Scientometrics (2016) 109:1725–1759
DOI 10.1007/s11192-016-2146-4

A study on construction and analysis of discipline knowledge structure of Chinese LIS based on CSSCI

Hao Wang[1] · Sanhong Deng[1] · Xinning Su[1]

Received: 26 April 2016 / Published online: 4 October 2016

Abstract This study proposes a method to automatically establish a narrow-sense knowledge structure for Chinese Library and Information Science (CLIS) using data from the Chinese Social Science Citation Index. The method applies multi-level clustering, using ontological ideas as theoretical guidance and ontology learning techniques as technical means. Knowledge categories generated are checked for cohesion and coupling through hierarchical clustering analysis and multidimensional scaling analysis in order to verify the accuracy and rationality of the narrow-sense knowledge structure of CLIS. Finally, the narrow-sense knowledge structure is expanded to a broad sense. Using scholars as objects in examples, this study discusses the semantic associations between topic knowledge and the other academic objects in CLIS from the micro-, meso-, and macro-levels, so as to fully explore the broad-sense knowledge structure of CLIS for knowledge analysis and applications.

Keywords Chinese Library and Information Science (CLIS) · Discipline knowledge structure (DKS) · Chinese Social Science Citation Index (CSSCI) · Multi-level clustering (MLC) · Hierarchical clustering analysis (HCA) · Multidimensional scaling analysis (MDSA) · Social network analysis (SNA)

Instruction

In a narrow sense, the knowledge of a discipline refers to the confident identification of the topics or content of the discipline. This concept represents the main driving force for discipline development and the core content for discipline innovation. The narrow sense of discipline knowledge can be revealed in two ways. One is to organize and describe the

✉ Hao Wang
ywhaowang@nju.edu.cn

[1] School of Information Management of Nanjing University, Nanjing 210023, China

static structure of knowledge (Song and Kim 2013), exploring the semantic association between different disciplines' knowledge in a certain period, so as to lay the foundation for displaying discipline knowledge distribution on maps and creating reasonable knowledge applications. The other is to imitate and track the evolution of knowledge (Chang 2012) and discuss the laws governing knowledge development and the corresponding causes of discipline knowledge nodes during a certain historical period, so as to instruct discipline researchers in knowledge innovation.

Clarifying the structure and evolution of narrow-sense discipline knowledge advances understanding of discipline connotations and motivates discipline research so as to perfect research content and promote innovation and development in the discipline. The process of clarification also contributes to exploring the semantic associations between discipline knowledge and other academic resources in the discipline, including scholars, institutions, and journals, so as to reveal their knowledge structures and distribution, and even mine and depict their potential associations. All of these measures help to outline the overall connotations of the discipline. In fact, such clarification forms the structure of discipline knowledge in a broad sense. Figure 1 illustrates divisions of distinct granularity knowledge in the Chinese Social Sciences Citation Index (CSSCI) and their relationships. In a narrow sense, "Keyword" represents the discipline knowledge, and the combination of keywords from different scopes and in different numbers form the structure of discipline knowledge. However, in a broad sense, "Scholar", "Institution", "Area", "Article" "Journal", and "Discipline" all stem from "Keyword" and represent distinct granularities in discipline knowledge. Therefore, the discussion of discipline knowledge structure in a narrow sense can be gradually expanded to a broad sense.

This study attempts to automatically establish a narrow sense of knowledge structure in the field of Chinese Library and Information Science (CLIS) using data from CSSCI. In order to achieve this goal, this study applies the method of multi-level clustering (MLC) with ontological ideas for theoretical guidance, using ontology learning (OL) techniques as technical means. Furthermore, knowledge categories (KCs) generated are checked for cohesion and coupling according to the methods of hierarchical clustering analysis (HCA) and multidimensional scaling analysis (MDSA) in order to verify the accuracy and rationality of the narrow-sense CLIS knowledge structure. Finally, the knowledge structure is expanded from the narrow sense to a broad sense; taking scholarly objects as examples, the semantic associations between topic knowledge and the other academic objects in CLIS

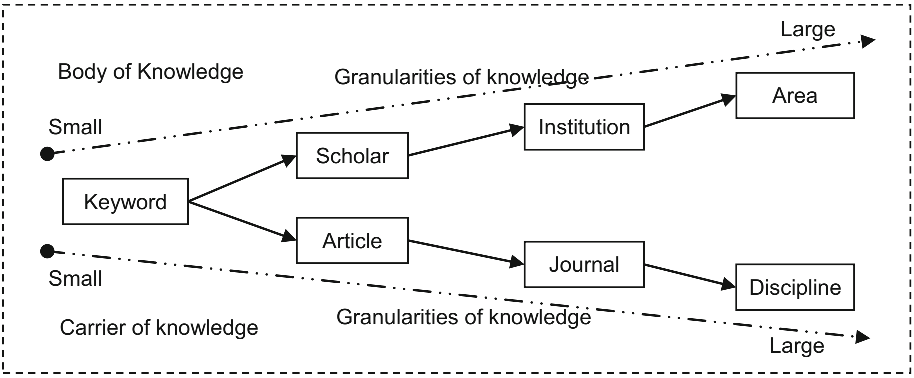


Fig. 1 Distinct granularities knowledge in CSSCI

are discussed from the micro-, meso-, and macro-levels, so as to fully explore the broad-sense knowledge structure of CLIS (CLIS_KS) in knowledge analysis and applications.

Related research

In current research, discipline knowledge structure (DKS) and its characteristics are discussed in two primary ways. First, domain experts subjectively integrate and qualitatively describe DKS according to their own background knowledge, research experience, and previous research results (Powers 1995; Sluyter et al. 2006; Hooper 2009). Second, DKS is objectively described and quantitatively analyzed based on bibliometric methods, either to explore various kinds of domain-oriented evaluation indexes for calculation and statistical description (Su 2007; Wang et al. 2014; Erserim 2016), or to reveal interactive associations between various research units in the domain (White and McCain 1998; Yoo et al. 2013; Ravikumar et al. 2015).

As bibliometrics have gradually matured, researchers are increasingly able to detect research hotspots and interrelations in their disciples through existing objective relationships between academic objects (Gonzalez-Alcaide et al. 2008; Galvagno 2011; Yan et al. 2015; Machado et al. 2016). This involves the following basic calculation mode: First, select data from a specific discipline; second, describe the academic units with vectors according to their objective relationships, such as co-word and co-citation; and third, cluster similar academic units in order to generate domain hotspots, with the goal of eventually demonstrating the interactive associations among these hotspots and their changes in a divided period with structural diagrams.

The above process includes the following characteristics:

The data for analysis comes from a plurality of types of data sources, including books (Torres-Salinas and Moed 2009), journal articles (Ma 2012; Garcia-Lillo et al. 2016), dissertations (Prebor 2010), letters, reviews, conference papers (Kurihara et al. 2013) and other academic events such as workshops, symposia, seminars, etc. (Jeong and Kim 2010).

The knowledge nodes with large granularity, such as directions and hotspots, in a discipline or a field even a journal, are all formed through academic units clustering together based on certain rules, in which the academic units may be words or terms (Hu et al. 2013; Darvish and Tonta 2016), authors (Chen and Lien 2011; Riviera 2015), articles (Pilkington and Meredith 2009; Hult 2016), journals (Pratt et al. 2012; Machado et al. 2016), etc., the similarities among academic units are the knowledge nodes, and the criterions judging similarities mainly are the relation of co-occurrence, including co-citation (Leydesdorff and Vaughan 2006; Pilkington and Meredith 2009; Chen and Lien 2011) and co-word (Zong et al. 2013; Yang et al. 2016) between the academic units.

The methods for gathering academic units based on similarities mainly include Hierarchical Cluster Analysis (HCA) (Liu 2005; Triventi 2014), Multidimensional Scaling Analysis (MDSA) (Calabretta et al. 2011; Wolfram and Zhao 2014), Factor Analysis (FA) (McCain 1990; Charvet et al. 2008; Hossain et al. 2013). The methods for further describing the association structure of knowledge nodes with large granularities primarily are social network analysis (SNA) (Otte and Rousseau 2002; Park and Leydesdorff 2008; Aleixandre et al. 2015), Pathfinder Net Analysis (PF-Net Analysis) (Kim and Lee 2008; Ma 2012) and Visualization of Similarities (VOS) (Olijnyk 2015; Pinto 2015).

The issues for discussing are mostly about narrow-sense subject or domain KS (Charvet et al. 2008; Dehdarirad et al. 2014; Naghizadeh et al. 2015), seldom with academic units such as scholars or institutions as main objects. In additional, the knowledge mainly refers to the knowledge nodes with large granularities, namely the research directions or hotspots in disciplines or domains (Tseng and Tsay 2013; Danell 2014; Rusk and Waters 2015). The discipline scope includes LIS (Tseng and Tsay 2013; Milojevic et al. 2011), Biomedical Informatics (Jeong and Kim 2010), E-Learning (Chen and Lien 2011), Strategic management (Nerur et al. 2008), International Marketing (Samiee and Chabowski 2012), Theology (Yoo et al. 2013), etc.

A variety of tools are used to analyze the knowledge structure, including Bibexcel (Persson et al. 2009) which is able to build co-citation/co-occurrence matrix and calculate the correlation coefficient, SPSS (Accessed by July 1st, 2015) and SAS (Accessed by July 1st, 2015) which are able to achieve HCA, MDSA and FA, Ucinet (Borgatti et al. 2002), Pajek (de Nooy et al. 2005), ORA (Meyer et al. 2011) and pathfinder algorithm (Chen and Paul 2001) used to construct correlation between knowledge nodes and to calculate characteristics of KS, VOSviewer (Van Eck and Waltman 2010) was able to implement clustering the co-occurrence matrix and visualize the clusters with density view and network view, and CiteSpace (Chen 2006; Seyedghorban et al. 2016) used to analyze the distribution of association between knowledge nodes and evolution with timeline.

Current explorations of CLIS and its characteristics have exposed problems of small scale, incompletion, and partial scope. The methods of HCA, MDSA, and FA have been extensively employed to generate discipline knowledge nodes in research, but the amount of data applied in these methods has been impossibly large; for example, HCA is a typical small-scale and high-precision clustering method. Additionally, the knowledge nodes generated are all research directions and hotspots for subjects or domains, with large granularity. In other words, the so-called DKS only refers to top-level knowledge categories and their associations, and analysis of the detailed circumstances in categories has been inadequate. The descriptions of semantic relationships between knowledge nodes with distinct granularities have therefore remained incomplete. Finally, discussions based on parts of KS are partial in their scope, which can only reveal the status of a given aspect or angle in a certain discipline. For instance, analysis of DKS based on high-citation authors or high-frequency terms can only yield the research directions in which these scholars dabble or which are described by this terms, rather than the KS of the entire discipline (Milojevic et al. 2011; Ma and Ni 2011).

To address these shortcomings, this study attempts to view DKS as a hierarchy of domain knowledge and to comprehensively analyze the semantic associations between knowledge nodes in CLIS from an ontological point of view, so as to construct a relatively complete narrow-sense KS to lay the knowledge foundation for exploring a broad-sense CLIS_KS. The novelty of this study lies in its attempt to build a relatively complete multi-level KS for the whole CLIS, on the basis of which some analysis and applications can be carried out. This requires the collection, processing, calculation, and analysis of large-scale data covering the entire discipline. In contrast, previous studies of DKS have mainly stayed in partial hotspots and two-levels structures of disciplines or domains. In addition, the introduction of the MLC method can solve the problem of large-scale data processing, which cannot be addressed by HCA because it focuses on precision.

Methodology

Research framework

The basic idea of this study is summarized graphically in Fig. 2. The entire research framework can be divided into four phases. First is the data pretreatment phase, which selects and cleans the LIS academic resources. Using CSSCI (2003–2012) as a data source, the CLIS document records are selected, and then the core keywords and distinct scholars who made a certain contribution to the discipline are identified through data-cleaning; this process forms triples formatted as <Keyword, Scholar, Weight>. Notably, when the task of data collection begins, the CSSCI (2013–2014) data has not yet been completed.

Second is the KS construction phase, in which the CLIS knowledge ontology is generated and stored. This process, taking the concept of ontology as theoretical guidance, converts the triples into a keyword-scholar matrix (KSM), with the keyword serving as an object, the scholar serving as the description factor for the keyword, and the correlation coefficient between them serving as the matrix value. Then, the keywords' hierarchical structure is constructed using the MLC method, so as to generate an ontology for the discipline that can be described and stored using Ontology Web Language (OWL) and graphical visualization. This lays the foundation for further applications of this knowledge ontology.

Third is the KC validation phase, which examines the correctness and rationality of CLIS_KC. Through modular reasoning, the cohesion and coupling aspects of KC demonstrate the quality of CLIS_KS. Cohesion confirms the degree of internal aggregation in parts of categories through the HCA method in order to discuss the possibility of further subdividing categories; meanwhile, coupling detects the spatial distribution of keywords belonging to different categories in order to discuss the discrimination of KCs and the rationality for dividing keywords.

