays an important role in the development of science. In order to combine the development of science with the current expert subject-classification system and further give a more appropriate description of scientific output analysis from subject level, We study the relationship between the natural science related sub-categories of Chinese library classification using objective computerized scientometrics, and give some modification to the first two level subjects of the existing Chinese library classification system. Taking Chinese Science Citation Database as our data source, this article studies the similarity of subjects based on journal coupling strength. Then we try to set up an improved subject-classification system whose top categories are relied on Chinese library classification system and sub-categories are the ensemble clustering result based on journal coupling measure. Further, in order to help identifying and interpreting the rationality of this improved classification system, we make use of some text mining methods, such as key words recognition and topic detection, to explain the cause of similarity between some subjects from the perspective of semantic. Our study shows that the improved subject-classification system constructed in this article not only conforms to previous experience and cognitive but also combines subject development knowledge.

Keywords Subject-classification · Journal coupling · Cluster analysis · Text mining · Chinese library classification

✉ Jing Zhang
xieyun-xieyun@163.com

[1] National Science Library, Chinese Academy of Sciences, No. 33, West Road, North 4th Ring Road, Beijing 100190, People's Republic of China

Published online: 15 February 2016

Introduction

Subject-classification is external expression of the internal structure of science and it reveals the inherent relationship among scientific fields, which has got important values in both theoretical research and practical application. Most existing classification systems belong to the so-called expert classification system, whose formation mainly depends on human judgment and longstanding experience. And such systems include 22 broad classification system used in the ESI database and the 179 system in JCR database of Web of Science. Also the Chinese library classification (fifth edition) system (hereafter CLC) whose construction relies on the scholars and experts coming from various fields, and it is a four level hierarchical knowledge map of social science and natural science, containing 22 top categories, 262 sub-categories, third level categories and fourth level categories. Other systems such as, Katz and Hicks (1995) developed a journal classification scheme to examine sectoral output and collaborative activity by reclassification of the SCI 154 subfields into the 10 broad Australian Standard Research Classification Scheme fields, which combines expert discussion and judgment, and the 15 subject system presented respectively by Glänzel and Schubert (2003) and Boyack et al. (2005).

Although expert classification systems indeed agglomerate much valuable human knowledge and subjective judgment, such systems also have some disadvantages for lacking of enough objectivity and being not able to reflect the development of science in time. In view of this, along with the development of computerized scientometrics, the study of subject-classifications and specialties based on objective scientometrics becomes popular, and most researches take journal-based cross-citation relationship[1] as a medium to study the relationship between subjects. In addition, author-based and article-based cross-citation relationship was also used in some study.

On journal-based research, Cason and Lubotsky (1936) was the earliest to study the interaction between subjects using journal cross-citation relationship. Then Daniel and Loutitt (1953) made use of journal cross-citation matrix in journal clustering analysis in the field of psychology for the first time. And Narin et al. (1972) firstly studied the relationship among journals which belong to different subject according to the cross-citation relationship between these journals. Lately, Narin (1976) and Leydesdorff (1993, 2002) had both applied the multivariate statistical analysis methods, such as PCA and FA, to a kind of bottom-up clustering analysis based on journal-based cross-citation analysis, who then made some confirmatory study on the consistency between the agglomerated subject system and the existing expert subject-classification system. Then, Leydesdorff (2004a, 2004b) applied graph theory methods such as bi-connected component analysis to the clustering analysis of journal cross-citation graph based on Web of Science's JCR database. In 2008, he (Leydesdorff 2008) then made some visualization study on different science fields derived from journal cross-citation relationship, which gives light to the development trend of scientific structure in the time dimension. Following, Zhang et al. (2010, 2012) studied the classification system based on journal cross-citation relationship, then its concordance with an existing expert system called SOOI, as well as made some

[1] In this study, we use term "cross-citation" to refer to the citing and cited behavior among articles, journals and authors and so on. Hereafter, we will also mention term "coupling", such as "journal coupling", and this refers to the measurement we used to study the similarity between different journals or different subjects based on the "cross-citation" behavior among them. That is to say, in this study we use "journal coupling" to study the "cross-citation" relationship among subjects, so the "cross-citation" relationship is the basis of the similarity measure "journal coupling".

advices to the adjustment and improvement of this expert systems. In 2011, Archambault et al. (2011) proposed a scientific journal ontology for production of bibliometric data, this method took three existing subject-classification as a start, and updated journal's subjects based on a classification engine which using not only citation relationship between journals but also some text level similarity. Along with the development of network study and visualization technology, many researches combing journal–journal citation (Leydesdorff and Rafols 2012) network classification or network classification based on journal similarity matrix (Börner et al. 2012)with network visualization in the construction of subject-classification map. Then, some network classification, like community detection algorithms: VOS Clustering and Louvain Method, based on journal cross-citation relationship were applied to the quantitative construction of subject-classification (Gómez-Núñez et al. 2014).

While, in the author-based and article-based research, making use of cross-citation behavior among articles in the mining of specialty was proposed by Braam et al. (1991). Ahlgren and Colliander (2009) studied the subject-classification from article level, and the similarity measures between different articles contain text similarity, bibliographical coupling and a combination of the two, then hierarchical clustering method was applied to construct a final classification. Chen et al. (2010) made some quantitative analysis of scientific structure using the author cross-citation behavior and article cross-citation behavior respectively. Then Waltman and Van Eck (2012) constructed a publication-level hierarchical subject-classification system in a million level articles by a iteration of citation based similarity measure between different article pair, and clusters of these article constitute the final classification system. White and McCain (1998), as well, studied the scientific structure and its development which hiding behind the author cross-citation data. Moreover, Luka Kronegger et al. (2013) also took advantage of the co-author relationship to study the changes of the subject along with time, then the development of the science.

As a classical expert subject-classification system, CLC has a wide application in practice, especially its sub-categories. But with the development of science, the sub-categories, especially those belonging to a same top category, have revealed a problem that some of them are too similar to remain as separate subjects. And this will lead to subject classifying ambiguous or repeated classification, which has great influence on the analysis of scientific output analysis. In view of this, this paper on the basis of journal cross-citation data collected from Chinese Science Citation Database (here after, CSCD)and under the premise of keeping the 10 nature science related top categories unchanged in the CLC, uses the journal-based coupling and clustering analysis to study the correlation of the sub-categories belonging to the same top category. By ensemble learning of the multiple clustering results of these sub-categories, we proposed an improved hierarchical subject-classification system. The system we constructed is a hybrid one because it is not only based on the cross-citation behavior, but also relies on the expert system CLC. This classification system both has got the qualitative characteristics and quantitative characteristics, that is to say, it retains the crystallization of expert's wisdom as well as integrates the development of subject. Then we construct topic models on the original bibliographies to help us validate the journal-based coupling similarity measure by checking the research topic coherent and semantic similarity of similar subjects calculated by our method.

