情报学报 ISSN 1000-0135
第34卷第2期 115-128, 2015年2月
JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION ISSN 1000-0135
Vol. 34 No. 2, 115-128 February 2015

doi:10.3772/j.issn.1000-0135.2015.002.001

我国图书情报学科知识结构的建立及其演化分析[1)]

王 昊[1,2] 邓三鸿[1,2] 苏新宁[1]

(1. 南京大学信息管理学院,南京 210093;
2. 威斯康辛-密尔沃基大学信息研究学院,威斯康辛州,美国 53211)

摘要 本文以本体思想作为理论指导,将狭义的学科知识结构理解为学科知识点的层次体系,进而借助本体学习技术从CSSCI期刊论文集合中衍生出CLIS的学科知识结构,并对其在2003~2010年10年间的发展轨迹进行了初步探测。具体过程包括:基于热点关键词及核心作者筛选的关键词-学者矩阵构建,基于多层k-means聚类的CLIS知识结构生成,基于OWL的CLIS知识结构本体描述和可视化展示,基于层次聚类和多维尺度分析的知识类目内聚性和耦合性检测以及基于不同粒度知识映射的CLIS知识类目发展轨迹分析等。经过验证,论文构建的CLIS知识体系具有一定的合理性和有效性,为后期该学科广义知识结构的研究奠定了知识基础。

关键词 CLIS 学科知识结构 本体 关键词学者矩阵 多层聚类 层次聚类 多维尺度分析 演化分析

Construction and Evolution Analysis on Disciplinary Knowledge Structure of Chinese LIS

Wang Hao[1,2], Deng Sanhong[1,2] and Su Xinning[1]

(1. *School of Information Management of Nanjing University, Nanjing* 210093;
2. *School of Information Studies of University of Wisconsin-Milwaukee, WI, USA* 53211)

Abstract Taking the idea of ontology as a theoretical guidance, this paper interprets the disciplinary knowledge structure in narrow sense as a hierarchical system of disciplinary knowledge points. Then the disciplinary knowledge struetare of CLIS is derived from CSSCI papers with the help of ontology learning technology, and its development track over ten years (2003-2010) is detected. The specific process includes the building of keywords-scholars matrix based on filtering for hot keywords and core authors in CLIS, the generation of CLIS knowledge structure based on multi-level k-means clustering, ontology description and visual display for CLIS knowledge structure based on OWL, the checking for cohesion and coupling of knowledge categories based on hierarchical clustering and multidimensional scaling analysis, and the analysis for development track of CLIS knowledge categories based on knowledge mapping on different granularity levels. After verification, the knowledge system of CLIS constructed in this paper has a certain rationality and effectiveness, and it could lay the foundation for the latter study on broad knowledge structure of this discipline.

Keywords Chinese Library and Information Science (CLIS), disciplinary knowledge structure, ontology, Keywords-Scholars Matrix (KSM), multi-level clustering, Hierarchical Clustering (HC), Multidimensional Scaling Analysis (MDSA), *evolution analysis*

收稿日期:2014年8月20日

作者简介:**王昊**,男,1981年生,南京大学信息管理学院情报学博士,副教授,在校主要从事知识本体构建及应用、数据挖掘技术应用,科学评价和引文分析等研究。E-mail:ywhaowang@nju.edu.cn。**邓三鸿**,男,1975年生,南京大学信息管理学院情报学博士,副教授,主要从事科学评价和引文分析,知识管理与知识地图等研究。**苏新宁**,男,1955年生,南京大学信息管理学院教授,博导,长江学者,主要从事智能信息处理与检索、科学评价和引文分析等研究。

1) 本文受国家社科重大招标项目"面向突发事件应急决策的快速响应情报体系研究"(13&ZD174)和"面向学科领域的网络信息资源深度聚合与服务研究"(12&ZD221)等的资助。

1 引 言

学科知识是对学科研究主题确信的认识，是学科发展的主要动力和学科创新的核心内容。对学科知识的揭示可以从两个方面进行，一是组织并描述知识的静态结构[1]，探索某一时期学科知识点间的语义关联，为学科知识分布的图谱展示以及知识的合理利用奠定基础；二是模拟并追踪知识的动态轨迹[2]，探讨学科知识点在一定历史时期内的发展规律及其可能的原因，从而引导学科研究人员的知识创新。