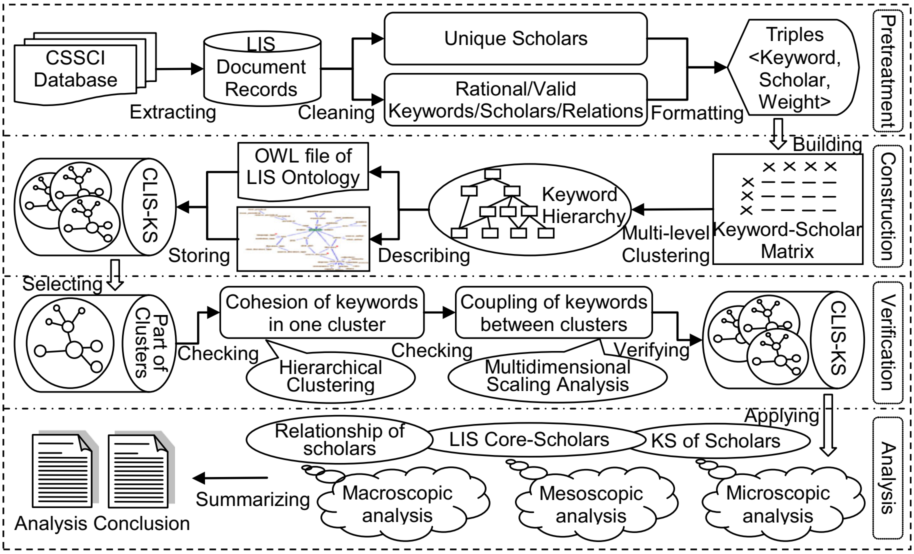


Fig. 2 Research framework of CLIS_KS

Fourth is the CLIS analysis phase, which explores the internal relationships among CLIS academic objects. As constructed in this study, CLIS_KS only refers to the hierarchical relationships among the contents of discipline research, reflecting the scope of narrow-sense DKS. This DKS can provide a wealth of knowledge reserves for analyzing and evaluating other academic objects in CLIS. This study proposes to detect academic objects on the microscopic, mesoscopic, and macroscopic levels. Taking the objects in the scholar category as an example, on a microscopic level, this study builds scholars' individual KS in order to analyze their research focuses; on a mesoscopic level, this study detects the core groups of scholars in each large category in CLIS in order to evaluate the distribution of important scholars in each research direction in CLIS; and on a macroscopic level, this study mines the content cross and study dependence between scholars in order to identify the interactions between scholars within a certain range from the point of view of research content. Of course, the methods for discussing scholars' research content could apply equally to other academic objects, such as journals, institutions, and areas; therefore, this study provides sufficient methodological support and a referential analysis concept for the construction of broad-sense CLIS_KS.

Methods for CSSCI data cleaning

Journal articles constitute the main carriers and sources of discipline knowledge. They not only contain the existing foundations of discipline knowledge, but also represent the important process of recording and sharing new discipline knowledge. CSSCI, widely recognized as China's leading, authoritative, and comprehensive database for scholarly citations in Chinese, includes more than 500 high-impact and high-quality journals in the fields of the humanities and social sciences, selected by the Social Sciences Evaluation Center of Nanjing University. This database includes articles, keywords, authors, cited documents, and other academic resources; therefore, it curates the most cutting-edge and complete knowledge of social science disciplines. This study intends to employ the inherent knowledge held in CSSCI journal articles in order to construct CLIS_KS.

A bibliographic record in the CSSCI database consists of multiple fields that describe the record, among which are title, keyword, and cited document. These fields contain subject terms or keywords used to describe the subject contents of an item. Due to limitations in the precision of Chinese word segmentation, the word group or phrase representing knowledge may be subdivided, and the title, citation, and other fields representing discipline knowledge may not be selected as knowledge sources. For this reason, this study use the keyword field with wide content coverage, significant segmentation signs and explicit knowledge as the data source for discipline knowledge.

However, directly using the original data as the experimental sample causes many problems, such as scholars having the same name, topics not belonging to the discipline, edge scholars, and accidental associations between knowledge nodes and scholars. Therefore, it is necessary to clean the original data in order to obtain more standardized and moderate-scale experimental samples. CSSCI data cleaning includes two main tasks, which are described in the following paragraphs.

The first is to identify distinct authors. We considered using name and area (the first two digits of the area code in CSSCI) as unique marks to distinguish scholars; however, it turns out that some individual scholars are associated with multiple marks. Some of this duplication is caused by indexing error. For example, the scholar "JP Qiu" from "Wuhan University" (42) was incorrectly indexed as "Xiangtan University" (43) twice, and the same scholar was also incorrectly indexed as "Chinese Academy of Social Sciences" (11) twice. Other instances of duplication are due to interprovincial changes in a scholar's workplace. For example, the scholar "CJ Suo" from the "Chinese National Library" (11) previously worked in "Zhengzhou University" (41) for a long time, and the scholar "X Xu" of "East China Normal University" (31) participated in doctoral studies at "Nanjing University" (32). In these cases, the area symbols do not help to distinguish authors with the same name; rather, they contribute to data corruption. Ultimately, we decided to employ "full name" as a mark to distinguish scholars, such that if scholars with the same name remain after data filtering, they are separated manually. The method proceeds as follows: All of the possible institutions and corresponding titles are extracted from the source data for the selected scholars, and then scholars with the same name are judged to be the same person based on their research content, supplemented by search engines that specifically check full names with common or short surnames. Fortunately, among all selected scholars in CLIS, only a dozen groups of scholars with the same name were identified. In fact, renowned scholars with the same name are actually very rare within a given discipline. Undeniably, this method introduces a certain subjectivity and therefore reduces the degree of automation possible; however, this is necessary to ensure precision in the case of without scholar ontologies.

The second task is to screen data. In order to ensure that non-domain topics, edge scholars, and accidental associations have little influence on the generation of domain knowledge and its structure, this study considers screening methods from three perspectives: keywords, scholars, and associations between them. Therefore, we set the following five parameters to ensure the rationality and scale of the selected data, with the main purpose of getting rid of noise data. (1) Keyword frequency factor (K). In order to ensure that the selected keywords cover a long study period and maintain a certain degree of novelty, we established the two frequency factors K10 and K5, which are used to control the number of keywords appearing in the past 10 and 5 years, respectively. When the two factors of a keyword exceed the thresholds T1 and T2, this keyword is accepted by the domain; in other words, the keywords selected need to satisfy formulas (1) and (2). However, this study did not standardize the selected keywords, because no related dictionary exists. This operation may affect the accuracy of clustering, such that synonymous keywords with different forms may be classified in different categories. However, within the same discipline, scholars tend to use certain established keywords. (2) Scholar issuing factor (A). The issuing number is only counted for the first author. Similar to the above method, the scholar is assumed to be an important scholar in the discipline if his or her total of recent issuing numbers reaches a certain level; this excludes a large number of scholars who make small contributions to the domain of CLIS. Therefore, we established the factors A10 and A5, and the selected scholars must meet the conditions of formulas (3) and (4), in which T3 and T4 are constants. It is worth noting that A10 and A5 are not evaluation indices measuring the influence of a scholar. Rather, they serve as a basis for excluding those authors who make small contributions to CLIS. A5 ensures that scholars are actively and currently conducting research, while A10 ensures that scholars have considerable research history. (3) The factor of association between a keyword and a scholar (*W*). A scholar has a semantic association with a keyword when a paper is published, and the strength of this association is determined by the rank of the scholar in the paper and the weights of the keywords. Table 1 lists the contribution rates of authors in CSSCI articles, determined by the author quantity and signature order; with very slight modifications, this table is based on a study by XN Su and ZR Zou, who founded CSSCI (Su and Zou 2011). If the weight of all keywords is set as 1, the strengths of associations between all scholars and all keywords can be calculated, and the total correlation coefficient between a scholar and a

Table 1 Author contribution rates based on author quantity and signature order

Author quantity	Signature order					
	1	2	3	4	5	6
1	1	–	–	–	–	–
2	0.6	0.4	–	–	–	–
3	0.6	0.25	0.15	–	–	–
4	0.6	0.2	0.1	0.1	–	–
5	0.6	0.2	0.1	0.05	0.05	–
6	0.6	0.1	0.1	0.1	0.05	0.05

The contributions of the corresponding author and authors 7 and later of an article are not considered. The total contribution rate of one article is 1

keyword can be obtained. The coefficient reflects the degree of a scholar grasping and applying a related knowledge node. However, a large number of accidental associations may occur. Therefore, the association between a scholar and a keyword is considered to be real and valid only when its coefficient satisfies formula (5), in which T5 is a constant threshold.

$$\mathrm{K10} \geq \mathrm{T1} \tag{1}$$

$$\mathrm{K5} \geq \mathrm{T2} \tag{2}$$

$$\mathrm{A10} \geq \mathrm{T3} \tag{3}$$

$$\mathrm{A5} \geq \mathrm{T4} \tag{4}$$

$$W > \mathrm{T5} \tag{5}$$

Methods for CLIS_KS construction and description

In this study, keywords are considered to be the smallest-granularity discipline knowledge, and multiple-granularity knowledge nodes can be generated after aggregating keywords in varying degrees. All of these knowledge nodes are integrated to form a complete DKS. This process is illustrated in Fig. 3. First, all of the core keywords in CLIS combined constitute the largest-granularity knowledge node, denoted as CLIS_KS. Then, based on the object similarity principle, those keywords having greater similarities are gathered together to form several clusters or classes, each of which is a discipline knowledge node with relatively large granularity, denoted as C1_KS. Clusters that contain more keywords or in which keywords have a high degree of dispersion can be further clustered, and keywords with greater similarities in the same cluster are gathered together, such that the knowledge nodes on the C1-level split into smaller-granularity knowledge nodes denoted as C2_KS. The above process is executed continually, so that large-granularity knowledge nodes constantly split into small-granularity knowledge nodes with stronger cohesion, until the number of keywords in a cluster falls to a specified threshold or the similarity of keywords in a cluster reaches a fairly high degree.

The process described above employs MLC to achieve the automatic generation of distinct-granularity knowledge nodes. However, prior to performing the specific operation, some issues need to be addressed. The first issue is the choice of clustering algorithm. To deal with large-scale keyword objects, the high-efficiency, division-based K-means clustering algorithm is employed to aggregate different-granularity knowledge nodes; then,

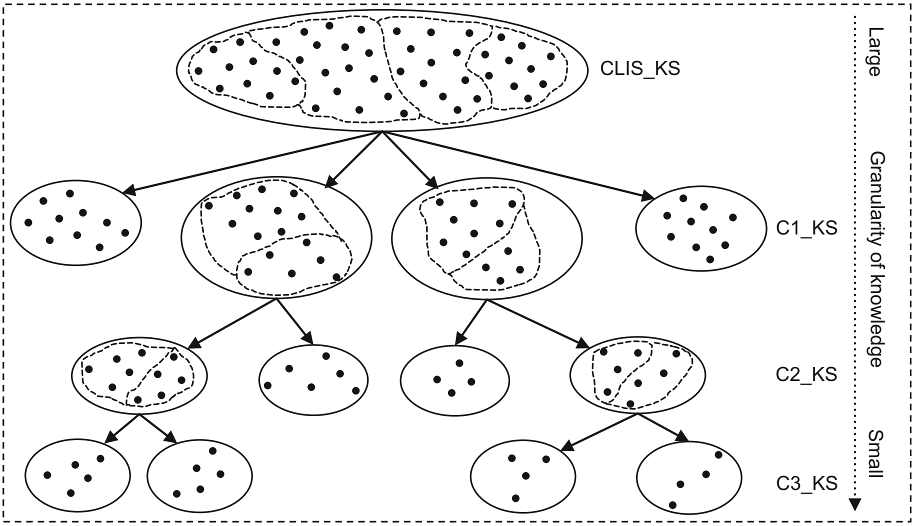


Fig. 3 Process of establishing CLIS_KS based on multi-level K-means clustering

HCA is adopted to subdivide the minimum knowledge sets and verify the effect of knowledge clustering. To address this issue, we considered a variety of clustering algorithms, including DBSCAN, BRICH, and K-means. DBSCAN is a pioneer among density-based clustering techniques; it can discover clusters of arbitrary shapes, it requires no prior knowledge of the cluster number, and it handles noise and outliers effectively. However, this algorithm does not work well in high-dimensional datasets, and it may rule out parts of keywords as noise (Ester et al. 1996; Sarafis et al. 2007; Kumar and Reddy 2016). Therefore we determined that it is not suitable for the large-scale, high-dimensional, sparse data in this study. BRICH (Zhang et al. 1996) is a bottom-up hierarchical clustering algorithm, and it is very suitable for constructing multi-level data structures. However, this algorithm requires specification of parameters beforehand, especially in setting the maximum number of leaf nodes, which is more difficult than setting the number of categories for each level. Because of this limitation and its few degrees of freedom, BRICH was not employed. K-means has a large number of practical applications, fast algorithm speeds, and a nice treatment effect for sparse matrices; furthermore, when the experimental data was precalculated, the clustering results were reliable. Therefore, the K-means algorithm was employed in this study. However, the algorithm cannot determine the cluster number, which must be manually set.

The second issue to address involves describing keywords with vectors. Keywords are used as clustered objects in this study, and they should therefore be described with features. The previous data-cleaning process produced triples of <Keyword, Scholar, Weight>, such that "Scholar" is taken to be description factor for "Keyword", and a KSM can be built as a clustered object. Generally speaking, more frequent co-occurrences of a pair of keywords in the literature or context indicate more similar themes (Cho 2014; Hong et al. 2016). This suggests that keywords frequently used by similar scholars may be correlated with each other. The third issue involves setting a category on every level. Clustering is a non-supervised classification method, and the category number (Cn_num) and name (Cn_name) must be artificially set by domain experts before clustering. The threshold of Cn_num is set according to the number of keywords in categories, which should not be too large or too small. In order to set a C1_num threshold, the experiment was conducted many times; this revealed that using too many categories leads to some small categories having fewer than 100 keywords, while using too few categories results in large categories with more than 1000 keywords. Therefore, we determined that setting C1_num = 10 guarantees that the number of keywords in each category will have three digits. Similarly, the C2_num threshold is dictated by the number of C1 category keywords and the maximum category keywords. The minimum value of the former is about 100, and the latter can be set between 5 and 20; therefore, setting C2_num = 5 ensures that the number of keywords in the second level will not be too low. Setting Cn_num for the third level and beyond must also consider the number of keywords in categories. Based on the subject characteristics of CLIS, we set Cn_num to 10 on C1-level and 5 on C2-level, while all subsequent levels are set according to formula (6):

$$\mathrm{Cn\_num} = \min(5,\ \mathrm{ceil}\,(\mathrm{m}/\mathrm{MaxNum})) \tag{6}$$

where m represents the number of knowledge nodes in the cluster needing to be clustered, MaxNum is the maximum number of keywords allowed in any lowest cluster, the function ceil (*X*) denotes the minimum integer greater than *X*, and min (*X*, *Y*) denotes the smaller of the parameters *X* and *Y*.