Methods

Subject reflects the commonality and difference of scientific knowledge, and is of great significance to the study of science's development as well as the promoting of scientific research. Previous researches on subject-classification mainly can be divided into two categories—quantitative research and qualitative research. And most of the existing subject-classification systems are the product of qualitative research—expert subject-classification systems. The quantitative research of subject-classification is still in an exploratory, validating, and supplementary stage. There is no doubt that expert systems contain a large number of subjective cognitive and knowledge, and have important leading role in the development of subject-classification system. While, being able to fully mining the objective knowledge embedded in the cross-citation relationship of bibliographies, subject-classification systems based on quantitative analysis has advantages of catching the development of science. In order to organically combine the qualitative and quantitative research in the study of subject-classification system, we set up an improved system which reflects both the stability and evolution of science.

We take the first two level subject of CLC as prototype and make ensemble clustering analysis of the sub-categories based on journal coupling. The correlation measure in our study is journal-based coupling and the further clustering method is an ensemble learning of several classical hierarchical or partition-based clustering methods. After this process, we construct a modified two-level hierarchical subject system whose top category is relied on CLC and its sub-category is the clustering result based on journal coupling strength.

Correlation measure

Cross-citation behavior can manifest the rule of scientific development and reflect the cumulativeness, continuity and inheritance of scientific knowledge, so it can catch the change of science structure. This paper studies the subject-classification system grounded on subject similarity which measured by journal coupling strength, and expects to reflect the subject similarity through cross-citation among journals which belong to different subjects.

Bibliographic coupling and co-citation are two main method to measure the correlation among nodes (here is journal) in a cross-citation network. Bibliographic coupling was firstly proposed by professor Kessler (1963) in MIT (Massachusetts Institute of Technology in American).Since then, the strength of coupling has been used in measuring the correlation among journals in a journal-based cross-citation network (Ni et al.2013; Qiu and Liu 2014), in article cross-citation network (Ahlgren and Colliander 2009), and among authors in an author-based cross-citation network gradually (Zhao and Strotmann 2008; Rousseau and Zuccala 2004; Qiu and Dong 2013).

The concept of co-citation was originated in 1973, proposed by former soviet information scientist Irina Marshakova (1973) and American information scientist Small (1973) respectively. Co-citation was firstly used in the measurement of similarity among articles in the cross-citation network, then further been introduced to study journal co-citation relationship (Chen et al. 2010) and author co-citation relationship (Braam et al. 1991, Rousseau and Zuccala 2004). Gómez-Núñez et al. (2014, 2015) also proposed to use a weighted combination of both the bibliographic coupling, cross-citation and direct citation as the similarity measure when performing the journal based subject-classification study.

Bibliographic coupling and co-citation are both measures that reflect the correlation of nodes, what's the difference is that they are from two direction of a citation relationship. Bibliographic coupling measures backward citation which reflects a kind of static and stable relationship. Oppositely, co-citation measures from the forward direction, reflecting a dynamic relationship. But from the perspective of cross-citation network, they both can be seen as methods of one-step network correlation measures. Although there exist differences, journal coupling strength and journal co-citation strength can all reflect the correlation between subjects. Compared with journal-based co-citation which measures the number of common citing journals two cited journal share, we choose the journal-based coupling method to measure correlation between subjects. We use this measure for the reason that journal-based coupling can keep the number of journals participated in our analysis increasing as much as possible through measuring the number of common cited journals shared by two citing journal, which will improve the accuracy of the subsequent analysis.

Clustering method

The nature of subject-classification research based on cross-citation relationship is the study of the commonality and difference between nodes (here is journal) in a cross-citation network, and subject or specialty is the abstract of a set of similar nodes, and the process of searching for coherent nodes can be seen as clustering analysis. In previous studies, clustering methods used in subject-classification analysis can be divided into three categories, methods grounded on multivariate statistics, such as principal component analysis (or factor analysis) (Leydesdorff and Cozzen 1993; White and McCain 1998; Leydesdorff 2006). Methods using the classical clustering analysis, for instance, hierarchical clustering (Zhang et al. 2010, 2012; Braam et al. 1991; Kronegger et al. 2013; Ahlgren and Colliander 2009), minimum spanning tree (Chang and Chen 2011), etc. The last one are methods belonging to the clustering of social network in graph theory (Chen et al. 2010; Qiu and Liu 2014; Waltman and Van Eck 2012; Leydesdorff and Rafols 2012; Börner et al.2012; Gómez-Núñez et al. 2014).Clustering methods based on multivariate statistical theory take the node pair which has citation behavior as the variable and case respectively, then clustering the cases with same characteristics using the idea of projection, but there is no definite standard on the division of clusters, this is also true when it comes to the choosing of cluster number, and the results of such methods can hardly form a clear hierarchical subject-classification system. The graph theory based clustering methods (such as, community detection) take the cross-citation network as a whole, and the similarity measure used in these methods are some kind that beyond one-step measure in a network. And this kind of methods can reflect the commonness and difference between each node in a network more comprehensively, while the quality of cross-citation data and the sample size have much to do with the ultimate clustering performance in this kind of methods, in addition, the determination of cluster number needs some experience judgment. The main research object of classical clustering method is not the whole cross-citation network but the nodes in the network, and the similarity between nodes is mainly measured by a kind of one-step method in graph theory. When compared to graph theory related clustering method, the classical clustering methods have problem of utilizing data inadequately, but the classical clustering method is not restricted by data quality as well as the sample size.

Considering that the objects for clustering analysis in this paper is the sub-categories belonging to a same top category, and the research object itself is a small sample. After taking the restrictions of the clustering methods referred above into consideration, we decide to use the classical clustering methods, such as hierarchical and partition-based

Table 1 Members of our clustering system

Clustering system	Clustering methods
Partition-based clustering algorithm	Kmeans (Hartigan and Wong 1979; MacQueen 1967)
	PAM (Reynolds et al.1992)
Hierarchical clustering algorithm	Agglomerated clustering algorithm (Everitt 1974) (including average method; complete linkage; median linkage; Ward's method; Mcquitty's method)
	AGNES (Kaufman and Rousseeuw 1990) (containing: average method; complete linkage; Ward's method; weighted average linkage)
	DIANA (Kaufman and Rousseeuw 1990)

clustering methods, listed in Table 1 which are not strict with sample size and the distribution of the original data. From the perspective of statistic, there exist no methods who can give best result, and the existing clustering methods may all have some instability performance. So following the clustering process, we use the ensemble learning theory in machine learning to make ensemble of the 13 clustering results in order to improve the accuracy and validity of the clustering results.