对学科知识结构及其发展脉络的梳理，不仅可以帮助理解学科内涵，树立并激励学科研究的信心，从而完善学科研究内容，促进学科创新和发展，而且在充分揭示学科知识语义关系的基础上，可以进一步探索学科内其他学术资源如学者、机构、期刊等与学科知识的语义关联，揭示它们的知识结构和分布，甚至借此挖掘并描绘出这些对象自身之间的潜在关联，从而实现对学科总体内涵的完整认识，这实际上就是广义的学科知识结构。图1展示出了CSSCI中不同粒度的知识及其相互关系。若从狭义角度而言，“关键词”是学科知识的主要代表，不同范围和数量“关键词”的组合可以形成学科知识结构；但从广义角度而言，“学者”、“机构”、“地区”以及“论文”、“期刊”和“学科”等都是以“关键词”为核心的不同粒度的学科知识。因此对学科知识结构的探讨可以从狭义逐步扩展到广义。

本文试图以本体思想为理论指导，借助本体学习技术，以多层聚类方法自动实现我国图书情报学科(Chinese Library and Information Science, CLIS)知识结构的建立；进而采用层次聚类(Hierarchical Clustering, HC)和多维尺度(Multidimensional Scaling, MDS)等分析方法对生成的知识类目进行内聚性和耦合性检测，以验证CLIS知识结构的正确性和合理性；在此基础上，对2003~2012年10年间学科不同粒度知识的发展轨迹进行探测，分析其变化的原因。鉴于篇幅，本文仍以学科狭义知识结构的建立和应用为主要内容，而广义知识结构的实质是狭义知识结构在学科其他学术对象上的扩展，笔者将另行撰文探讨。

2 近期相关研究

当前对学科知识结构及演化的探讨多采用两种方式。一是领域专家根据自身知识背景、研究经验以及前人研究成果，对本学科知识结构进行主观综合和定性描述[3]；二则是引入文献计量学方法，面向领域探索各类评价指标的计算[4]，或揭示领域中各种研究单元之间的交互关联[5]等对领域知识结构进行客观描述和定量分析。随着文献计量学的逐渐成熟，研究人员越来越倾向于利用学术对象之间存在的客观关系来揭示学科研究热点及其相互关联[6,7]。其基本计算模式如下：选择具体学科数据，根据客观关系对学术单元进行向量描述，聚集相似学术单元以生成领域热点，最终利用结构图分时段展示学科热点之间的交互关联及其变化情况。该过程存在以下几个特点：

(1)分析的基础来源于多种类型的数据源，包括图书[8]、期刊论文[9]、学位论文[10]、信件、评论、会议论文[11]及其他学术事件如workshops、symposia、seminars等[12]。

(2)学科内大粒度知识点，如研究方向等都是通过学术单元按照一定规则聚类而成。其中学术单元包括术语[13]、作者[14]、论文[15]、期刊[16]等，学术单元之间的相似处即为学科知识点，而判断相似的

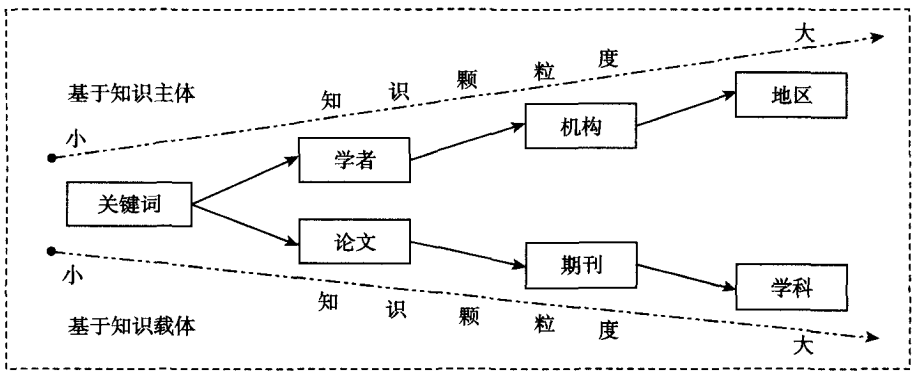

图1 CSSCI中知识的粒度变化情况

标准主要有学术单元之间的同被引(Co-citation)[14,15]和共现(Co-occurrence)关系[13,17]。

(3)实现学术单元间基于相似而聚集的方法主要有层次聚类分析(HCA)[18]、多维尺度分析(MDSA)[19]以及因子分析(Factor Analysis, FA)[20];而进一步描述大粒度知识点间关联结构的方法主要是社会网络分析(Social Network Analysis, SNA)[21]和 Pathfinder 网络分析(PF-Net Analysis)[22]等。