The fourth issue concerns the conditions required to finish MLC. As Fig. 3 shows, this study adopts MLC to achieve the division of basic knowledge nodes and the generation of distinct-granularity knowledge nodes; therefore, the conditions to end clustering must be preset. To address this, we introduced two parameters to control the clustering process. One is MaxNum, which denotes the maximum number of nodes allowed in any lowest cluster. If the number of keywords in a cluster exceeds MaxNum, the clustering continues; otherwise, the clustering stops [see formula (6)]. The other parameter is SumD, which denotes the minimum distance allowed for a cluster. If the sum of the distance between each node in the cluster and the center node exceeds SumD, the clustering continues; otherwise, the clustering stops. As long as one of these two conditions is satisfied, the clustering process is finished. MaxNum is used for the purpose of amplifying the degree of coupling between clusters, in order to prevent semantic bias caused by excessively small granularity in a knowledge node; meanwhile, SumD is used for the purpose of controlling the degree of cohesion within a cluster, in order to prevent the production of an error category or excessive node density in a cluster that could inhibit clustering. Variations in these two parameters' values lead directly to changes in CLIS_KS. The width and depth of a reasonable CLIS_KS must be effectively controlled, such that the CLIS_KS is neither too wide nor too deep; this requires the user to detect the thresholds of the MaxNum and SumD parameters.

The CLIS_KS generated can be stored as text in the form of OWL for visualization. <Owl: Class> and <Owl: subClassOf> are the main labels for describing IS_A relations. There are two basic syntaxes, as follows:

```
<owl:Class rdf:ID="Class Name">Content</owl:Class>        (7)
```

```
<owl:Class rdf:ID="Subclass Name">
    <rdfs:subClassOf rdf:resource="#Superclass Name"/>
    ...                                                   (8)
</owl:Class>
```

```
<owl:Class rdf:ID="Subclass Name">
    <rdfs:subClassOf><owl:Class rdf:ID=" Superclass Name">Content</owl:Class>
    </rdfs:subClassOf>
    …
</owl:Class>
```

(9)

Together, formulas (7) and (8) describe two classes and their parent–child relationship. The parent class is defined first, and the child class is defined in relation to its designated parent class. Formula (9) merges these two steps into one, such that the child class is defined and related to a designated parent class at the same time that the parent class is defined. Figure 4 uses these two approaches to define the classes of “Ontology” and “Semantic Web”, as well as their IS_A relation. The first approach is depicted on the left, and the second is depicted on the right.

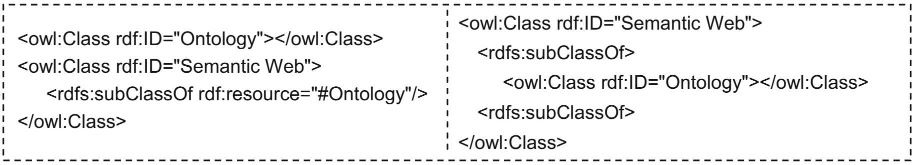
Protégé (Accessed July 1, 2015) is a plug-in tool for ontological editing and visualization display. It is able to read and write OWL files and to transform them into visual graphics. A variety of ontology visualization plug-ins currently exist for Protégé, among which Ontograf (Falconer 2015) demonstrates comprehensive visualization capabilities. It not only shows ontological relationships in a variety of layouts and effectively filters them, but also retrieves ontological concepts and quickly locates and locally displays them.

The generation of OWL files and realization of graphical display indicate that the construction of the knowledge ontology of CLIS_KS is complete, and the ontology can be further applied to the validation and analysis of discipline knowledge.

Methods for CLIS_KS Verification

CLIS_KS is automatically generated by K-means clustering. However, due to the instability of the K-means clustering algorithm, some deviations may appear in the automatic MLC results for keywords. On the whole, CLIS_KS roughly reflects the overall situation of CLIS, and it is certainly credible. Using the methods of HCA and MSDA, partial investigation for cohesion and coupling in special categories goes further in demonstrating the rationality of CLIS_KS generated by clustering.

When MLC was performed in this study, the two parameters SumD = 3 and MaxNum = 15 were set to control whether clustering would continue. However, in the final CLIS_KS generated, the lowest clusters had up to 25 objects, far more than the MaxNum threshold; instead, clustering ended because the total distance between each object within the cluster and its centroid (denoted as Sum_dis) was less than the SumD threshold. In contrast, some of the bottom knowledge categories had fewer objects than MaxNum, while



```
<owl:Class rdf:ID="Ontology"></owl:Class>
<owl:Class rdf:ID="Semantic Web">
    <rdfs:subClassOf rdf:resource="#Ontology"/>
</owl:Class>
```

```
<owl:Class rdf:ID="Semantic Web">
  <rdfs:subClassOf>
      <owl:Class rdf:ID="Ontology"></owl:Class>
  <rdfs:subClassOf>
</owl:Class>
```

Fig. 4 OWL codes on hyponymy of LIS knowledge classes “Ontology” and “Semantic Web”

Sum_dis was far greater than SumD. If the former situation is denoted as A, and the latter is denoted as B, the following rational hypotheses can be stated:

- $H1_0$: In situation A, objects inevitably gather in the vicinity of the centroid; therefore, cohesion is good, and the differences between objects within clusters are not obvious.
- $H2_0$: In situation B, objects are highly dispersed relative to the centroid, which directly leads to larger total distance and poorer cohesion, However, because the clusters have few objects, they can be further clustered using more accurate methods, depending on the needs.

The purpose of stating these two hypotheses is to provide reasons for additional research and lay the groundwork for further outlining and analyzing the internal structure of the lowest clusters, including the specific distribution of keyword nodes. MLC involves a question of when clustering should stop. The conditions for stopping clustering have different results, and the thresholds can be adjusted. Therefore, the distributions in the lowest clusters must be understood in order to answer these questions. HCA, which can describe the process of clustering in detail, can be employed to make an argument for understanding the relative position of objects within a cluster and their degree of cohesion.

Another question is whether knowledge categories in CLIS_KS obtained by K-means clustering are relatively independent; in other words, how is the coupling between knowledge categories? To address this question, MDSA is employed to scatter the nodes in different knowledge categories over a two-dimensional plane. Then, the rationality of category division is verified according to the position distribution of the nodes on the plane, as well as the coupling between categories. The two samples from category A with the most objects (denoted as A1 and A2) and the two from category B with the largest Sum_dis (denoted as B1 and B2) are used to carry out coupling analysis. Prior to the specific operation, the hypotheses can also be stated as follows:

- $H3_0$: The objects from the four categories are independent of each other; therefore, the coupling between categories is very low, and clustering has a positive effect on category division.
- $H4_0$: The hierarchy of the objects from A1 and A2 is poor, while further clustering may be possible for the objects from B1 and B2.

Methods for CLIS_KS Analysis

According to the hierarchical structure of discipline knowledge, the semantic relationships between discipline knowledge and other academic objects can be further understood by means that include sketching out the knowledge structure of individual academic objects, discovering the core academic object groups for various categories in CLIS, and even depicting associations of content containing and crossing academic objects. Academic objects mainly include scholars, journals, institutions, and areas. Analysis of narrow-sense KS including only research content transforms to exploration of broad-sense KS containing all academic objects in a given discipline. This study adopts only scholars as example objects, assessing from microscopic, mesoscopic, and macroscopic levels in order to fully detect and discuss the semantic associations between the scholars and CLIS_KS.

Microscopic analysis detects the internal structure of academic objects and examines their research interests and contributions to the discipline based on the content details. This analysis specifically refers to exploration of the KS of individual scholars. Given the length of this study, the scholar “JP Qiu” was chosen for a detailed detection of his microscopic

KS; he has the largest knowledge coefficient in CLIS from 2003 to 2012. The knowledge coefficient is the total correlation coefficient between a scholar and all his related keywords, which are the basic knowledge nodes. Of course, this analysis mode is also applicable to other scholars and academic objects, and it can help create a deep understanding of their research focuses and corresponding causes within a certain range of time, as well as providing a factual basis and referential suggestions for the further development and improvement of academic objects in the CLIS study.

Mesoscopic analysis comprehensively investigates the same or similar external performances of academic objects groups. This analysis specifically refers to probing the associations between scholars and category knowledge nodes (CKNs), so as to analyze the core scholar groups on distinct, hierarchical category levels. The core scholar groups are the main creators and users of CKNs. First, keywords are generalized to CKNs based on CLIS_KS. In this way, the triples <Scholar, Keyword, Weight> are converted to <Scholar, Sub-domain, Correlation Coefficient>. Then, in accordance with the correlation coefficient, the rank of scholars is calculated, and the collection of the most relevant scholars in every CKN is identified as the core scholar group for this category. Finally, the relationships between CKNs and core scholar groups are described and visually displayed through social network analysis (SNA). Given the restrictions on length, this study only depicts the core scholar groups for sub-domains (categories) C1 and C2. The core scholars are identified for every sub-domain according to formula (10), in which CC_Rank means the rank of the sum of correlation coefficient in sub-domain, S_Num means the number of scholars in sub-domain and n is an indefinite value empirically determined based on the level and size of the sub-domain. Similar to the microscopic analysis described above, this study provides a model for analyzing the relationships between academic objects and CKNs with distinct generalization degrees based on CLIS_KS. This model can conveniently single out the most relevant academic objects for every research sub-domain in a discipline, and then offer knowledge services for evaluating academic objects, retrieving reference information, submitting academic papers, and surveying a region's study domain.

$$\frac{\text{CC\_Rank}}{\text{S\_Num}} \geq n\% \tag{10}$$

Macroscopic analysis discusses the research characteristics of academic objects to a more comprehensive extent. CLIS_KS contains discriminative generalized levels, including basic knowledge nodes (BKNs) at the bottom in addition to distinct, comprehensive CKNs. Because of the complexity of Chinese compound words and less normative Chinese keywords, the BKNs for scholars are complicated, with basically no rules to follow. However, if a scholar's BKNs are generalized to a certain extent, such that the non-core study fields of the scholar are excluded, the scholar's true CKNs can be obtained. At the same generalized level, the KS of scholars can exhibit some regularity. Therefore, this study addresses the associations between scholars on level C1, which includes 10 sub-domains, and analyzes the possibility of knowledge exchange and integration among scholars so as to promote research innovation in the discipline. In detail, the methods are as follows: First, BKN that all scholars are related to are generalized to knowledge nodes of the C1 sub-domains; this allows the number of knowledge nodes to drop from 3081 to 10, and the content crossing between scholars increases substantially. Second, scholars' non-core C1 sub-domains are removed. The threshold K_C1 is introduced; if the knowledge coefficient of a scholar in a C1 sub-domain is larger than K_C1, this C1 sub-domain is

Table 2 Description of variables in formula (11)

Variable	Description	Variable	Description
$\lambda_{Y \to X}$	The association degree of scholar *Y* with *X*, where the latter is the central object	cf_{XY}	The number of C1 sub-domains in which scholar *X* and *Y* co-occur
tf_{iXY}	The shared knowledge coefficient of scholars *X* and *Y*	N	The total number of C1 sub-domains
tf_{iX}	The knowledge coefficient of scholar *X* in C1 sub-domains	cf_X / cf_Y	The number of C1 sub-domains in which scholar *X*/*Y* occur

considered to represent the scholar's core research field. A scholar's K_C1 threshold is set to the average of all of the scholar's knowledge coefficients with related C1 sub-domains. Third, the cross-correlation between scholars is calculated based on the research crossing between scholars on level C1. The improved TF-IDF (Wang 2010) is employed to describe different interactions between associated scholars'. The calculation method is shown in formula (11), and the variables are described in Table 2. When scholar *Y* has an important effect on scholar *X*, their shared knowledge coefficient tf_{iXY} should be much higher, and scholar *Y*'s research should be much more converged. Finally, the most important *M* scholars for scholar *X* can be identified as associated scholars of scholar *X*. Similarly, when discussing the associations between scholars on level C2, the BKNs can be generalized to the CKNs on level C2.

$$\lambda_{Y \to X} = \text{Relevancy}(S_X, S_Y) = \frac{\sum_{i=1}^{n} C_{iXY}}{\sum_{i=1}^{n} C_{iX}} \times WeightingFactor\,(S_Y) \tag{11}$$

$$C_{iXY} = tf_{iXY} \times \log_{10}\left(\frac{N}{cf_{XY}} \times (N - cf_X)\right) \tag{11 - 1}$$

$$C_{iX} = tf_{iX} \times \log_{10}\left(\frac{N}{cf_X} \times (N - cf_X)\right) \tag{11 - 2}$$

$$WeightingFactor\,(S_Y) = \frac{\log_{10} \frac{N}{cf_Y}}{\log_{10} N} \tag{11 - 3}$$

Results and discussion

Results of CSSCI data cleaning

Starting from a subject category of 870 (the serial number of CLIS in CSSCI), this study retrieved records for all papers and their authors from CSSCI between 2003 and 2012, using this data as a foundation to construct DKS. This dataset included a total of 58,281 papers, 34,222 scholars, and 67,351 keywords. Obviously, this represents a huge amount of data; therefore, the data were screened according to preset parameters. Thresholds were set for each parameter, denoted as T1 through T5 as described in "Methods for CSSCI data cleaning" section, according to the following principles. (1) The laws of the distribution of source data, including the frequency and number of keywords, and the number and issuing number of scholars, must meet certain criteria. (2) The size of the calculated data, including the selected keywords and scholars, must cover the entire discipline, while the size of the experimental data must be controlled within calculability. (3) Repeated testing is necessary to ensure that all important keywords and well-known scholars are selected. (4) The actual situation of CSSCI papers is considering. The degree of correlation between the content of papers and authors other than the first author is small. Based on these four considerations, the distributions of keywords and scholars were evaluated, with the results shown in Fig. 5.