Experiment and result

Data

In this paper, the data source is CSCD. 932, 429 publications and more than 12.5 million references of 1286 source journals from 2009 to 2011 are retrieved. Due to CSCD's focusing on natural science, this study pays main attention to classification of the 126 nature science related sub-categories from CLC. The number of CSCD source journal is small, while cited journals set is large. When measuring subject similarity, compared with the journal-based co-citation method, calculation of journal coupling can cover greater amount of journals. In order to improve the accuracy of our final results, this study chooses the journal-based coupling as method to measure the similarity between subjects. In addition, subject-classification system adopted by CSCD is CLC, which provides data foundation for our following research.

Calculation of journal-based coupling

Step 1: The statistical object is the journal-typed data among the bibliographies and references from 2009 to 2011 in CSCD.
Step 2: For bibliographies coming from the source journals of CSCD, we firstly mapping their subject- classification all to the sub-categories in CLC.
Step 3: Constructing the subject-journal matrix through the citation behavior of the citing and cited bibliographies, Firstly, in order to decrease the sparse degree of the matrix, for each sub-category, we intercept its cited journals so that the cumulative percentage of the journal citations is less than 80 %.[2] Then we make our matrix whose rows contain the citing sub-categories belonging to a same top category of CLC, and columns are the cited journals by these sub-categories.

[2] 80 % is determined by repeated trials so that sparse degree of the adjacency matrix can be reduced significantly and the raw information cannot be loss much.

Step 4: In order to avoid the influence of journals' volume, period number, the article type and number of articles to the calculation of times cited from the subjects to journals, this study transforms the matrix in step 3 to a 0–1 matrix,[3] this then can eliminate the interference to journal coupling strength calculation by the above factors.
Step 5: Convert the 0–1 matrix derived in step 4 to the subject–subject distances matrix for clustering analysis. From step 4 we derived a 0–1 matrix whose columns are 0–1 vector which is binary, so when calculating the distances matrix we choose the Gower's coefficient,[4] and the distance of a subject to itself is not considered.

The clustering results

Process of clustering

We apply the clustering methods listed in Table 1 to the distance matrix of the sub-categories derived in step 5 of 'Calculation of journal-based coupling' section, and then made ensemble learning of the 13 classification results.

Clustering analysis is a kind of unsupervised method, and the number of the clusters is not explicitly given in the final result, so the determination of cluster number has great influence on research performance. In order to reduce subjective judgment to the clustering results, this article uses the Gap statistic (Tibshirani et al. 2001) to help determining the cluster number. Following we will take the top category Q (Biological sciences) and its 17 sub-categories as example to show our clustering process. We take the results of the DIANA methods on the 17 sub-categories as an example and the implementation of other 12 clustering methods is similar, the Gap statistic[5] of different clustering number is presented in Fig. 1, the figure shows that agglomerating to 5 or 6 cluster is better, we then check the value of the Gap statistic to seek for cluster number which makes the $log\ (W\ (k))$ decrease most fast, and finally we determine the optimal clustering number is 6.

From the dendrogram of DIANA method (Fig. 2), the final result of 6 clusters is: sub-category—Q (comprehensive biology) is a separate class; Q-(Q-0 theories and methods of biology science, Q-1 status and development of biological sciences, Q-3 biological scientific research methods and technology, Q-4 biological science education and popularization, Q-9 biological resources survey)and Q2 (cytobiology), Q3 (genetics), Q4 (physiology), Q5 (biochemistry), Q6 (biophysics), Q7 (molecular biology), Q81 (biological engineering (biotechnology)), Q93 (microbiology) for a class; Q1 (general biology), Q94

[3] In subject-journal matrix we derived from step 3, the number in row i and column j indicate times cited of journal j by subject i. In order to avoid the influences indicated in step 4, we choose to calculate the journal-based coupling strength of different subject with the simple method (a basic method of calculating coupling strength), which only consider the number of journals coupled by two subjects not cites. So we change the original cites in matrix to 0–1 which indicated if the citation from subject to journal is exist or not. Well, the simple method of coupling has problems of using original information insufficiently. But compared with bias coming from the sensitive cites, bias coming from the insufficient data usage is smaller, so we eventually choose this method, and further we will make great effort to improve our data quality and try to apply other coupling calculation method, such as the binary one proposed by Rousseau et al. (2004).

[4] We choose the general Gower's coefficient for the reason that it is suitable for handling of nominal, ordinal, and binary data. Moreover, due to including weights to different variable, the calculation of distance is more robust.

[5] For each number of clusters k, it compares $log\ (W\ (k))$ with $E^*[log\ (W\ (k))]$ where the latter is defined via bootstrapping, i.e. simulating from a reference distribution. The optimal number of cluster is the one who make the $log\ (W\ (k))$ decrease most fast, that is make the Gap statistics increase most fast to its maximum.

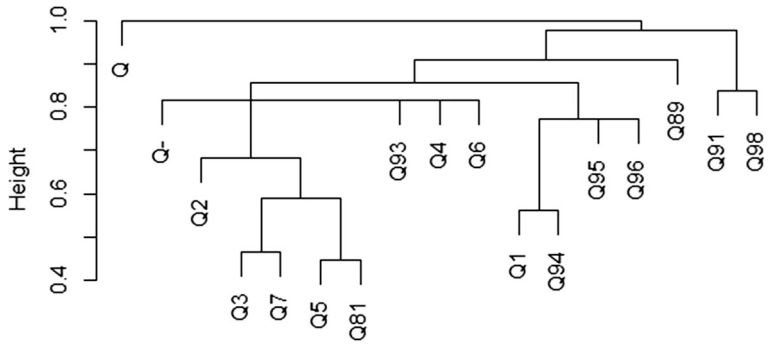

Fig. 1 Gap statistic of different clustering number based on DIANA clustering of Q (Biological sciences)'s 17 sub-categories, the vertical line segment of each point indicates the standard deviation of this point's gap statistic

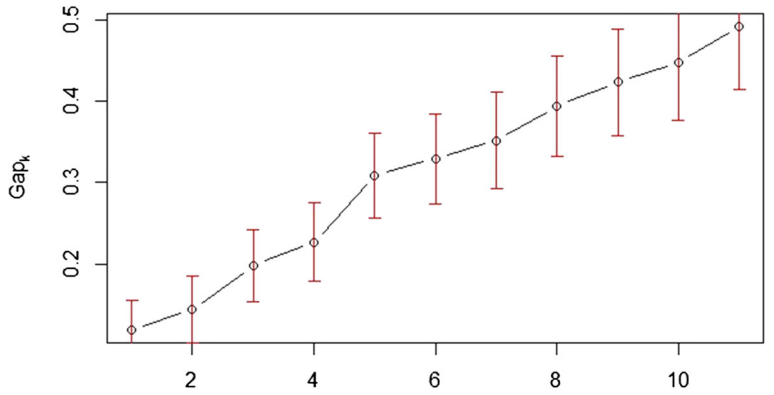

Fig. 2 The dendrogram of DIANA clustering method of the Q (Biological sciences)'s 17 sub-categories

(botany), Q95 (zoology), Q96 (entomology) gathered into a category; Q91 (paleontology) as a separate class; Q98 (anthropology) a separate one; Q89 (environmental biology) is another separate one.