(4)探讨的多是狭义的学科或领域知识结构[20],甚少以学者或机构等学术单元作为主体;其中知识主要指大粒度知识点,即学科或领域内的研究方向或热点[23];范围则包括了 LIS[23,24]、生物医学信息学[12]、E-Learning[14]、战略管理[25]、国际营销[26]、神学[5]等。

(5)知识结构分析常借助于各种工具,包括能够构建同被引/共现矩阵并计算相关系数的 Bibexcel[27];能够实现 HCA、MDSA 和 FA 等的 SPSS[28]和 SAS[29];用于构建知识点间关联并可计算结构特征的 Ucinet[30]、Pajek[31]、ORA[32]以及 pathfiner 算法[33]等;可实现知识点关联分布和时间轴演化分析的 CiteSpace[34]等。

综上所述,笔者发现目前对 LIS 学科知识结构及其发展脉络的探索具有规模小,欠完整,局部性等问题。①在研究中大量使用了 HCA、MDSA 以及 FA 等方法来实现学科知识点的生成,而这些方法所针对的数据量都不可能太大,例如 HCA 就是典型的小规模高精度的聚类方法;②生成的知识点都是学科/领域的研究热点或方向,粒度较大,或者说所谓的学科知识结构仅仅指学科顶层知识类目及其关联,而对类目中的具体情况则分析不足,即对学科不同粒度知识点间的语义关系描述欠完整;③基于部分知识结构进行的分析具有局部性,只能了解学科中某个指定方面或角度的状况。例如基于高被引作者或高频术语的学科知识结构分析[24,35],得到的只能是这部分学者所涉猎或这部分术语所描述的研究方向,而非整个学科的知识结构。因此,本文试图将学科知识结构理解为领域知识层次体系,以本体的视角全面解析学科内知识点间的语义关联,从而构建较为完整的狭义知识结构,为探讨学科广义知识结构奠定知识基础。

3 基于 CSSCI 的 CLIS 学科知识结构建立

期刊论文既包含了学科中原有的知识基础,同时也是记录学科新知识的重要方式,是学科知识的主要载体和来源。本节即利用 CSSCI 期刊论文中所蕴含的知识,来构建 CLIS 的学科知识结构。

3.1 CSSCI 数据清洗

CSSCI 全称"中国社会科学引文索引"(Chinese Social Sciences Citation Index),是南京大学社会科学评价中心遴选出 400 ~ 500 种我国人文社会科学精品期刊,收录与其相关的论文、关键词、作者、被引文献及其他学术资源等所构成的引文数据库。可以认为,CSSCI 中蕴含了我国人文社科科学中最前沿、最完整的学科知识。本文从 CSSCI 中检索出了2003 ~ 2012 年"学科"为"870"(即 LIS)的所有论文及其作者作为数据基础来构建学科知识结构,共计论文 58 281 篇,学者 34 222 个,关键词 67 351 个。

然而,直接以原始数据作为实验样本存在学者重名、非领域性主题、边缘学者以及知识点与学者之间存在偶然性关联等诸多问题。因此,有必要对原始数据进行清洗,以获得较为规范并规模适中的实验样本。CSSCI 数据清洗的思路主要有两个方向。

(1)学者的重名处理

笔者曾考虑采用"人名 + 省份(即邮编前两位)"作为区分学者的唯一标记,结果发现大量同一学者出现了多个标记。有的是由于错误标引造成的,如"武汉大学"(42)学者"邱均平"被 2 次错误标引为"湘潭大学"(43),另 2 次错误标引为"中国社会科学院"(11);也有的是由于单位跨省变动导致的,如"国家图书馆"(11)学者"索传军",曾长期就职于"郑州大学"(41),"华东师范大学"(31)学者"许鑫"曾在"南京大学"(32)攻读博士学位;本为了消除重名现象而引入地区符号,反而增加了数据的混乱程度,再加上一个学科中的重名学者相对较少的客观因素,本文最终决定以"人名"作为区分学者的标记,即若经过数据筛选后仍然存在重名现象,则将其视为一个学者处理。