Figures 5a, b show the keyword frequency trends over the last 10 and 5 years, respectively, under different grades. Figure 5a divides the range 0 through 25 into 6 grades, and Fig. 5b divides the range 0 through 15 into 6 grades. When the grades of both are 2, or $K10 \geq 5$ ($T1 = 5$) and $K5 \geq 3$ ($T2 = 3$), changes in keyword frequency begin conforming with the trendline with regularity; at this point, the keywords satisfying these conditions are recognized by CLIS. Figures 5c, d show the trends for numbers of scholars over the last 10 and 5 years, respectively, under different grades. Figure 5c divides the range 10 through 20 into six grades, and Fig. 5d divides the range 5 through 10 into 6 grades. Published papers are counted only for the first author, and the grades are divided based on the issuing number. The parameters are set at $A10 \geq 10$ ($T3 = 10$) and $A5 \geq 6$ ($T4 = 6$), such that only authors who have published at least ten papers in past 10 years and at least six papers in the past 5 years are selected as CLIS scholars. The final parameter is set as $T5 = 0.6$ and $W > 0.6$; in other words, only a keyword used by a scholar who is the first author in more than one multi-author paper is inevitably association with this scholar. To refine the data further, scholars with the same name were checked based on data filtering. A dozen groups of scholars with the same name were found, and the full names were all short, with common surnames. After these names were marked with different tokens, data filtering was performed again. It is worth noting that some errors may remain in the results due to the strong subjectivity and rough operation of the cleaning process. It is possible that some scholars with same name were not separated.

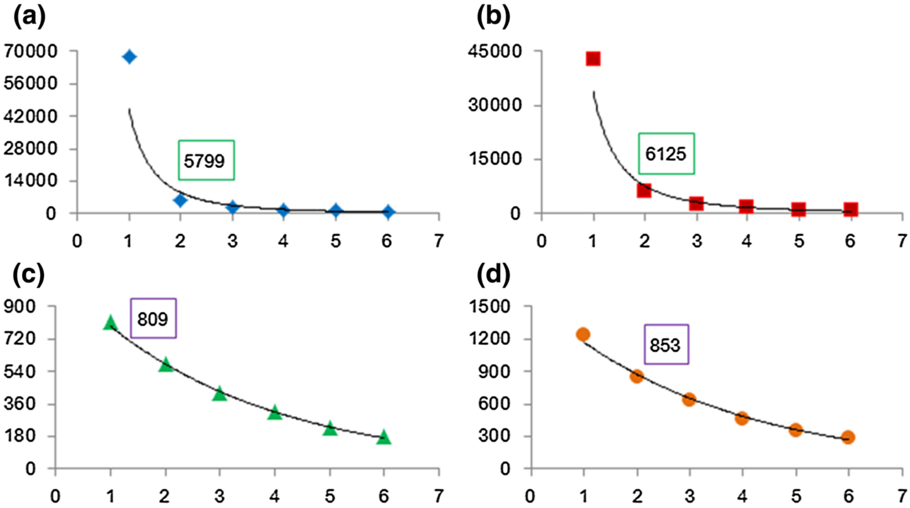


Fig. 5 The trends for keywords frequencies (**a**, **b**) and numbers of scholars (**c**, **d**) in the most recent 10 (**a**, **c**) and 5 (**b**, **d**) years under the different grades

After data cleaning, CLIS ultimately recognized 3081 valid keywords, 575 important scholars in the discipline, and 12,005 semantic associations between keywords and scholars. Therefore, the obtained triples <Keyword, Scholar, Correlation coefficient> were used as the data foundation to construct CLIS_KS.

Results of CLIS_KS construction and description

Results of CLIS_KS construction

In order to identify final values for MaxNum and SumD, this study set several values for MaxNum {5, 10, 15, 20} and SumD {2, 3, 4, 5, 6} and carried out MLC experiments 20 times to test all combinations. The detailed process proceeded as follows. First, according to the rules for setting the number of categories on all levels, KSM was clustered with MLC using the K-means algorithm. However, the K-means algorithm produces instability in clustering results by selecting different initial centroids; therefore, every K-means clustering operation was executed ten times, and the final result was the one that minimized the sum of the distance between all nodes in the cluster and the centroid. The final clustering results are shown in Table 3.

As Table 3 proceeds from top to bottom, the conditions for clustering become gradually stricter, and the generated knowledge hierarchical structures also undergo a subtle change. The number of generated clusters (Num_C) decreases significantly, and the overall width of the hierarchy (MaxWid_H, the number of nodes on the level with most nodes) becomes smaller. As the number of nodes allowed in a cluster increases, the maximum depth of the hierarchy (MaxDep) generally increases; meanwhile, the minimum depth (MinDep) shows a decreasing trend, such that the whole structure is essentially elongated. As the conditions for clustering become stricter, the maximum number (MaxWid_C) and the minimum number (MinWid_C) of clusters both demonstrate an upward trend, with the maximum changing very significantly. Generally speaking, in a reasonable class hierarchy, the overall width, depth, and size of the clusters must be moderate. Therefore, MaxNum can reasonably equal 15, and SumD can reasonably equal 3; in other words, when there are more than 15 nodes in a cluster, and the sum of distances between every node and the centroid exceeds 3, clustering will continue so as to ensure that the nodes in every cluster gather in a small space or the number of nodes is less than 15. At this moment, the depth of CLIS_KS (MaxDep) is less than 10, while the width (MaxWid_H) is between 160 and 200. At the same time, the largest number of keywords in a cluster (MaxWid_C) is 25, which is not excessive compared with the threshold of MaxNum, and the total number of categories (Num_C) is as small as possible. Relatively speaking, this structure is more reasonable.

The MLC experiment was performed an additional ten times with the selected parameters. Then, combining the opinions from domain experts with the characteristics of hierarchical structure including width, depth, and category number, one of the most reasonable clustering results was chosen to represent CLIS_KS. The chosen result has a depth of 9, a width of 178, and a total of 347 CKNs. The first level of CKN (C1) is shown in Table 4, in which the keywords with the highest frequency are used as names for the categories (C1_name). In the CLIS_KS shown in the table, "University Library" and related research forms the largest category, with 741 keywords. This is more than twice the number of keywords in the "Digital Library" category, which is ranked second. Therefore, this CKN currently constitutes the main research content of CLIS. The study scales for "Informatics" and "Philology" are relatively small, with the number of related keywords being fewer than 200. "Philology" seems to represent a declining trend as a traditional

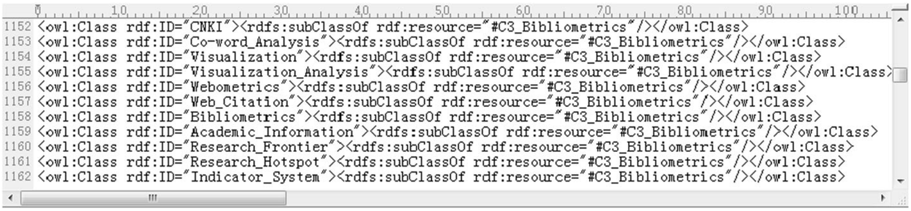
Table 3 Characteristics of generated CLIS_KS with different values for MaxNum and SumD

No.	Cn_num	MaxNum	SumD	MaxDep	MinDep	MaxWid_H	MinWid_H	MaxWid_C	MinWid_C	Num_C
1	C1 = 10	5	2	6	3	441	10	20	1	710
2	C2 = 5		3	9	3	376	5	24	1	617
3	Cn = min (5, ceil (m/MaxNum)) (n > 2)		4	7	3	322	9	26	1	557
4			5	7	2	262	3	32	1	494
5			6	6	2	239	5	33	1	471
6		10	2	8	3	264	2	22	1	506
7			3	7	3	261	4	23	1	471
8			4	8	2	226	2	27	1	418
9			5	7	3	221	4	31	1	387
10			6	8	3	223	2	33	1	358
11		**15**	2	8	2	178	2	25	2	347
12			**3**	**8**	**2**	**180**	**4**	**25**	**2**	**335**
13			4	9	2	187	2	27	2	327
14			5	8	2	183	2	31	2	317
15			6	8	2	189	2	33	1	299
16		20	2	10	2	152	4	22	1	278
17			3	9	2	147	4	24	2	265
18			4	9	2	152	2	27	2	267
19			5	11	2	153	2	33	2	256
20			6	7	2	150	4	34	2	245

Bold values are the final values for the parameters of MaxNum and SumD, and the results of clustering

Table 4 First level of CKN (C1) in CLIS_KS

C1_no.	C1_name	Number of BKN	C1_no.	C1_name	Number of BKN
11	C1_University_Library	741	16	C1_Communication	244
12	C1_Philology	195	17	C1_Digital_Library	356
13	C1_Public_Library	261	18	C1_Library	298
14	C1_Competitive_Intelligence	316	19	C1_Knowledge_Management	282
15	C1_Search_Engine	254	20	C1_Informatics	134



```
<owl:Class rdf:ID="CNKI"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Co-word_Analysis"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Visualization"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Visualization_Analysis"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Webometrics"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Web_Citation"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Bibliometrics"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Academic_Information"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Research_Frontier"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Research_Hotspot"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
<owl:Class rdf:ID="Indicator_System"><rdfs:subClassOf rdf:resource="#C3_Bibliometrics"/></owl:Class>
```

Fig. 6 Partial OWL codes on CLIS_KS

research direction for CLIS, while "Informatics" has experienced a great diversion of research contents since the rise of research classified as "Competitive Intelligence" and "Search Engine". The ten C1 categories, four of which are related to the term "library", indicate that up to 2012, research into "Library Science" has remained the focus of knowledge distribution in CLIS.

Results of CLIS_KS description

The entire CLIS_KS can be coded using OWL labels; then, the CLIS knowledge ontology can finally be formed, containing only hierarchical relationships. After the OWL ontology file is generated, the file can be read with Protégé in order to perform class retrieval and visual display using visualization plug-ins such as OntoGraf. Figure 6 lists partial OWL codes for CLIS_KS, which mainly describe the hypernym/hyponym relationships between the fourth-level category "C3_Bibliometrics" and its terminal KN. Figure 7 structures all the KN in the three fourth-level categories "C3_Semantic_Web", "C3_Library_Cause", and "C3_Bibliometrics" into a spring shape, in which the node for CLIS_KS is taken as the ceiling-level CKN. Figure 8 is a tree-shaped graph on DKS constructed from parts of categories in CLIS_KS.

Results of CLIS_KS verification

Results of cohesion analysis

Two cases were chosen from A and B (see also "Methods for CLIS_KS Verification" section) to represent two kinds of extreme CKN. This study attempted to explore and analyze the distribution and cohesive characteristics of keywords in these CKNs, using the HCA method so as to verify the above hypotheses listed in "Methods for CLIS_KS Verification" section.

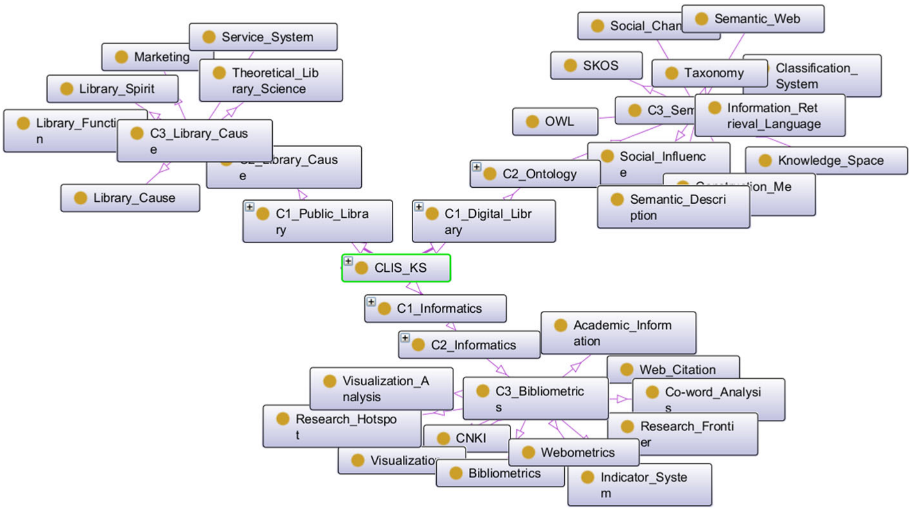
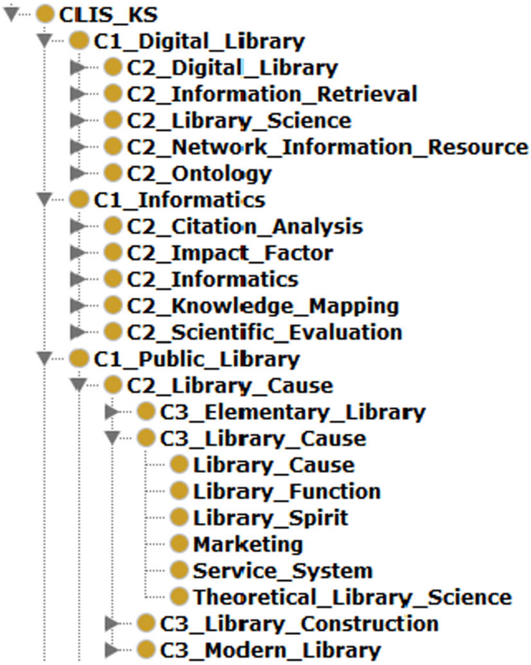


Fig. 7 Spring-shaped visualization of DKS using parts of categories in CLIS_KS

Fig. 8 Tree-shaped DKS using parts of categories in CLIS_KS


- CLIS_KS
  - C1_Digital_Library
    - C2_Digital_Library
    - C2_Information_Retrieval
    - C2_Library_Science
    - C2_Network_Information_Resource
    - C2_Ontology
  - C1_Informatics
    - C2_Citation_Analysis
    - C2_Impact_Factor
    - C2_Informatics
    - C2_Knowledge_Mapping
    - C2_Scientific_Evaluation
  - C1_Public_Library
    - C2_Library_Cause
      - C3_Elementary_Library
      - C3_Library_Cause
        - Library_Cause
        - Library_Function
        - Library_Spirit
        - Marketing
        - Service_System
        - Theoretical_Library_Science
      - C3_Library_Construction
      - C3_Modern_Library

(1) A: CKNs with the most objects, both including 25 keywords

A1: LIS_KS>C1_Public_Library>C2_Library_Management>C3_Reader_Services>C4_Library_Architecture>C5_Library_Architecture>C6_Library_Architecture, Sum_dis = 2.787157.