We then implement the other 12 clustering methods listed in Table 1 to the 17 sub-categories of Q (Biological sciences) and the ensemble learning result of the 13 clustering results is showed in Table 2.

Through the clustering result of Q (biological sciences) in Table 2, we find that there indeed exists highly correlated phenomenon between sub-categories of Q (Biological sciences). And compared with the actual definition of each sub-category in CLC, the clustering result also showed our subject clustering analysis based on journal coupling strength is reasonable.

Modified subject-classification system

Then this study applies the above clustering process to the whole 10 nature science related top category in CLC. Finally, we proposed the improved nature science related hierarchical

Table 2 The clustering results of Q (Biological sciences)'s 17 sub-categories

Top category	Sub-category
Q (biological sciences)	Q (comprehensive biology)[a]
	Q1 (general biology), Q94 (botany), Q95 (zoology), Q96 (entomology)
	Q-(Q-0 theories and methods of biology science, Q-1 status and development of biological sciences, Q-3 biological scientific research methods and technology, Q-4 biological science education and popularization, Q-9 biological resources survey) and Q2 (cytobiology), Q3 (genetics), Q4 (physiology), Q5 (biochemistry), Q6 (biophysics), Q7 (molecular biology), Q81 (biological engineering (biotechnology)), Q93 (microbiology)
	Q89 (environmental biology)
	Q91 (paleontology)
	Q98 (anthropology)

The sub-category Q (comprehensive biology) does not real exist in CLC; we define this subject to contain journals whose content is comprehensive biology that cannot be classified to the specific sub-categories in Q (Biological sciences). And we keep this defined sub-category to make our study more practical. This is true for other comprehensive sub-categories in Table 3 like P (comprehensive astronomy, earth science), R (comprehensive medicine)

subject-classification system as showed in Table 3 whose top categories relies on the first-level subject-classification of CLC and the sub-categories is the results of subject clustering. And in Table 4 we give some comparison analysis of the sub-category number before and after our improvement.

Through the calculation of journal-based coupling, we find the number of CSCD journals belonging to U (communications and transportation) and V (aeronautics and astronautics) subject is too small to carry out the clustering analysis of its sub-categories, that is to say there exist too little research output to carry out so much detailed subject division, so the sub-categories of these two top category can be regarded as a whole one. And compared with the original nature science related subjects of CLC in Appendix 1, the journal coupling strength of all sub-categories of N (general natural science), besides N9 (system science) who is a more specific subject compared with the other general and abstract sub-categories (such as, N0 (natural science theory and methodology), N1 (present situation and development of natural science) in N, are extremely high and can be clustered into one super cluster. And the journal coupling strength of all sub-categories of X (environmental science and safety science), besides X9 (safety science), are extremely high and can be clustered into one environmental related super cluster. And the sub-categories of O (mathematical science and chemistry), P (astronomy, earth science) have few changes because their sub-categories have little overlapping.

While the sub-categories of Q (biological science) and S (Agricultural Science) have changed much compared to the original classification system in Appendix 1, there exist strong correlations between the sub-categories of these two top categories respectively. The 15 sub-categories of Q (biological science) has been clustered into 6 super sub-categories as showed in Table 2 and Table 3, In Q, sub-categories (like Q94 botany, Q95 zoology) whose research object is macro are clustered to a sub-category; sub-categories whose research object is micro (like Q3 genetics, Q2 cytobiology, Q7 molecular biology) are clustered to a sub-category; environmental related subject Q89 (environmental biology) as a separate sub-category; Q91 (paleontology) as a separate one; Q98 (anthropology) as another separate one. And the 11 sub-categories of S (Agricultural Science) has been

Table 3 Improved nature science related hierarchical subject-classification system

Top category	Sub-category
N (general natural science)	N0 (natural science theory and methodology), N1 (present situation and development of natural science), N2 (natural science institutions and organizations, meeting), N3 (methodology for natural science), N4 (science education and popularization), N5 (Natural science series, corpus), N6 (natural science reference books), N7 (natural science literature retrieval reference books), N79 (book materials, audio-visual materials), N8 (natural science investigation and inspection), N91 (Nature study, natural history), N93 (nonlinear science), N99 (Information science, intelligence work)
	N94 (system science)
O (mathematical science and chemistry)	O1 (mathematic)
	O3 (mechanics)
	O4 (physics)
	O6 (chemistry)
	O7 (crystallography)
P (astronomy, earth science)	P (comprehensive astronomy, earth science)
	P1 (astronomy)
	P2 (topography)
	P3 (geophysics)
	P4 (atmospheric science (meteorology))
	P5 (geology)
	P7 (oceanography)
	P9 (physical geography)
Q (biological sciences)	Q (comprehensive biology)
	Q1 (general biology), Q94 (botany), Q95 (zoology), Q96 (entomology)
	Q-(Q-0 theories and methods of biology science, Q-1 status and development of biological sciences, Q-3 biological scientific research methods and technology, Q-4 biological science education and popularization, Q-9 biological resources survey) and Q2 (cytobiology), Q3 (genetics), Q4 (physiology), Q5 (biochemistry), Q6 (biophysics), Q7 (molecular biology), Q81 (biological engineering (biotechnology)), Q93 (microbiology)
	Q89 (environmental biology)
	Q91 (paleontology)
	Q98 (anthropology)
R (medicine)	R (comprehensive medicine)
	R4 (Clinical medicine), R5 (internal medicine)
	R1 (preventive medicine, hygiene)
	R2 (Chinese medicine)
	R3 (preclinical medicine)
	R6 (chirurgery)
	R71 (gynecotokology)
	R72 (pediatrics)
	R73 (oncology)
	R74 (neurology and pathergasiology)