(2)数据的合理性筛选

本文从关键词、学者以及两者间关联等 3 个角度来消除非领域性主题、边缘学者以及偶然性关联对领域知识及其结构生成的影响。①关键词频率阈值(K)。图 2 显示了不同等级下近 10 年(0 ~ 25 分为 6 等级)及近 5 年(0 ~ 15 分为 6 等级)关键词频次的变化情况。由图可知,当两者等级均为 2 时,即 K10≥5 且 K5≥3 时,关键词频次的变化开始符合趋势线,其变化呈现出一定的规律,笔者认为满足该条

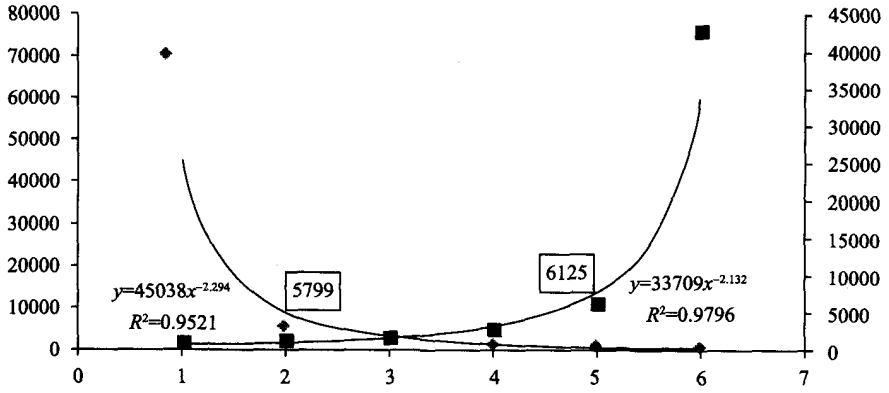

图2 不同数量等级下近10年(左)和近5年(右)关键词频次的变化情况

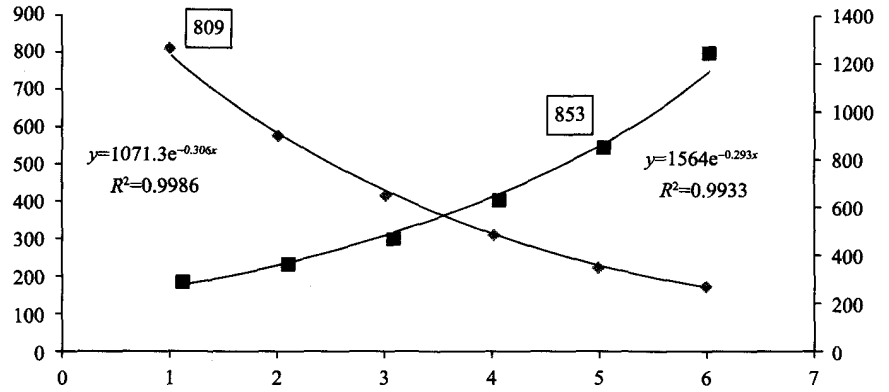

图3 不同发文量等级下近10年(左)和近5年(右)学者数量的变化情况

件的关键词被领域认可。②学者发文阈值(A)。仅以第一作者统计发文量,可得近10年(10~20分为6等级)和近5年(5~10分成6等级)发文量级别与学者数量变化如图3所示。同理,取A10≥10且A5≥6,即唯有10年间发表论文10篇及以上,且最近5年间发表6篇及以上的作者才能入选为CLIS的学者。③关键词与学者的关联阈值(W)。学者通过论文发表与关键词之间存在语义关联,其关联强度由学者在论文中的排名及关键词权重所决定;表1列出了作者人数及署名顺序与其对论文的贡献度,若假设论文中关键词的权重均为1,那么据此可计算出论文中所有学者与各关键词之间的关联强度,并最终获得学者和关键词之间的总关联系数。该系数反映了学者对相关知识点的掌握和应用程度,但其中可能存在大量的偶然性关联。为此,笔者设定$W>0.6$,即学者至少在1篇以上多作者论文中以第一作者身份所使用的关键词,与学者之间才存在必然关联。