A2: LIS_KS>C1_Informatics>C2_Informatics>C3_Bibliography_Metering>C4_Bibliography_Metering>C5_Knowledge_Exchange, Sum_dis = 2.833931.

The KSMs of A and B, which constitute 25 keywords, are clustered using HCA, and the results are shown as A1 and A2 in Fig. 9. A1 contains 11 completely similar objects, accounting for about half of the total. The maximum difference between objects within the cluster (MaxDiff) is only slightly larger than 0.7. The distribution of BKNs in A2 is very similar to A1; there is also a group of 11 objects among which the distances are all 0, and the MaxDiff is slightly larger than 0.7. The maximum differences of BKNs in A1 and A2 are quite small, even with no differences between a large number of BKNs, and the clustering hierarchies are not obvious. This implies that there is no significance in classifying BKNs within A1 and A2 further, and the BKNs of these categories maintain a great cohesion.

(2) B: CKNs with the largest Sum_dis
B1: LIS_KS>C1_Library>C2_Information_Sharing>C3_Knowledge_Integration, Sum_dis = 8.502761
B2: LIS_KS>C1_Philology>C2_Four_Books_Comprehensive_Table_of_Contents_Abstract>C3_Intangible_Cultural_Heritage, Sum_dis = 7.464072.

The keywords in B1 and B2 are clustered using HCA, and the results are shown as B1 and B2 in Fig. 10. In B1, MaxDiff between the BKNs is close to 1.6, almost twice as large as A1' and A2' MaxDiff. Additionally, the BKNs in B1 are distributed in a wider range of space. The 15 BKNs are divided into three distinct groups, making the clustering hierarchy extremely clear. The distribution of BKNs in B2 is similar to that in B1, in which No. 5 and No. 12 are identical, as are No. 2 and No. 6. These 14 objects are also divided into three distinct categories. Because MaxDiff exceeds 1.6, the dispersion among the objects is even greater than in B1.

An analysis of the four bottom CKNs yields the following results:

- $R1_0$: $H1_0$ should be accepted. Category A generally contains more BKNs, but due to the great similarity between internal objects and the strong cohesion resulting from the internal objects gathering in a small space, it is impossible to continue the subdivision.
- $R2_0$: $H2_0$ should be accepted. The dispersion among the BKNs in category B is greater, and the hierarchy of internal objects is fairly clear; therefore, it is obvious that the internal objects have relatively poor cohesion. However, because of the small number of internal objects, the categories should be further subdivided using the cluster algorithms, which are more accurate and suitable for small-scale data such as HCA.

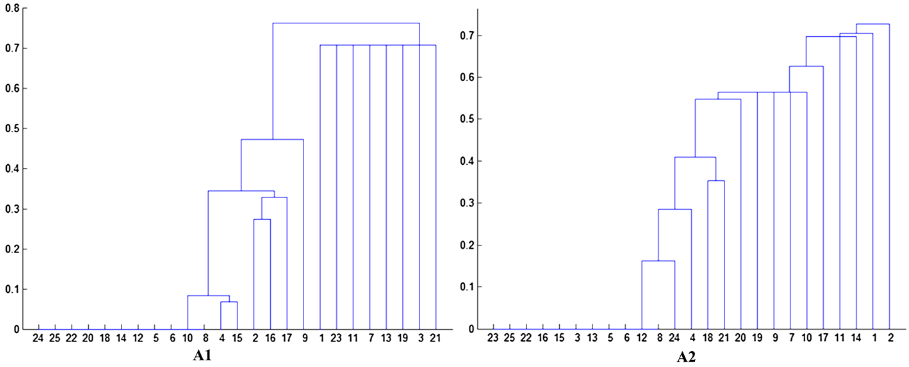


Fig. 9 Results of HCA on the CKN containing the most keywords

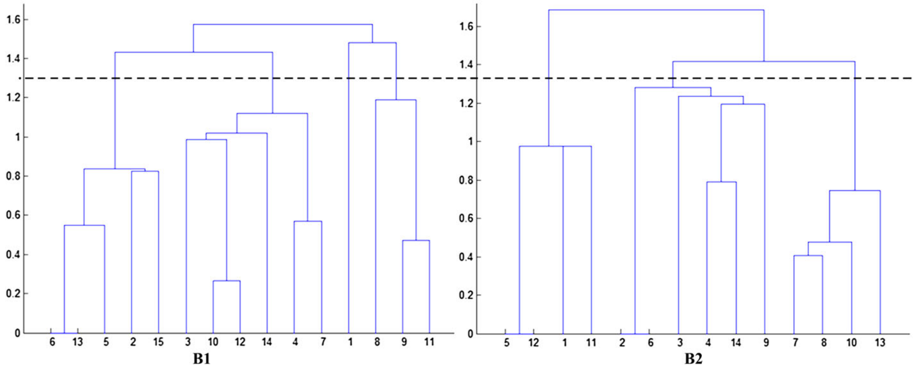


Fig. 10 Results of HCA on the CKNs with the largest Sum_dis

In summation, microscopic verification of the internal distribution of special categories makes it clear that the CKNs obtained by MLC through the K-means algorithm in CLIS have either strong cohesion or less embedded BKNs; therefore, the CLIS_KS possesses a certain rationality.

Results of coupling analysis

MDSA is performed on the KSM, a 79-by-76 matrix formed by the keywords from A1, A2, B1, and B2 to calculate the pairwise relative distances between all of the keywords. It involves dimensionality reduction for compression, depending on the distances. Ultimately, the KSM demonstrates that credibility stress equals 0.10893, and values greater than 0.1 are recognized as generally credible. Additionally, the validity RSQ equals 0.98240, and values larger than 0.6 convey validity. Evidently, dimensionality reduction has good effects, but the credibility is general. According to the position coordinates of the keyword objects after reducing the dimensions from 76 to 2, the relative positions of the BKNs from the four categories are shown in Fig. 11.

Overall, objects from these four clusters are divided into three groups. The blue circles in the second quadrant represent the BKNs from A1; the pink points from A2 lie in the fourth quadrant; and the black forks from B1 and the red diamonds from B2 are piled together in the first quadrant, with B1 objects occupying the gaps between B2 objects. These results suggest that coupling among these CKNs is low, because A1 and A2 are independent from the others. Dimensionality reduction distorts the associations between objects to some extent, with the result that B1 and B2 are integrated into one cluster in the two-dimensional plane. It seems that the characteristics distinguishing B1 and B2 objects are lost in the process of reducing dimensions. Of course, compared with A1 and A2, the association between B1 and B2 is much closer; B1 is from “C1_Library” and B2 is from “C1_Philology”, so the two groups may be correlated along some dimension.

There are 25 BKNs in A1 and A2, but fewer nodes appear in the figure. This suggests that many objects are covered by others, and the BKNs in these two clusters are relatively concentrated and highly cohesive. Of course, there may be individual objects far away from the center of the collection, such that the BKNs show a larger difference; this could be attributed to less similarity with other BKNs in the same cluster, or dimensionality reduction could magnify distortions. The distances between dissimilar BKNs show a linear growth trend in A1 and A2, such that the clustering hierarchy of BKNs is not clear and it is

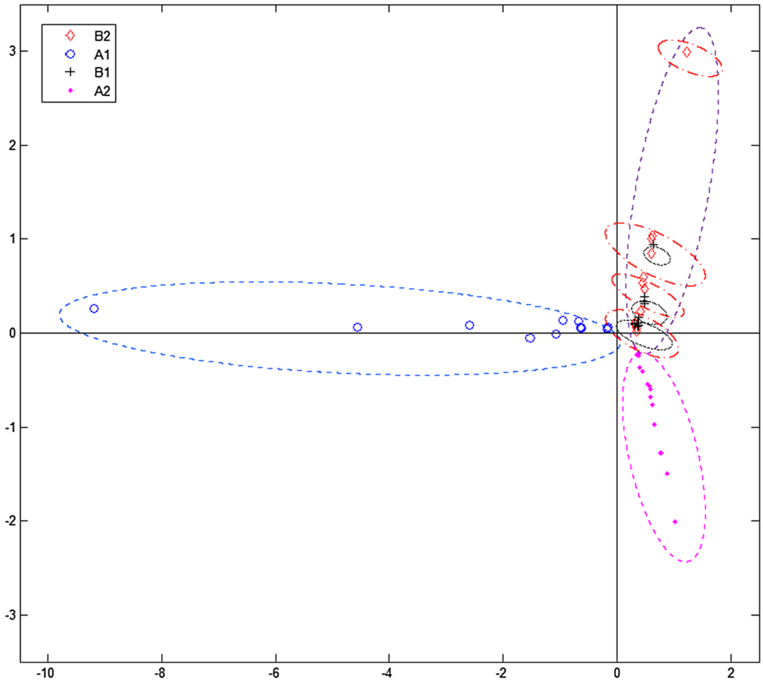


Fig. 11 Results of MDSA on four CKNs with greatest Sum_dis and most keywords in CLIS

difficult to subdivide A1 and A2 further. In other words, the current classification has reached the optimum level.

The BKNs are also concentrated in B1 and B2; this indicates that the division of KC has reached a good effect based on K-means clustering, such that similar BKNs are gathered together. However, in contrast to the linear distribution of the BKNs in A1 and A2, the BKNs in B clusters show a more significant hierarchy. In fact, the BKNs can be subdivided into several groups. B1 can easily be divided into three groups, as marked with black-dot circles in Fig. 11, while B2 can be divided into four groups, marked with red-dot-line circles. This implies the potential for further clustering in B1 and B2, which could be performed depending on actual demand. However, it should be noted that the MDSA method displays differences between objects on a flat surface by reducing dimensions for compression; this introduces some distortion, such that the results cannot be used as a basis for further classification.

Based on the analysis above, the coupling of CLIS_KS is low, and the results are as follows:

- $R3_0$: $H3_0$ should be accepted. Viewed as a whole, the four categories are independent from each other, which reflects poor coupling among the categories. The discrimination between the B categories is relatively worse.
- $R4_0$: $H4_0$ should be accepted. The objects in A1 and A2 are concentrated, with strong cohesion and poor hierarchy. In contrast, the possibility of further clustering exists for B1 and B2, consistent with the conclusions of the previous section.

In summary, the CKNs obtained by HCA and MDSA have high cohesion and low coupling; therefore, the CLIS_KS has strong credibility and rationality.

Results of CLIS_KS analysis

Results of microscopic analysis: scholar's knowledge structure

This study used the scholar "JP Qiu", who has the largest knowledge coefficient in CLIS at 220.10, as an example. CLIS_KS and the associations between scholars and related BKNs can be employed to build the complete KS of individual scholars, as well as to analyze the scope and depth of their research. According to the CLIS_KS, the individual KS of "JP Qiu" contains eight levels and 188 KNs. Because the structure of the diagram is very complex, it will be stratified and discussed separately in Figs. 12 and 13.

The double-bordered box in Fig. 12 represents CKN. Figure 12 shows only the top two levels of the KS, and the nodes' shading indicates the extent of the scholar's research in the CKN. A knowledge coefficient between 0 and 250 is averagely divided into five levels. The deeper the color, the more of the scholar's studying appears in the CKN. The total knowledge coefficient of the scholar is listed below the CKN. Figure 12 clearly shows a wide research scope, which covers seven of ten total C1 sub-domains in CLIS, in addition to 20 C2 sub-domains. Although the research scope is wide, this scholar only scratches the surface in the majority of the seven C1 sub-domains, and six C1 sub-domains have knowledge coefficients no larger than 10; this indicates that these C1 sub-domains are not focuses of his research. More than 80 % of this scholar's studies are focused on "C1_Informatics", and this CKN also shows that nearly 80 % of his studies are performed in "C2_Informatics", followed by "C2_Knowledge_Mapping" and "C2_Citation_Analysis". The total studies represented in these three C2 sub-domains constitute 95 % of his entire body of work in "C1_Informatics".