Table 3 continued

Top category	Sub-category
	R75 (dermatology and venereology)
	R76 (otorhinolaryngology)
	R77 (ophthalmology)
	R78 (stomatology)
	R79 (foreign national medicine)
	R8 (special medicine)
	R9 (pharmacy)
S (agricultural science)	S (comprehensive agricultural science)
	S-(S-0general theory, S-1present situation and the development of agricultural science and technology, S-3agricultural science research and experiment, S-9agricultural economy)
	S1 (agricultural basic science), S2 (agricultural engineering)
	S3 (agriculture (agronomy)), S5 (crops), S6 (gardening), S4 (plant protection)
	S7 (forestry)
	S8 (animal husbandry, animal medicine, hunting, silkworms and bees)
	S9 (aquaculture, fishing)
T (industrial technology)	T-(T-0 theory of industrial technology, T-1present situation and the development of technology, T-2Institutions, organizations, conferences, T-6reference book, T-9industrial economy)
	TB (General industrial technology)
	TD (mineral engineering)
	TE (Oil and gas industry)
	TF (metallurgical industry), TG (metallographic and metal crafts)
	TH (machinery and instrument industry)
	TJ (arms industry)
	TK (energy and power engineering)
	TL (atomic energy technology)
	TM (electro-techniques)
	TN (electronic technology, communication technology), TP (automation technology, computer technology)
	TQ (chemical industry)
	TS (light industry, handicraft industry and living services)
	TU (building science)
	TV (hydraulic engineering)
U (communications and transportation)	U-9 (transportation economy), U1 (integrated transportation), U2 (railway transportation), U4 (highway transportation), U6 (waterway transportation), U8 (air transport)
V (aeronautics and astronautics)	V1 (aviation, space technology research and exploration), V2 (aviation), V4 spaceflight (astronavigation), V7 (aerospace medicine)

Table 3 continued

Top category	Sub-category
X (environmental science and safety science)	X-(X-0environmental scientific theory, X-1environmental science and development, X-2environmental protection organizations, institutions, meeting, X-4environmental protection publicity, education and popularization, X-6environmental protection reference books), X1 (Environmental science basic theory), X2 (Society and Environment), X3 (environmental protection and management), X4 (disaster and prevention), X5 (environment pollution and protection), X7 (industry pollution, waste disposal and comprehensive utilization), X8 (environmental quality assessment and environmental monitoring)
	X9 (safety science)

Table 4 The number of sub-category before and after improvement

Top category	Number of origin sub-categories	Number of improved sub-categories
N (general natural science)	14	2
O (mathematical science and chemistry)	5	5
P (astronomy, earth science)	8	8
Q (biological sciences)	17	6
R (medicine)	18	17
S (agricultural science)	11	7
T (industrial technology)	17	15
U (communications and transportation)	6	1
V (aeronautics and astronautics)	4	1
X (environmental science and safety science)	9	2

clustered into 7 as showed in Table 3, Comprehensive agricultural sub-categories (S1 agricultural basic science and S2 agricultural engineering) have been grouped into a sub-category in the improved classification system; plant related sub-categories (like S5 crops, S3 agronomy, S6 gardening)have been grouped into one; forestry related sub-category S7 as a separate one; animal and insect related sub-category S8 as a separate one; S9 (aquaculture, fishing) as a separate sub-category. And according to its definition in the CLC, we found their qualitative definition and the correlation measured based on journal coupling has greatly consistency.

Finally, the variation of the sub-categories of R (medicine) and T (Industrial Technology) are not so much. Among the sub-categories of R (medicine), R4 (Clinical medicine) and R5 (internal medicine) agglomerates to a super subject due to that more and more internal medicine research have been containing clinical methods and much more clinical researches are related to diseases belonging to internal medicine. And among the sub-categories of T (Industrial Technology), TF (metallurgical industry) and TG (metal science and metal techniques) has been gathered into a super sub-category due to their strongly relationship with metal. And the strong intersection has made sub-category TN (electronic technology, communication technology) and TP (automation technology, computing technology) aggregated to a super computer technology related subject.

Semantic verification of our clustering results

In order to verify the rationality of the subject study method proposed in this paper, we make some text mining research to analyze semantic similarity between similar sub-categories calculated by our method. In clustering process of the nature science related sub-categories of CLC, we find some similar sub-categories that beyond our current knowledge. For instance, during the clustering analysis of R (medicine) subject, we find that though not being aggregated into a super sub-category after the optimal clustering number is determined, the distance matrix of sub-categories of R (medicine) has showed that R74 (neurology and psychiatry) and R76 (otolaryngology science) have stronger correlation than with the other sub-categories. Given that, we try to make some text mining on such cognitive fuzzy super sub-category both from the citing and cited direction. And we hope to understand the similarity between these sub-categories from the perspective of semantic.

Specifically this study has tried to extract the research topic embedded in the title and keywords of the bibliographies pair which have bibliographical coupling relationship and the cited bibliographies having same citing bibliography in the super sub-category, and then topic detection is accomplished through the establishment of LDA (Latent Dirichlet Allocation) topic model in text mining area. By learning from the research topics of the sub-category that have strong coupling strength, we try to find the cause of such phenomenon, and expect to provide reference for the development of subject-classification study.

From Table 5, we find that research point of the citing bibliography pairs in coupling data related to sub-category R74 and R76 have great consistency and can form to complete topics; most are researches such as swallowing disorder after stroke, respiratory sleep disorders and related swallowing dysfunction complications. This is true with the research topic of the cited bibliographies in coupling data, which are also about Swallowing disorder after stroke. And the reason why these two sub-categories have showed stronger similarity may be the nerve function research (in R74) involved in the process of the otolaryngology disease treatment (in R76).

Research topics coming from the coupling data of the sub-category TD and TU subject in Table 6 show that, research topics of the citing bibliography pairs mainly focus on the application of rock mechanics theories from TU (building science) subject to TD's coal mine construction, coal mining, experimental study of coal mine safety; the application of other architectural theory belonging to TU subject to the study of coal mine design in TD subject. Moreover the research topics extracted from the cited bibliographies are coincident with those from citing bibliography pairs.