经过数据清洗,最终被CLIS认可的有效关键词有3081个,学科重要学者575名,以及两者之间的语义关联12 005对;从学者和关键词两个角度来筛选数据,同名学者往往因研究主题不同而均入选的概率相对较低。由此获得的<关键词,学者,关联系数>三元组可作为CLIS学科知识结构生成的数据基础。

3.2 基于多层聚类的CLIS学科知识结构生成

关键词是粒度最小的学科知识,对关键词进行不同程度的聚合可以生成不同粒度大小的知识点,将所有知识点整合在一起即形成了完整的学科知识结构,整个过程如图4所示。首先,CLIS学科内的所有关键词一起构成了最大粒度知识点,记为CLIS_KS;然后根据对象相似原理,分别将具有较大相似度的关键词聚集在一起,形成若干个簇(类),每个簇即为粒度相对较小的学科知识点,记为C1_KS;接着具有较多关键词或关键词离散程度较大的簇可以进一步聚类,簇内相似程度较大的关键词被进一步聚集在一起,于是C1级别的知识点被拆分为粒度更小的知识点C2_KS;上述过程持续执行,于是大粒度知识点被不断拆分为内聚性更强的低粒度知识点,直到簇内关键词数量减少到指定值或簇内关键词相似程度相当高为止。

表 1　作者人数及署名顺序与作者对论文贡献度的对照表 *

作者人数 \ 署名顺序	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

* 没有考虑通讯作者，第 7 作者及其以后不考虑其贡献，论文总贡献为 1。

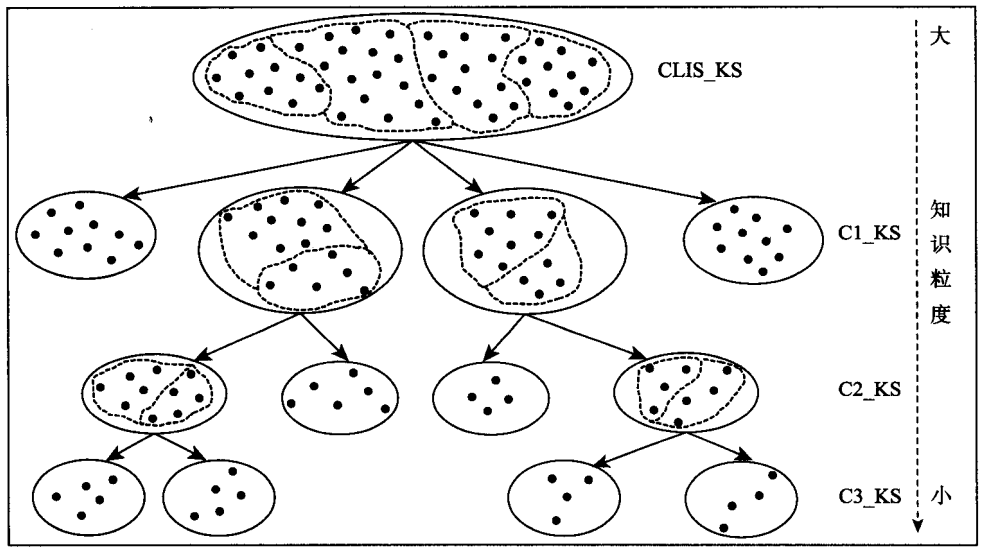

图 4　CLIS 学科知识结构的建立过程

根据上述基本思路，笔者采用多层聚类的方法来实现不同粒度知识点的自动生成。在具体操作之前，需要解决若干个问题。①首先是聚类算法的选择。由于本文面对规模较大的关键词对象，可采用具有较高效率的基于划分的 K-means 聚类算法来实现不同粒度知识点的聚合，再以 HC 对最小的知识集合进行细分以验证知识聚类的效果。②关键词的向量描述。本文以关键词作为聚类对象，因此需要对关键词进行特征描述。之前在数据清洗中，已然形成 < 关键词，学者，关联系数 > 三元组，因此可以学者作为关键词的描述特征，构造关键词 - 学者矩阵（Keywords-Scholars Matrix，KSM）作为聚类对象，即认为具有相似学者集合的关键词之间存在相似性。③各层次类目的设置。聚类是一种无监督的分类方法，类目个数（C_num）以及类目名称（C_name）都需要领域专家在聚类前人为设定。笔者根据 CLIS 的学科特点，设置 C1 层 C_num = 10，C2 层则为 5，其后各层均设定为：