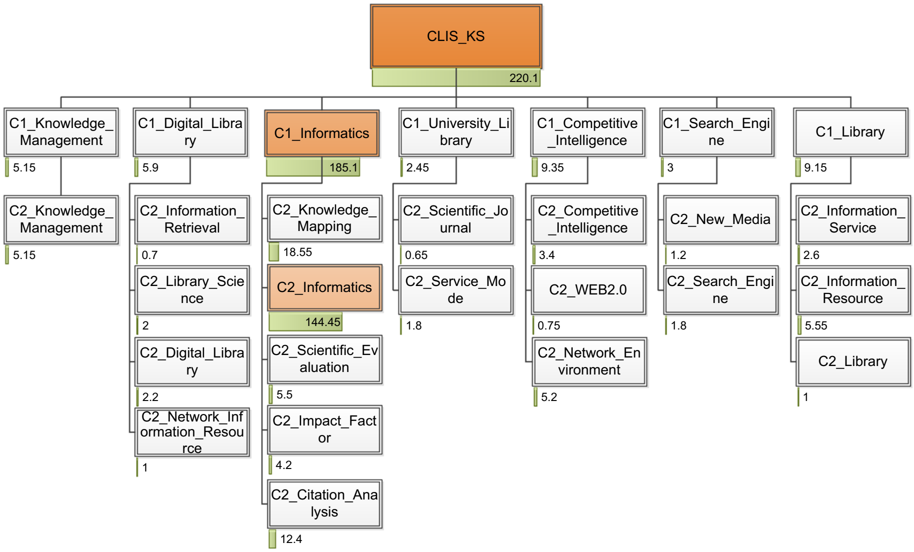


Fig. 12 Top two levels of the KS of "JP Qiu"

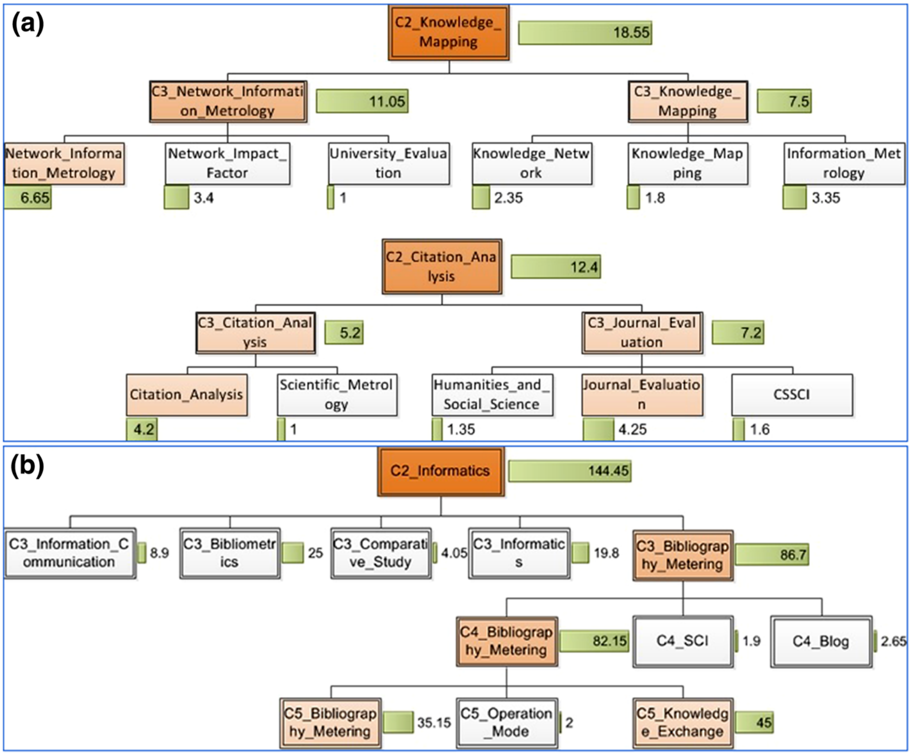


Fig. 13 Details of the three largest sub-domains in JP Qiu's KS

From this point, the scholar's research focuses can be further pinpointed. In his key research domains, he has achieved great depths of research. In order to further explore the scholar's KS, Fig. 13 shows his KS in "C2_Informatics", "C2_Knowledge_Mapping", and "C2_Citation_Analysis."

In Fig. 13, the double-bordered box represents CKN, and the single-bordered box represents BKN. The other graphic symbols are similar to the ones used in Fig. 12. Part A of Fig. 13 is the KS of "C2_Knowledge_Mapping" and "C2_Citation_Analysis" for "JP Qiu", and the background color of each node corresponds with the level of the knowledge coefficient divided from 0 to 20 into five levels. Part B is the KS of "C2_Informatics", and the background color of each node corresponds with the level of the knowledge coefficient divided from 0 to 150 into five levels. This sub-domain has too many BKNs to list. In Part A of Fig. 13, the two C2 sub-domains are the secondary research directions, and the total knowledge coefficient does not exceed 20. However, a number of BKNs used often by the scholar, such as "Network_Information_Metrology", "Journal_Evaluation", and "Citation_Analysis" have knowledge coefficients exceeding 4. This suggests that the scholar used these BKNs in at least four papers. "Information_ Metrology" and "Network_Impact_Factor" are also similar to the BKNs listed above. The C2 sub-domain in Part B represents the scholar's core research contents, consisting of five C3 sub-domains. Three of these, "C3_Bibliography_Metering", "C3_Bibliometrics", and "C3_Informatics", represent the scholar's major focus. The first one can be divided further to yield a C5 level. The BKNs related to this scholar mostly derive from "C2_Informatics", which can be assumed to represent the scholar's research focus. Although numerous BKNs appear in the KS, few are frequently used by the scholar. In fact, only six BKNs knowledge coefficients exceed 4: "Informatics", "Bibliometrics", "Social_Network_Analysis", "Bibliography_Metering", "Content_Analysis_Method", and "Link_Analysis". "C5_Knowledge_Exchange" is the scholar's highest C5 sub-domain in terms of knowledge coefficients, containing 25 BKNs; however, only three BKNs have knowledge coefficients larger than 3, "Multidimensional_Scaling_Analysis", "Author_Co-citation_Analysis", and "Knowledge_Exchange". Some domains have high knowledge coefficients due to the large number of BKNs, while other domains have medium knowledge coefficients but contain a small number of BKNs frequently used by the scholar. Looking at the high-frequency BKNs from the different categories listed above, it is clear that "metrology", "scientific evaluation", and "citation analysis" are this scholar's absolute core research subjects. These contents are classified into different CKNs for two main reasons. One is that the Chinese keywords are irregular and fully processed by machine; for example, "Bibliography_Metering" and "Bibliometrics" should actually be merged. The other reason is that some errors occurred in the unsupervised machine clustering. Further discussion should also address using related scholars as descriptive features of keyword objects.

"JP Qiu" is the scholar with the highest knowledge coefficient in CLIS. He not only has a wide knowledge scope reaching almost all the domains in CLIS, but also has a strong study depth in the domain of metrology. In the most recent 10 years, he has primarily focused on knowledge application and knowledge accumulation in this domain. These are the characteristics of this scholar's individual KS.

Results of mesoscopic analysis: core-scholars of CLIS

This section analyzes the core scholars in sub-domains C1 and C2 of CLIS. For C1, n is set to 5, and for C2, n is set to 3. However, it is important to note that this study focuses on the general results and typical modes of data analysis based on the CLIS; therefore, the value of n is only an example. It can be set to other values based on the analyst's actual priorities.

There are a total of ten C1 sub-domains in CLIS_KS. Furthermore, 174 relationships involving 159 core scholars meet the condition that the association coefficients between scholars and C1 sub-domains rank in the top 5 % in every C1 sub-domain. The specific distribution of core scholars in C1 sub-domains is shown in Fig. 14. In Fig. 14, red circles indicate the ten C1 sub-domains; their names, all beginning with C1, are written beside the circles. The size of each circle represents the sum of the association coefficients between the core scholars and the C1 sub-domain. The ratio of this sum to the total association coefficients of that C1 sub-domain is in the name label. Blue squares indicate scholars, and they are labeled with each scholar's name. The links between blue squares and red circles represent associations in which the scholar is a core scholar of the corresponding sub-domain. The size of the square represents the strength of the association between the scholar and the sub-domain; if a scholar associates with multiple sub-domains, the size of his square represents the sum of all the associations.

Figure 14 shows that in the ten C1 sub-domains, the overall core scholar knowledge references are highest in the library-related sub-domains, such as "C1_Library", "C1_University_Library", "C1_Digital_Library", and "C1_Public_Library". This result indicates that many more studies are carried out in these sub-domains. To some extent, this suggests that these sub-domains are the main research objects in CLIS, and they all contain a large number of discipline KNs. As for the four sub-domains listed above, in addition to "C1_Public_Library", the associations of other sub-domains account for about 30 % of the

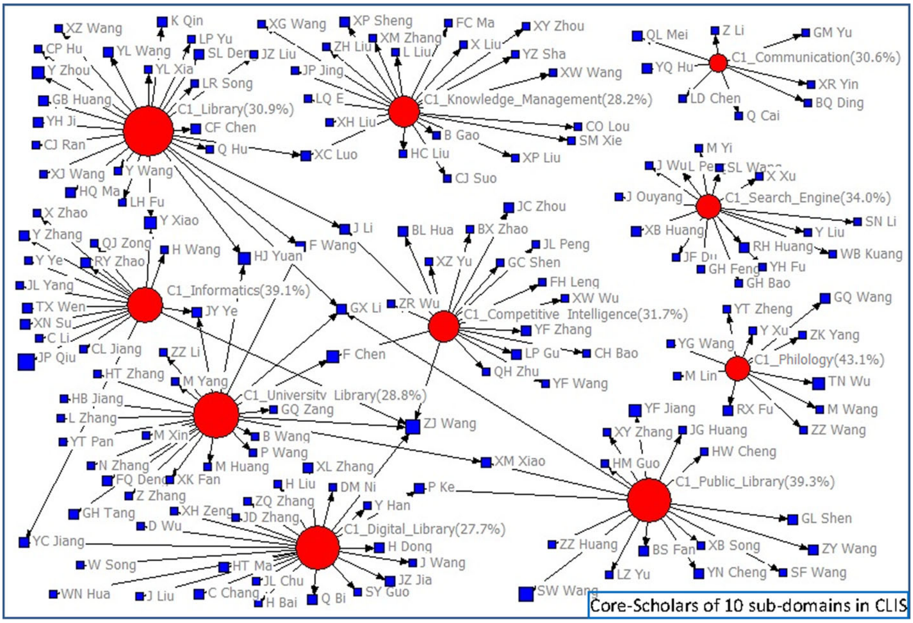


Fig. 14 Core scholars (top 5 %) in the C1 sub-domains in CLIS

total associations; in other words, “C1_Knowledge_Management”, “C1_Communication”, “C1_Competitive_Intelligence”, and “C1_Search_Engine” are all similar in that the sum of the core scholars’ studies in a sub-domain account for about 30 % in the sub-domain. This reveals a major feature of CLIS, in that generally speaking, about 5 % of scholars account for about 30 % of the work in their research field. However, the ratio is close to 40 % for “C1_Public_Library” and “C1_Informatics”, while the ratio for “C1_Philology” reaches 43 %. The research tasks in these three sub-domains are much more concentrated, such that 5 % scholars perform about 40 % of the research work. These sub-domains all seem to represent traditional research directions for CLIS, and the literature is relatively mature. Long-term research results in a certain number of core characters, and it is hard to achieve research innovation in these directions.

On the whole, the studies in “C1_Communication”, “C1_Search_Engine”, “C1_Philology”, and “C1_Knowledge_Management” are relatively independent, and their core scholars have very little crossover with other categories. These sub-domains are either traditional research directions that are currently developing slowly, such as “C1_Philology”, or emerging fields in their discipline that are the subjects of a great deal of knowledge and technology outside CLIS. Research in emerging fields is special, such that little crossover occurs with other fields in CLIS. Examples of such crossovers include “C1_Communication”, “C1_Search_Engine”, and “C1_Knowledge_Management”, which are the products of content crossing between CLIS and journalism, computer science and management science. A great deal of crossover exists among the core scholars from the four library-related and two intelligence-related (“C1_Informatics” and “C1_Competitive_Intelligence”) fields, suggesting that these sub-domains are not completely separated. They may have some common research, and they may represent traditional research fields with a great deal of knowledge exchange. This finding also confirms the discipline characteristics of CLIS that “library science and intelligence science are always one family”.

From a scholar’s point of view, the size of a square reflects the total quantity of the scholar’s research work. Figure 14 shows that the scholars “JP Qiu”, “ZJ Wang”, “SW Wang”, “HJ Yuan”, “YF Jiang”, “TN Wu”, “F Chen”, “Y Zhou”, and “Y Xiao” have the most discipline KNs and the greatest research breadth and depth in CLIS. These scholars represent the main sources of discipline innovation. Some of these scholars cross several sub-domains, indicating that they have a wide range of knowledge and can become a core figure in multiple sub-domains. One example of this phenomenon is “ZJ Wang”, whose knowledge covers the four sub-domains of “C1_University_Library”, “C1_Digital_Library”, “C1_Informatics”, and “C1_Competitive_Intelligence”. “HJ Yuan”, “F Chen”, “Y Xiao”, “GX Li”, “JY Ye”, “XC Luo”, “XM Xia”, “P Ke”, “YC Jiang”, “J Li”, and “F Wang” are other examples. The extent of this crossover also indicates the presence of many fields in related sub-domains among which knowledge can be integrated and exchanged, conducive to achieving subject knowledge innovation. In contrast, other scholars focus on one sub-domain and achieve astonishing depth of research, such as “JP Qiu” in “C1_Informatics”, “SW Wang” and “YF Jiang” in “C1_Public_Library”, “TN Wu” in “C1_Philology”, and “Y Zhou” in “C1_Library.”

Similarly, CLIS has 50 C2 sub-domains. The top 3 % of scholars associated with every C2 sub-domain are taken as core scholars; this produces 208 associations between scholars and sub-domains, encompassing 172 scholars. The results are shown in Fig. 15. In Fig. 15, C2 sub-domains from the same C1 sub-domain are represented with circles of the same color, separated with dashed lines and named with abbreviations using the first letters of the words in full names as the new name. Figure 15 looks incredibly complicated, because it contains too many sub-domains and scholars. With the exception of the absolutely independent “C1_Communication” (C1_C) shown in green nodes on the bottom-right corner, the C1 sub-domains are associated with each other, which is a different result from that indicated in Fig. 14. A few overlaps exist between core scholars in the C2 sub-domains of “C1_Search_Engine” (C1_SE), “C1_Philology” (C1_P), and “C1_Knowledge_Management” (C1_KM). This suggests that of all CLIS C1 sub-domains, “C1_C” is the most marginalized sub-domain, whose main researchers may not be core scholars in CLIS. “C1_C” is followed in this regard by “C1_P” and “C1_SE”. “C1_P” is a traditional CLIS field, and its sub-domain “C2_Bibliography” has some associations with “C2_Library_Science” in “C1_Digital_Library” (C1_DL) and “C2_Resource_Sharing” in “C1_KM”, suggesting that prominent “C1_P” researchers may transfer in the latter two directions. “C1_SE” is related to the study of information technology, and it is the product of crossover between CLIS and computer science; its sub-domain “C2_Personalization_service” has some associations with the studies of knowledge management and digital libraries.