Through the above text level study of the two super sub-category we find that, firstly, the subject clustering study based on journal coupling has practical significance, the research topics in these two super sub-category all show that these research topics are consistent with each other, that is, from the perspective of knowledge and semantic, subjects that are highly journal coupling related are also correlated in their research content, then journal-based coupling can reflect the correlation between subjects. Secondly, in interpretation of the newly clustered subjects, research topics extracted from the citing bibliography pairs in coupling data are consistent with those derived from the cited bibliographies, which indicates both the cited bibliographies and the citing bibliography pairs in coupling data have the ability to explain and label newly constructed subject. But compared with the cited bibliographies in coupling data, the citing bibliography pairs have

Table 5 Research topic of R74 (neurology and psychiatry) and R76 (otolaryngology science)

Data source	Research topics
Citing bibliography pairs in coupling data	Meniere's disease; BPP; Anxiety self-assessment scale; Depression self rating scale; vertigo
	Cells; interleukins; rhinitis; allergic; Sudden deafness
	Validity; reliability; Swallowing disorder evaluation; stroke; swallowing disorder
	Fibrinogen; tinnitus; transient ischemic attack; children; evoked potentials
	Cerebral infarction; obstructive sleep apnea hypopnea syndrome; continuous positive airway pressure; transient ischemic attack; predict
	The quality of life; rhinitis; allergic; tomography; line computer
	Sensorineural; hypertensive cerebral hemorrhage; deaf; hearing loss; large vestibular conduit syndrome
	Stroke; swallowing disorder; related factors; line fluoroscopic examination; fiber nose throat swallowing function
	Newborn; gene; congenital cytomegalovirus infection; fluorescence quantitative polymerase chain reaction; acute cerebral infarction; deafness
	Obstructive; sleep apnea; subarachnoid hemorrhage; magnetic resonance imaging (fmri); cerebral vasospasm
	Stroke; swallowing disorder; acute cerebral infarction; line fluoroscopic examination; related factors
	Rehabilitation treatment; the brain stem swallowing disorder after stroke; cerebral infarction; ischemic stroke; stroke
	Swallowing disorder; stroke; video swallowing angiography examination; ring pharyngeal muscle relaxation loss; polymorphism
	Complications; depression after stroke; magnetic resonance imaging (fmri); the treatment results; cochlear implantation
	Dizzy; vertigo; Balance function; Blink reflex; tinnitus
	Stroke; swallowing disorder; acute cerebral infarction; related factors; fiber nose throat swallowing function
	Swallowing disorder; stroke; neuromuscular electrical stimulation; rehabilitation training; rehabilitation
	Cerebralinfarction; hydrogensulfide; hyperbaricoxygen; cystathionine; cerebrovascular disease
Cited bibliographies in coupling data	Stroke; swallowing disorder; treatment; clinical; swallowing; function; curative effect; comprehensive rehabilitation
	Deafness; hearing; syndrome; diagnosis; gene; mutation; children; characteristics; brain paralysis; curative effect
	Diagnostic point; cerebrovascular disease; sleep; breathing; suspended; syndrome; blocking; nerve; ventilation
	Ear; worm; artificial; implant; image; hair cell death; nerve; deformity; rat; effect of rehabilitation
	Protein; animal; facial nerve; damage; rat; fiber; brain; express; function; reaction; change
	Facial nerve; reaction; brain; Listen; surface; chirurgery; diagnosis; tumor; bone; collapsed; treatment; artery; short; clinical; pharyngeal muscle

Table 6 Research topic of TD (mining engineering) and TU (building science)

Data source	Research topics
Citing bibliography pairs in coupling data	Rock mechanics; the numerical simulation; Similar material; neural network; anchor
	Unified strength theory; intermediate principal stress; analytical solutions; elastic–plastic analysis; factors affecting
	Rock mechanics; deep mining; constitutive model; unloading; high temperature
	Numerical simulation; anchor; stress distribution; anchor cable; prestressed anchor cable
	Triaxial compression; rock mechanics; creep property; numerical simulation; soft rock
	Acoustic emission; rock mechanics; rock; damage; uniaxial compression
	Mining engineering; high stress; directional fracture; rock; quick drilling and blasting
	Numerical simulation; mine earthquake; stability evaluation; key layer; micro seismic
	Deep roadway; partition burst; deep rock mass; support pressure; shear sliding failure theory
	Model; creep; pile; stack preloading; critical sedimentation method
	Rock mechanics; crispy; soft rock; dissipation can; shape pit
	Compressive strength; rock mechanics; pore pressure; failure process; mining engineering
	Stress; acoustic emission; mining engineering; confining pressure; strain curve
	Creep; creep model; rock mechanics; coal and rock; distribution
	In-situ stress measurement; Rock burst; Rock mechanics; Hydraulic fracturing method; In-situ stress
	Rock mechanics; blasting vibration; deep rock mass engineering; Dynamic problems; volume strain
	Rock burst; mining engineering; mined-out area; numerical simulation; type of pit
	Safety factor; strength subtraction; slope stability; slope; rock mechanics
	Mining engineering; permeability; rock mechanics; gas seepage; temperature
Cited bibliographies in coupling data	Tunnel; deep; foundation pit; construction; underground; excavation; surrounding rock; buried; engineering
	Features; test; rock; test; shaft; deformation; intensity; mechanics; soil
	Coal; stress; seepage; impact; crack; Rock mass; strain; penetration; features
	Rock; Under the; damage; Acoustic emission; deformation; rupture; test
	Impact; blasting; concrete; damage; reinforced; Mined-out area; stability; cracks; speed
	Foundation; pile; soft; layer; soil; composite; model; strengthening; calculate; grouting
	Structure; strain; nonlinear; elastic–plastic; mechanics; consider; surrounding rock
	Landslide; sudden; mechanism; floor; Coal mine; monitoring; Optical fiber; forecast; hydraulic
	Predict; evaluation; Rock mass; model; Mined-out area; Rock burst; T quality
	Deep; In-situ stress; surrounding rock; roadway; stress; engineering; in-situ stress measurement; mechanics; mine
	Supporting; roadway; anchor; control; technology; soft rock; surrounding rock; strengthening; parameter; anchoring; deep

stronger explanation ability to the newly clustered subjects due to richer data and relationship it contains. Finally, the correlation of each subject is in changing, and unexpected similarity between the above two subject pairs has reflected development of scientific structure. Also the text level study of these two super subjects indicates that the direction of scientific research is always to solve practical problem.

Conclusion and discussion

The development of science has made the quantitative description of science structure developing. Subject-classification is an important part of scientific structure research, previous studies that related to subject system or specialties mostly focused on making some quantitative study, and gave some comparative research of these newly calculated systems with existing expert system, and take the quantitative results and the existing qualitative result separately. On the basis of previous studies, this paper combines qualitative research and quantitative research in the process of subject-classification system study, and proposes an improved subject-classification system which adds scientific development into expert classification. We expected our study will serve as a new resource for the future study of scientific structure research and for the practical application, such as better journal subject assignment, more appropriate scientific output calculation and analysis.