$$C\_num = \min(5, \text{ceil}(m/MaxNum)) \quad (1)$$

其中，m 为需要聚类的簇中的基本知识点数目，MaxNum 则为允许聚类的最小关键词数，函数 ceil（X）表示取大于 X 的最小整数，min（X，Y）表示取参数 X，Y 中的较小者。④多层聚类结束的条件。根据图 4 所示，本文采用了多层聚类方式来实现基本知识点的划分和不同粒度知识点的生成，因此需要事先设定聚类算法结束运行的条件。笔者设置了两个参数来控制聚类过程，MaxNum：允许聚类的最小结点数，即若簇中关键词数大于该值，则继续聚类，否则停止，同公式（1）；SumD：允许聚类的最小簇内距离，即若簇中各结点距离中心结点的总和大于该值，则继续聚类，否则停止。两个条件只要满足其中一个，则聚类过程结束。MaxNum 用于防止学科知

识点粒度过小而导致语义偏差，目的是放大簇间的耦合程度；而 SumD 用于防止关键词结点过于密集而导致错误类目生成或无法聚类现象发生，目的是控制簇的内聚程度。这两个参数的取值变化会直接导致 CLIS 学科知识结构的变动。

为了获得 MaxNum 和 SumD 的取值，笔者令 MaxNum = {5, 10, 15, 20}，SumD = {2, 3, 4, 5, 6}，共进行了 20 次实验。根据各层次的类目设置规则，采用 K-means 算法对 KSM 执行多层聚类操作；鉴于 K-means 算法初始中心选择不同可能导致的结果不稳定性，对每次 K-means 聚类均执行 10 次运算，并选择簇内结点与中心结点距离和最小的运算作为聚类结果。最终聚类结果如表 2 所示。表中自上而下，允许聚类的条件逐步放宽，由此生成的知识层次结构也发生了微妙的变化。①生成的簇（Num_C）明显减少，层次结构的总体宽度（MaxWid_H，即结点最多层次上的结点数）也越来越窄；②很明显，随着簇内允许结点数的不断升高，层次结构的平均最大深度（MaxDep）不断变大，而最小深度（MinDep）也有减少的趋势，其实质就是整个结构被拉长了；③允许聚类的条件变宽松了，簇内结点数最大值（MaxWid_C）和最小值（MinWid_C）均呈现出上升趋势，前者变化甚为明显。一般来说，一个合理的类层次结构中，整体的宽度、深度以及簇的大小均要适中。于是，笔者选择了比较合理的参数为 MaxNum = 15，SumD = 3，即当簇内结点数大于 15 而且簇内结点与中心点距离和大于 3 时，则继续聚类，以确保簇内结点要么聚集在较小空间内，要么少于 15 个。

以选定的参数再进行 10 次实验，并结合领域专家的意见以及层次结构的宽度、深度、类目数等基本特征，最终选择其中最为合理的聚类结果作为 CLIS 的知识体系结构。其深度为 9，宽度为 178，共计 347 个知识集合，第一层的 10 个知识类目如表 3 所

表 2 MaxNum 和 SumD 参数值的不同组合所生成的 CLIS 知识结构

No.	各层簇数	MaxNum	SumD	MaxDep	MinDep	MaxWid_H	MinWid_H	MaxWid_C	MinWid_C	Num_C
1	C1 = 10 C2 = 5 $Cn = \min(5, \text{ceil}(m/\text{MaxNum}))$	5	2	6	3	441	10	20	1	710
2			3	9	3	376	5	24	1	617
3			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