Most scholars do not cross between C2 sub-domains, because only the top 3 % of scholars in a C2 sub-domain are defined as core scholars. In other words, this analysis highlights only the most critical researchers in each C2 sub-domain, resulting in reduced probability of domain-crossing. Additionally, most scholars’ research efforts are so limited that the majority of them concentrate their studies on only one field within a decade, because research concentration is the basic law of general scientific research. Of course, some C1 sub-domains feature a great deal of crossover in their C2 sub-domains, including “C1_Informatics” (C1_I) and especially “C1_Public_Library” (C1_PL), in which associations exist among all five C2 sub-domains. This feature suggests that these two sub-domains are smaller in scope, and the differences between them are not obvious. The

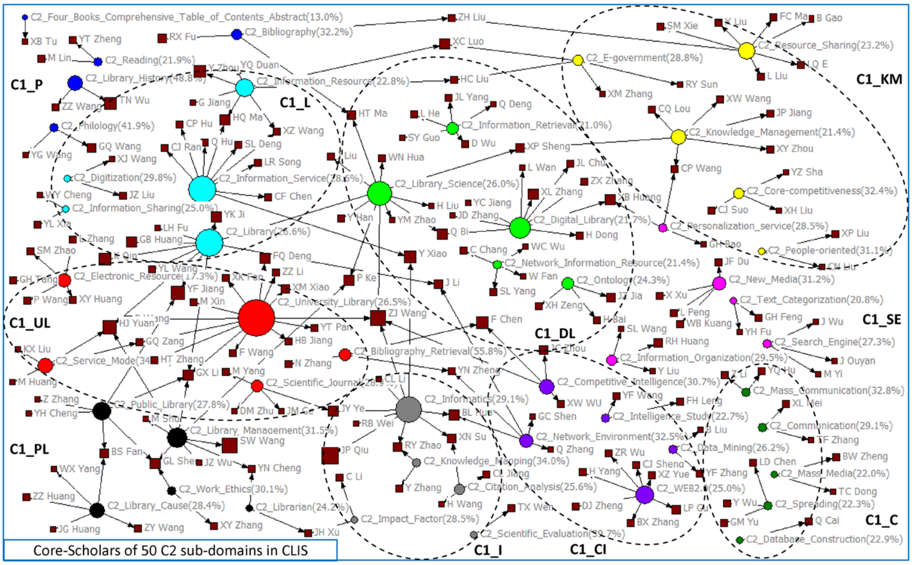


Fig. 15 Core scholars (top 3 %) of the C2 sub-domains in CLIS

content of these two sub-domains overlaps significantly, such as “C2_Infomatics”, “C2_Knowledge_Mapping”, and “C2_Citation_Analysis”. The latter two examples can almost be merged into one C2 sub-domain.

Core scholars account for 20–40 % of the studies in these C2 sub-domains, similar to the results at the C1 level. However, some special sub-domains are exceptions to this rule, such as “C2_Four_Books_Comprehensive_Table_of_Contents_Abstract” in “C1_P”. This is a C2 sub-domain defined by specific Chinese characteristics. The scope of research is very narrow, and only one scholar was selected. This suggests that researchers in this sub-domain are very scarce, and the distribution of their studies is relatively uniform. “C2_Library_History” in “C1_P” and “C2_Bibliography_Retrieval” in “C1_University_Library” (C1_UL) are just the opposite, and the research coverage of their core scholars is close to or even above 50 %. The studies in these C2 sub-domains are relatively concentrated, such that some outstanding core figures appear and there are few related scholars.

Results of macroscopic analysis: relationships among scholars

Following the methods described above and using CKNs on level C1 as the basis for associations, all of the correlation coefficients between each pair of scholars can be calculated. M is set to 15, such that the 15 scholars with the highest correlation coefficients will be identified as the associated scholars for every scholar. Of course, M also can be set to other values, according to the analyst’s actual demand. In this example, 11,553 associations on level C1 were built among 575 scholars, in which the maximum correlation is 1 and the minimum is 0.052954. The number of C1 sub-domains was small, yielding a small value for N in formula (11) and generating many identical values in terms of correlation coefficients. No fewer than 15 associated objects were identified for every scholar, with one having 73 associated objects with the same coefficient. Calculating and mining associated scholars on level C1 is a useful exercise for scholars who want to seek content-related peers for the purposes of academic discussion or research collaboration within the range of a C1 sub-domain.

Associated scholars on level C1 were identified for the two authors of this study, "XN Su" and "H Wang", with the results shown in Fig. 16. Then, the "School of Information Management of Nanjing University" was taken as an example to depict associations among 17 scholars selected in this study. These results are shown in Fig. 17, to aid in further analysis of the association status of their studies on level C1 level in CLIS.

In Fig. 16, the circles at the center are the central scholars (*X*), and the squares distributed around them are the associated scholars (*Y*). The sizes of the shapes represent the total knowledge coefficient of a scholar in his level C1 research focus, and the thicknesses of the lines linking scholars represent the strength of associations between them. Additionally, the associated scholars are coded with three colors. Associated scholars shared by the two central scholars are colored blue, scholars associated only with "XN Su" are colored pink, and those associated only with "H Wang" are colored green. The two central scholars have a great deal of common ground, in that they share 13 associated scholars, excluding themselves. "XN Su" is also associated with "ZZ Wang" and "TX Wen", while "H Wang" is associated with "YC Jiang" and "Y Ye". However, the unshared associated scholars have little effect on the central scholars, judging from the thickness of the lines connecting them. In other words, all of the most important associated scholars are shared, and the two central scholars are also associated with each other. These results demonstrate that the two central scholars have a great deal of commonality in their studies, such that they can cooperate closely and explore together on level C1 topics.

Looking at the unshared associated scholars, the knowledge coefficients of the two scholars associated only with "XN Su" are far larger than those of the two scholars associated with "H Wang". The former's associated nodes are also larger than the latter's, indicating that the current research depth or intensity of "XN Su" is much larger than that

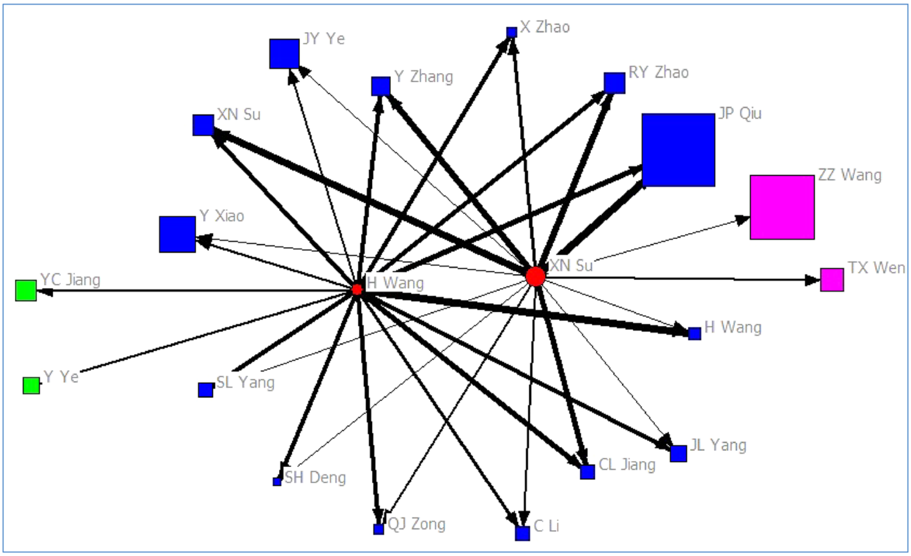


Fig. 16 Associated scholars for XN Su and H Wang in C1 sub-domains of CLIS_KS

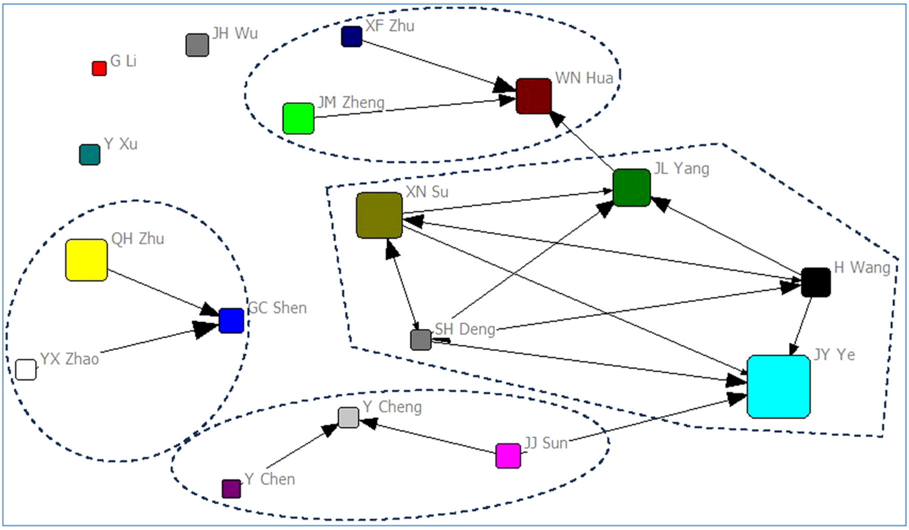


Fig. 17 Social network diagram of 17 scholars from the School of Information Management of Nanjing University in C1 sub-domain of CLIS_KS

of “H Wang”. This means that “XN Su” is more strongly associated with scholars who have larger research efforts. In other words, scholars associated with “XN Su” have more specialized and deeper research than those associated with “H Wang”. In terms of association strength, the method used in this study to calculate association strength is global in nature. Therefore, the strengths of different scholars are comparable. On the whole, the association coefficients between “XN Su” and the shared associated scholars are larger than the ones between them and “H Wang”. In particular, “JP Qiu”, “RY Zhao”, “Y Zhang”, and “CL Jiang” all have very close contact with the “XN Su” in terms of content, and they are his most important associated scholars; in contrast, the influence of these scholars on “H Wang” is not as great. This finding suggests that these scholars have more common studies with the former; furthermore, if these common studies are all co-authored by both “XN Su” and “H Wang”, the former’s research efforts were much larger than the latter’s. The other scholars with the greatest associations with “H Wang” are “JL Yang”, “QJ Zong”, “SH Deng”, “SL Yang”, and “Y Xiao”. These results show that scholars with more significant research efforts are more easily influenced by a scholar with a similar research level, while associated scholars with low research levels have little influence on them.

In Fig. 17, the round-cornered rectangular boxes represent scholars, and their sizes express the scholar’s research efforts or research level in C1 sub-domains, as expressed by knowledge coefficients. The links show associations between scholars, with the arrows pointing to the associated scholar and the sizes representing the degree of dependence or association. The 17 scholars from the School of Information Management of Nanjing University have few associations with one another, indicating that their studies are quite dispersed on level C1. This school has certain representative figures in distinct C1 sub-domains, and their research has a certain systematicness. “JH Wu”, “Y Xu”, and “G Li” form three islands, indicating that their studies have some independence and that they are the representative figures for “C1_Library”, “C1_Philology”, and “C1_Knowledge_Management”, respectively. In particular, “JH Wu” is the representative figure for “Archival Science”, but this subject has such a small research scale that it has not been automatically clustered into one class in CLIS; instead, it is included in “C1_Library”. These three research directions are so minor in this school that they deserve more attention for enhancing comprehensiveness of the school’s research.

On the left side of Fig. 17, “QH Zhu”, “YX Zhao”, and “GC Shen” form a group characterized by unidirectional dependence; in other words, “QH Zhu” and “YX Zhao” are associated scholars for “GC Shen”, who is a representative figure in “C1_Competitive_Intelligence”. The core research fields of the other two scholars have been included in this domain, allowing them to form a small group in competitive intelligence research. At the bottom of Fig. 17, “Y Cheng”, “Y Chen”, and “JJ Sun” form a dependent group that is also a unidirectional dependent association. “Y Cheng” represents the core of this group, whose representative BKN are all included in “C1_University_Library”. The main sub-domains of the other two are also “C1_University_Library”, creating the dependence on “Y Cheng”. However, this group is not entirely independent; in addition to this sub-domain, “JJ Sun” has an important connection to the sub-domain “C1_Informatics”, giving him another dependence on “JY Ye”. At the top of Fig. 17, “XF Zhu”, “JM Zheng”, and “WN Hua” form a small group whose research field is “C1_Digital_Library”. It should be noted that “XF Zhu” and “JM Zheng” have other sub-domains, but the total number of their studies is small, such that they have not become the dependent objects of other scholars and no scholar has become their dependent objects.

The largest research group in this school appears in the middle of Fig. 17. This group is made up of five scholars with many complicated dependent associations among themselves. The core field of this group is “C1_Informatics”. The calculation method used means that the larger and the more concentrated the total studies of a scholar are, the easier for the scholar to become the dependent object of other scholars. Therefore, the interdependences of this group suggests that this school concentrates significant research efforts in “C1_Informatics”, and there are also a number of scholars involved in “C1_Digital_Library”. Therefore, this group takes “C1_Informatics” and “C1_Digital_Library” as core research fields. Finally, the School of Information Management of Nanjing University has other sub-domains with some representative figures, in addition to the small contributions of “C1_Search_Engine” and “C1_Public Library”. Especially in regards to “C1_Informatics”, the school has attained a certain scale of study, with a plurality of scholars demonstrating research strengths.

Conclusions

Using keywords from CSSCI (2003–2012) papers and co-occurrence associations between keywords and scholars, this study completely constructed a narrow-sense CLIS_KS with keywords as BKN. Then, this study applied cohesion and coupling from part of CKN to verify the rationality of CLIS_KS. Subsequently, this study deeply analyzed and discussed the KS of renowned scholars, the related core-scholars of CKNs, and the associations among the scholars in CLIS from the microscopic, mesoscopic and macroscopic levels. This study provides a reference for the construction and applications of DKS based on academic resources, and it plays an enlightening and driving role in summarizing and combing discipline knowledge from CLIS. Moreover, applying these modes and conclusions to the analysis of other academic objects could promote discipline construction and development.