In our study, similarities between subjects are measured through the journal-based coupling strength; further the relationships of subjects are analyzed with the help of clustering analysis. The subject system constructed in this paper is grounded on the expert system CLC, and the study covers the whole 10 nature science related top categories, so the result subject-classification system has universal applicability in the field of natural science.[6] The cross-citation data in this study is from CSCD, and utilization of journal coupling method fully ensures the sample size of the analysis. The relationship between subjects are studied through clustering analysis, different from previous studies, the determination of the whole clustering scheme takes the characteristics of data and clustering methods into consideration, and the final classification result is a ensemble learning consequence of the those selected clustering methods, which ensure the stability and accuracy of clustering results. Finally, an improved subject-classification system that conforms to previous experience and cognitive as well as combines subject development knowledge is proposed in our study.

Further, for some cognitive fuzzy super sub-categories appeared in the clustering process, we conduct topic mining to explain the result from the perspective of semantic, and we find these cognitive fuzzy subjects indeed have crossover research point, which result in two subjects' strong similarity. Furthermore, in the process of trying some semantic level interpretation of such super subject, we also find that the interpretation derived from the citing bibliography pairs in coupling data and cited bibliographies in coupling data has consistency. When it comes to the explanation ability, the citing bibliography pairs may perform better than cited bibliographies in coupling data, which gives demonstration on the

[6] We believe the subject-classification we derived in this paper is applicable to other situation for the reason that the journals in CSCD source list are all nature science related core journals. And according to Garfield's Law of Concentration, the citation behavior of these core journals have strong representation, so the modified subject system based on citation can be commonly adopted by situations using CLC to some extent.

selection of interpretation source for labeling the newly calculated subjects in the previous studies (Chen et al. 2010).

Though puts forward a revised subject-classification system, this paper also have some shortcomings Firstly, in the subject relationship study rather than take the whole citation network into account, it only using a kind of one-step correlation in network, information loss may affect the accuracy of results to some extent. Secondly, semantically similarity research among different subjects is not included. Finally, not breaking the top category of CLC may bring some constraints for discovering new interdisciplinary. In the future, we hope combining citation behavior with Text Mining and Natural Language Processing (NLP) technique, and study the development of subject-classification system from the aspects of citation and semantic at the same time. We also want to conduct some further research in terms of finding a wider range of interdisciplinary. And we hope to study the development of scientific structure more scientifically, and then give some reference to future researches on science structure and development.

Appendix 1

See Table 7.

Table 7 Original subject-classification system of CLC

Top category	Sub-category
N (general natural science)	N0 (natural science theory and methodology)
	N1 (present situation and development of natural science)
	N2 (natural science institutions and organizations, meeting)
	N3 (methodology for natural science)
	N4 (science education and popularization)
	N5 (Natural science series, corpus)
	N6 (natural science reference books)
	N7 (natural science literature retrieval reference books)
	N79 (book materials, audio-visual materials)
	N8 (natural science investigation and inspection)
	N91 (Nature study, natural history)
	N93 (nonlinear science)
	N94 (system science)
	N99 (Information science, intelligence work)
O (mathematical science and chemistry)	O1 (mathematic)
	O3 (mechanics)
	O4 (physics)
	O6 (chemistry)
	O7 (crystallography)

Table 7 continued

Top category	Sub-category
P (astronomy, earth science)	P (comprehensive astronomy, earth science)
	P1 (astronomy)
	P2 (topography)
	P3 (geophysics)
	P4 (atmospheric science (meteorology))
	P5 (geology)
	P7 (oceanography)
	P9 (physical geography)
Q (Biological sciences)	Q (comprehensive biology)
	Q-(Q-0 theories and methods of biology science, Q-1 status and development of biological sciences, Q-3 biological scientific research methods and technology, Q-4 biological science education and popularization, Q-9 biological resources survey)
	Q1 (general biology)
	Q2 (cytobiology)
	Q3 (genetics)
	Q4 (physiology)
	Q5 (biochemistry)
	Q6 (biophysics)
	Q7 (molecular biology)
	Q81 (biological engineering (biotechnology))
	Q89 (environmental biology)
	Q91 (paleontology)
	Q93 (microbiology)
	Q94 (botany)
	Q95 (zoology)
	Q96 (entomology)
	Q98 (anthropology)
R (medicine)	R (comprehensive medicine)
	R1 (preventive medicine, hygiene)
	R2 (Chinese medicine)
	R3 (preclinical medicine)
	R4 (Clinical medicine)
	R5 (internal medicine)
	R6 (chirurgery)
	R71 (gynecotokology)
	R72 (pediatrics)
	R73 (oncology)
	R74 (neurology and pathergasiology)
	R75 (dermatology and venereology)
	R76 (otorhinolaryngology)
	R77 (ophthalmology)

Table 7 continued

Top category	Sub-category
	R78 (stomatology)
	R79 (foreign national medicine)
	R8 (special medicine)
	R9 (pharmacy)
S (agricultural science)	S (comprehensive agricultural science)
	S-(S-0general theory, S-1present situation and the development of agricultural science and technology, S-3agricultural science research and experiment, S-9agricultural economy)
	S1 (agricultural basic science)
	S2 (agricultural engineering)
	S3 (agriculture (agronomy))
	S4 (plant protection)
	S5 (crops)
	S6 (gardening)
	S7 (forestry)
	S8 (animal husbandry, animal medicine, hunting, silkworms and bees)
	S9 (aquaculture, fishing)
T (Industrial Technology)	T-(T-0 theory of industrial technology, T-1present situation and the development of technology, T-2Institutions, organizations, conferences, T-6reference book, T-9industrial economy)
	TB (general industrial technology)
	TD (mineral engineering)
	TE (oil and gas industry)
	TF (metallurgical industry)
	TG (metallographic and metal crafts)
	TH (machinery and instrument industry)
	TJ (arms industry)
	TK (energy and power engineering)
	TL (atomic energy technology)
	TM (electro-techniques)
	TN (electronic technology, communication technology)
	TP (automation technology, computer technology)
	TQ (chemical industry)
	TS (light industry, handicraft industry and living services)
	TU (building science)
	TV (hydraulic engineering)

Table 7 continued

Top category	Sub-category
U (Communications and transportation)	U-9 (transportation economy)
	U1 (integrated transportation)
	U2 (railway transportation)
	U4 (highway transportation)
	U6 (waterway transportation)
	U8 (air transport)
V (aeronautics and astronautics)	V1 (aviation, space technology research and exploration)
	V2 (aviation)
	V4 spaceflight (astronavigation)
	V7 (aerospace medicine)
X (Environmental science and safety science)	X-(X-0 environmental scientific theory, X-1 environmental science and development, X-2 environmental protection organizations, institutions, meeting, X-4 environmental protection publicity, education and popularization, X-6 environmental protection reference books)
	X1 (Environmental science basic theory)
	X2 (Society and Environment)
	X3 (environmental protection and management)
	X4 (disaster and prevention)
	X5 (environment pollution and protection)
	X7 (industry pollution, waste disposal and comprehensive utilization)
	X8 (environmental quality assessment and environmental monitoring)
	X9 (safety science)

References

Ahlgren, P., & Colliander, C. (2009). Document–document similarity approaches and science mapping: Experimental comparison of five approaches. *Journal of Informetrics, 3*(1), 49–63. doi:10.1016/j.joi.2008.11.003.

Archambault, É., Beauchesne, O. H., & Caruso, J. (2011). Towards a multilingual comprehensive and open scientific journal ontology. In E. C. M. Noyons, P.Ngulube, & J. Leta (Eds.), *Proceedings of the 13th international conference of the international society for scientometrics and informetrics* (pp. 66–77).

Börner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., Light, R. P., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS One, 7*(7), e39464. doi:10.1371/journal.pone.0039464.

Boyack, K. W., Klavans, R., & Börner, K. (2005). Mapping the backbone of science. *Scientometrics, 64*(3), 351–374.

Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis: I: Structural Aspects. *Journal of the American Society for Information Science and Technology, 42*(4), 233–251.

Cason, H., & Lubotsky, M. (1936). The influence and dependence of psychological journals on each other. *Psychological Bulletin, 33*(2), 95–103.

Chang, Y. F., & Chen, C.-M. (2011). Classification and visualization of the social science network by the minimum span clustering method. *Journal of the American Society for Information Science and Technology, 62*(8), 2404–2413.

Chen, C. M., Ibekwe-SanJuan, F., & Hou, J. H. (2010). The structure and dynamics of co-citation clusters: A multiple-perspective co-citation analysis. *Journal of the American Society for Information Science and Technology, 61*(7), 1386–1409.

Daniel, R. S., & Loutitt, C. M. (1953). *Professional problems in psychology*. New York: Prentice Hall.

Everitt, B. (1974). *Cluster analysis*. London: Heinemann Educ.

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

Gómez-Núñez, A. J., Batagelj, V., Vargas-Quesada, B., Moya-Anegón, F., & Chinchilla-Rodríguez, Z. (2014). Optimising SCImago journal & country rank classification by community detection. *Journal of Informetrics, 8*(2), 369–383.

Gómez-Núñez, A. J., Vargas-Quesada, B., & Moya-Anegón, F. (2015). Updating the SCImago journal and country rank classification: A new approach using Ward's clustering and alternative combination of citation measures. *Journal of the Association for Information Science and Technology,67*(1), 178–190.

Hartigan, J. A., & Wong, M. A. (1979). A K-means clustering algorithm. *Applied Statistics, 28*(1), 100–108.

Katz, J. S., & Hicks, D. (1995). The classification of interdisciplinary journals: A new approach (Version 2.0). In M.E.D. Koenig & A. Bookstein (Eds.), *Proceedings of the Fifth Biennial Conference of the International Society for Scientometrics and Informatics* (pp. 245–254). Medford: Learned Information.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. New York: Wiley.

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

Kronegger, L., Mali, F., & Ferligoj, A. (2013). Classifying scientific disciplines in Slovenia: A study of the evolution of collaboration structures. *Journal of the American Society for Information Science and Technology, 66*(2), 321–339.

Leydesdorff, L. (2002). Dynamic and evolutionary updates of classificatory schemes in scientific journal structures. *Journal of the American Society for Information Science and Technology, 53*(12), 987–994.

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 term of aggregated journal—Journal citation relations using the journal citation reports. *Journal of the American Society of Information and Technology, 57*(5), 601–603.

Leydesdorff, L., & Cozzen, S. E. (1993). The delineation of specialties in terms of Journals using the dynamic journal set of the SCI. *Scientometrics, 26*(1), 135–156.

Leydesdorff, L., & Rafols, I. (2008). A global map of science based on the ISI discipline categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348–362.

Leydesdorff, L., & Rafols, I. (2012). Interactive overlays: A new method for generating global journal maps from Web-of-Science data. *Journal of Informetrics, 6*(2), 318–332. doi:10.1016/j.joi.2011.11.003.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, (pp. 281–297) University of California Press, Berkeley, Calif.

Marshakova, S. I. (1973). System of Document Connections Based on References. *Scientific and Technical Information Serial of VINITI, 6*(2), 3–8.

Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. Washington, DC: National Science Foundation.

Narin, F., Carpenter, M., & BerltN, C. (1972). Interrelationships of scientific journals. *Journal of the American Society for Information Science, 23*(5), 323–331.

Ni, C., Sugimoto, C. R., & Jiang, J. (2013). Venue-author-coupling: A measure for identifying disciplines through author communities. *Journal of the American Society for Information Science and Technology, 64*(2), 265–279.

Qiu, J., & Dong, K. (2013). A Comparative study on the ability of author co-occurrence network in revealing scientific structure. *Journal of library science china, 39*(1), 15–24. **(In Chinese)**.

Qiu, J., & Liu, G. (2014). Research of discipline knowledge aggregation based on the journal-author coupling method. *Journal of intelligence, 33*(4), 17–22. **(In Chinese)**.

Reynolds, A., Richards, G., de la Iglesia, B., & Rayward-Smith, V. J. (1992). Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modeling and Algorithms, 5*(4), 475–504.

Rousseau, R., & Zuccala, A. (2004). A classification of author co-citations: Definitions and search strategies. *Journal of the American Society for Information Science and Technology, 55*(6), 513–529.

Small, H. (1973). Co-citation in the Scientific Literature:A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science, 24*(4), 265.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B, 63*(2), 411–423.

Waltman, L., & Van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the Association for Information Science and Technology, 63*(12), 2378–2392. doi:10.1002/asi.22748.

White, H. D., & McCain, K. W. (1998). Visualizing a Discipline: An Author Co-Citation Analysis of Information Science, 1972–1995. *Journal of the American Society for Information Science, 49*(4), 327–355.

Zhang, L., Janssens, F., Liang, L., & Glänzel, W. (2010). Journal cross-citation analysis for validation and improvement of journal-based discipline classification in bibliometric research. *Scientometrics, 82*(5), 687–706.

Zhang, L., Liang, L., Liu, Z., & Glänzel, W. (2012). The analysis of science structure based on journal clustering and SOOI classification system. *Study in science of science, 30*(9), 14–22. **(In Chinese)**.

Zhao, D. Z., & Strotmann, A. (2008). Evolution of research activities and intellectual in information science 1996–2005: Introducing author bibliographic -coupling analysis. *Journal of the American Society for Information Science and Technology, 59*(13), 2070–2086.