This study makes novel contributions through the automatic construction and subsequent application of a complete DKS. This kind of multi-level knowledge structure has not been fully presented in past studies; instead, co-citation analysis or co-word analysis have formed the main basis for judging similarity among academic objects such as words, literature, scholars, institutions, and disciplines. Author co-citation analysis provides a basis for establishing similarity between authors, while co-word analysis provides a reference for similarity in the word use of academic objects. In essence, this study is a variant of co-word analysis. However, in the past, methods of co-citation analysis or co-word analysis have mostly been employed in detecting and analyzing hotspots related to the most famous scholars or the most commonly used terms in a certain discipline. These hotspots only represent the partial knowledge structure of a discipline, and they are completely unable to cover the entire discipline or present a multi-level structure.

Some points in this study require further refinement and improvement in future studies, including eliminating the uncertainty of clustering, naming the CKNs and detecting the scholars' KS trend in microscopic analysis, determining the core-scholars in mesoscopic analysis, and judging the associations between scholars on distinct-granularity levels in macroscopic analysis.

Acknowledgments This work was supported by Jiangsu Province's Natural Science Foundation Project named "Study on Chinese Ontology Learning-Oriented Patent Forewarning" (No. BK20130587), as well as a major program of the National Social Science Foundation of China named "Studies on Deep Polymerization and Services of Network Information Resource-Oriented Discipline Field" (No. 12&ZD221).

References

Aleixandre, J. L., Aleixandre-Tudo, J. L., Bolanos-Pizarro, M., et al. (2015). Mapping the scientific research in organic farming: A bibliometric review. *Scientometrics, 105*(1), 295–309.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for windows: Software for social network analysis*. Harvard, MA: Analytic Technologies.

Calabretta, G., Durisin, B., & Ogliengo, M. (2011). Uncovering the intellectual structure of research in business ethics: A journey through the history, the classics, and the pillars of Journal of Business Ethics. *Journal of Business Ethics, 104*(4), 499–524.

Chang, Y. W. (2012). Tracking scientometric research in Taiwan using bibliometric and content analysis. *Journal of Library and Information Studies, 10*(2), 1–20.

Charvet, F. F., Cooper, M. C., & Gardner, J. T. (2008). The intellectual structure of supply chain management: A bibliometric approach. *Journal of Business Logistics, 29*(1), 47–73.

Chen, C. M. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology, 57*(3), 359–377.

Chen, L. C., & Lien, Y. H. (2011). Using author co-citation analysis to examine the intellectual structure of e-learning: A MIS perspective. *Scientometrics, 89*(3), 867–886.

Chen, C. M., & Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *Computer, 34*(3), 65–71.

Cho, J. (2014). Intellectual structure of the institutional repository field: A co-word analysis. *Journal of Information Science, 40*(3), 386–397.

Danell, J. A. B. (2014). Reception of integrative and complementary medicine (ICM) in scientific journals: A citation and co-word analysis. *Scientometrics, 98*(2), 807–821.

Darvish, H., & Tonta, Y. (2016). Diffusion of nanotechnology knowledge in Turkey and its network structure. *Scientometrics, 107*(2), 569–592.

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. UK: Cambridge University Press.

Dehdarirad, T., Villarroya, A., & Barrios, M. (2014). Research trends in gender differences in higher education and science: A co-word analysis. *Scientometrics, 101*(1), 273–290.

Erserim, A. (2016). Intellectual structure of accounting research: A historical review on the Journal of Accounting Organization Society. *Accounting and Finance Research, 5*(2), 1–9.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM SIGKDD*, Portland, Oregon (pp. 226–231).

Falconer, S. (2015). OntoGraf. http://protegewiki.stanford.edu/wiki/Onto-Graf. Accessed by 2015-07-01.

Galvagno, M. (2011). The intellectual structure of the anti-consumption and consumer resistance field: An author co-citation analysis. *European Journal of Marketing, 45*(11–12), 1688–1701.

Garcia-Lillo, F., Ubeda-Garcia, M., & Marco-Lajara, B. (2016). The intellectual structure of research in hospitality management: A literature review using bibliometric methods of the journal International Journal of Hospitality Management. *International Journal of Hospitality Management, 52*, 121–130.

Gonzalez-Alcaide, G., Castello-Cogollos, L., Navarro-Molina, C., et al. (2008). Library and information science research areas: Analysis of journal articles in LISA. *Journal of the American Society for Information Science and Technology, 59*(1), 150–154.

Hong, Y., Yao, Q., Yang, Y., et al. (2016). Knowledge structure and theme trends analysis on general practitioner research: A co-word perspective. *BMC Family Practice, 17*, 10.

Hooper, R. P. (2009). Towards an intellectual structure for hydrologic science. *Hydrological Processes, 23*(2), 353–355.

Hossain, M. G., Reza, A. H. M. S., Lutfun-Nessa, M., et al. (2013). Factor and cluster analysis of water quality data of the groundwater wells of Kushtia, Bangladesh: Implication for arsenic enrichment and mobilization. *Journal of the Geological Society of India, 81*(3), 377–384.

Hu, C. P., Hu, J. M., & Deng, S. (2013). A co-word analysis of library and information science in China. *Scientometrics, 97*(2), 369–382.

Hult, G. T. M. (2016). JAMS 2010–2015: Literature themes and intellectual structure. *Journal of the Academy of Marketing Science, 43*(6), 663–669.

Jeong, S., & Kim, H. G. (2010). Intellectual structure of biomedical informatics reflected in scholarly events. *Scientometrics, 85*(2), 541–551.

Kim, H., & Lee, J. Y. (2008). Exploring the emerging intellectual structure of archival studies using text mining: 2001–2004. *Journal of Information Science, 34*(3), 356–369.

Kumar, K. M., & Reddy, A. R. M. (2016). A fast DBSCAN clustering algorithm by accelerating neighbor searching using groups method. *Pattern Recognition, 58*, 39–48.

Kurihara, T., Tomari, N., & Aratani, T. (2013). Trend of EASTS research in the past 20 years. In *Proceedings of the eastern Asia society for transportation studies* (Vol. 9).

Leydesdorff, L., & Vaughan, L. (2006). Co-occurrence matrices and their applications in information science: Extending ACA to the Web environment. *Journal of the American Society for Information Science and Technology, 57*(12), 1616–1628.

Liu, Z. (2005). Visualizing the intellectual structure in urban studies: A journal co-citation analysis (1992–2002). *Scientometrics, 62*(3), 385–402.

Ma, R. M. (2012). Discovering and analyzing the intellectual structure and its evolution of LIS in China, 1998–2007. *Scientometrics, 93*(3), 645–659.

Ma, R., & Ni, C. (2011). The intellectual structure and its evolution of LIS in China based on the co-citation analysis. *Journal of Library Science in China, 6*, 17–26.

Machado, R. D., Vargas-Quesada, B., & Leta, J. (2016). Intellectual structure in stem cell research: Exploring Brazilian scientific articles from 2001 to 2010. *Scientometrics, 106*(2), 525–537.

McCain, K. W. (1990). Mapping authors in intellectual space: A technical overview. *Journal of the American Society for Information Science, 41*(6), 433–443.

Meyer, M., Zaggl, M. A., & Carley, K. M. (2011). Measuring CMOT's intellectual structure and its development. *Computational and Mathematical Organization Theory, 17*(1), 1–34.

Milojevic, S., Sugimoto, C. R., Yang, E. J., et al. (2011). The cognitive structure of library and information science: Analysis of article title words. *Journal of the American Society for Information Science and Technology, 62*(10), 1933–1953.

Naghizadeh, R., Elahi, S., Manteghi, M., et al. (2015). Through the magnifying glass: An analysis of regional innovation models based on co-word and meta-synthesis methods. *Quality & Quantity, 49*(6), 2481–2505.

Nerur, S. P., Rasheed, A. A., & Natarajan, V. (2008). The intellectual structure of the strategic management field: An author co-citation analysis. *Strategic Management Journal, 29*(3), 319–336.

Olijnyk, N. V. (2015). A quantitative examination of the intellectual profile and evolution of information security from 1965 to 2015. *Scientometrics, 105*(2), 883–904.

Otte, E., & Rousseau, R. (2002). Social network analysis: A powerful strategy, also for theinformation sciences. *Journal of Information Science, 28*(6), 441–453.

Park, H. W., & Leydesdorff, L. (2008). Korean journals in the Science Citation Index: What do they reveal about the intellectual structure of S & T in Korea? *Scientometrics, 75*(3), 439–462.

Persson, O., Danell, R., & Wiborg Schneider, J. (2009). How to use Bibexcel for various types of bibliometric analysis. In Åström, F., Danell, R., Larsen, B., & Schneider, J. (Eds.), *International society for scientometrics and informetrics*, Leuven, Belgium (pp. 9–24).

Pilkington, A., & Meredith, J. (2009). The evolution of the intellectual structure of operations management-1980–2006: A citation/co-citation analysis. *Journal of Operations Management, 27*(3), 185–202.

Pinto, M. (2015). Viewing and exploring the subject area of information literacy assessment in higher education (2000–2011). *Scientometrics, 102*(1), 227–245.

Powers, J. H. (1995). On the intellectual structure of the human-communication discipline. *Communication Education, 44*(3), 191–222.

Pratt, J. A., Hauser, K., & Sugimoto, C. R. (2012). Defining the intellectual structure of information systems and related college of business disciplines: A bibliometric analysis. *Scientometrics, 93*(2), 279–304.

Prebor, G. (2010). Analysis of the interdisciplinary nature of library and information science. *Journal of Librarianship and Information Science, 42*(4), 256–267.

Protégé. http://Protege.stanford.edu. Accessed by July 1st, 2015.

Ravikumar, S., Agrahari, A., & Singh, S. N. (2015). Mapping the intellectual structure of scientometrics: A co-word analysis of the Journal Scientometrics (2005–2010). *Scientometrics, 102*(1), 929–955.

Riviera, E. (2015). Testing the strength of the normative approach in citation theory through relational bibliometrics: The case of Italian sociology. *Journal of the Association for Information Science and Technology, 66*(6), 1178–1188.

Rusk, R. D., & Waters, L. (2015). A psycho-social system approach to well-being: Empirically deriving the five domains of positive functioning. *Journal of Positive Psychology, 10*(2), 141–152.

Samiee, S., & Chabowski, B. R. (2012). Knowledge structure in international marketing: A multi-method bibliometric analysis. *Journal of the Academy of Marketing Science, 40*(2), 364–386.

Sarafis, I. A., Trinder, P. W., & Zalzala, A. M. S. (2007). Nocea: A rule-based evolutionary algorithm for efficient and effective clustering of massive high-dimensional databases. *Applied Soft Computing, 7*(3), 668–710.

SAS. http://www.sas.com/en_us/home.html. Accessed by July 1st, 2015.

Seyedghorban, Z., Matanda, M. J., & LaPlaca, P. (2016). Advancing theory and knowledge in the business-to-business branding literature. *Journal of Business Research, 69*(8), 2664–2677.

Sluyter, A., Augustine, A. D., Bitton, M. C., et al. (2006). The recent intellectual structure of geography. *Geographical Review, 96*(4), 594–608.

Song, M., & Kim, S. Y. (2013). Detecting the knowledge structure of bioinformatics by mining full-text collections. *Scientometrics, 96*(1), 183–201.

SPSS. http://www-01.ibm.com/software/analytics/spss/. Accessed by July 1st, 2015.

Su, X. (2007). *Report on the academic influence of research achievement in China's humanities and social sciences*. Beijing: China Social Sciences Publishing House.

Su, X., & Zou, Z. (2011). *Report on the academic influence of research achievement in China's humanities and social sciences (2011)*. Beijing: Higher Education Press.

Torres-Salinas, D., & Moed, H. F. (2009). Library catalog analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics. *Journal of Informetrics, 3*(1), 9–26.

Triventi, M. (2014). Higher education regimes: An empirical classification of higher education systems and its relationship with student accessibility. *Quality & Quantity, 48*(3), 1685–1703.

Tseng, Y. H., & Tsay, M. Y. (2013). Journal clustering of library and information science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics, 95*(2), 503–528.

Van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics, 84*(2), 523–538.

Wang, H. (2010). *Information resource network model and its application*. Nanjing: Nanjing University Press.

Wang, C. H., Lee, Y. D., Chou, H. L., & Kuo, J. H. (2014). Identifying the intellectual structure of risk management studies. In *Proceedings of IEEE workshop on electronics, computer and applications (IWECA)*, Ottawa, Canada (pp. 964–968).

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.

Wolfram, D., & Zhao, Y. H. (2014). A comparison of journal similarity across six disciplines using citing discipline analysis. *Journal of Informetrics, 8*(4), 840–853.

Yan, B. N., Lee, T. S., & Lee, T. P. (2015). Mapping the intellectual structure of the internet of things (IoT) field (2000–2014): A co-word analysis. *Scientometrics, 105*(2), 1285–1300.

Yang, S. L., Han, R. Z., Wolfram, D., et al. (2016). Visualizing the intellectual structure of information science (2006–2015): Introducing author keyword coupling analysis. *Journal of Informetrics, 10*(1), 132–150.

Yoo, Y. J., Lee, J. Y., & Choi, S. (2013). Intellectual structure of Korean theology 2000–2008: Presbyterian theological journals. *Journal of Information Science, 39*(3), 307–318.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD international conference on management of data* (Vol. 25, pp. 103–114).

Zong, Q. J., Shen, H. Z., Yuan, Q. J., et al. (2013). Doctoral dissertations of Library and Information Science in China: A co-word analysis. *Scientometrics, 94*(2), 781–799.