表3 CLIS知识体系结构中第一层知识类目

C1_No.	C1_name	基本知识点数	C1_No.	C1_name	基本知识点数
11	C1_高校图书馆	741	16	C1_传播学	244
12	C1_文献学	195	17	C1_数字图书馆	356
13	C1_公共图书馆	261	18	C1_图书馆	298
14	C1_竞争情报	316	19	C1_知识管理	282
15	C1_搜索引擎	254	20	C1_情报学	134

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

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

图5 具有上下位关系的知识类"Ontology"和"Semantic Web"的OWL编码

示,以簇中出现频率最高的关键词作为其类目名称。从表中可以发现,在CLIS知识体系中,①"高校图书馆"及其相关研究是其中最大的类目,有741个关键词来自该大类,比排名第二的"数字图书馆"高出了一倍以上。可见,该类目是目前CLIS的主要研究内容。②"情报学"和"文献学"研究规模相对较小,相关关键词少于200个,"文献学"作为CLIS的一个传统研究方向似乎出现了没落的趋势,而"情报学"则由于"竞争情报"、"搜索引擎"等方向的兴起,在研究内容上出现了较大的分流。③在10个一级类目,与"图书馆"相关的类目占了4个,可见到2012年为止,图书馆学的研究依然是CLIS知识分布的重点。

3.3 CLIS学科知识结构的存储和展示

采用OWL可以将生成的CLIS学科知识体系以本体形式进行文本存储并实现可视化展示。OWL中用于描述IS_A关系的标签主要有<Owl: Class>和<Owl: subClassOf>,其基本语法有两种方式:

<owl:Class rdf:ID = "Class Name" >Content </owl:Class> (2)

<owl:Class rdf:ID = "Subclass Name" >
    <rdfs: subClassOf rdf: resource = " #Superclass Name"/> (3)
    …
</owl:Class>
<owl:Class rdf:ID = "Subclass Name" >
    <rdfs: subClassOf > <owl: Class rdf: ID = " Superclass Name" >Content </owl:Class > (4)
    </rdfs:subClassOf >
    …
</owl:Class >

公式(2)和公式(3)一起描述了两个类及其父子关系,先定义父类,然后在定义子类的同时指定其父类;而公式(4)则将上述两个过程合并为一个,即在定义子类并指明其父类的同时,定义父类。图5分别采用这两种方式定义了"Ontology"和"Semantic Web"类及其IS_A关系,其中左侧采用了第一种方式,而后侧采用了第二种方式。

类似的,采用上述OWL标签可以将整个CLIS知识体系进行编码,最终形成仅包含层次关系的CLIS知识本体。图6列出了CLIS知识结构的部分OWL编码,主要描述的是一级类目和二级类目之间的上下位关系;图7则是以spring图形的方式显示了CLIS学科中C3_语义网、C3_图书馆事业和C3_文献计量学等三级类目中的所有基本知识点,其中以LIS_KS作为顶层知识点。

4 CLIS学科知识结构的验证

CLIS的知识结构是根据K-means聚类自动生成的,那么这个结果是否具有合理性呢,笔者下面试图通过对一些特殊类目的内聚和耦合分析来进行局部验证。

```
79 <owl:Class rdf:ID="C2_文本分类"><rdfs:subClassOf rdf:resource="#C1_搜索引擎"/></owl:Class>
80 <owl:Class rdf:ID="C2_新媒体"><rdfs:subClassOf rdf:resource="#C1_搜索引擎"/></owl:Class>
81 <owl:Class rdf:ID="C2_信息组织"><rdfs:subClassOf rdf:resource="#C1_搜索引擎"/></owl:Class>
82 <owl:Class rdf:ID="C2_数字化"><rdfs:subClassOf rdf:resource="#C1_图书馆"/></owl:Class>
83 <owl:Class rdf:ID="C2_图书馆"><rdfs:subClassOf rdf:resource="#C1_图书馆"/></owl:Class>
84 <owl:Class rdf:ID="C2_信息服务"><rdfs:subClassOf rdf:resource="#C1_图书馆"/></owl:Class>
85 <owl:Class rdf:ID="C2_信息共享"><rdfs:subClassOf rdf:resource="#C1_图书馆"/></owl:Class>
86 <owl:Class rdf:ID="C2_信息资源"><rdfs:subClassOf rdf:resource="#C1_图书馆"/></owl:Class>
87 <owl:Class rdf:ID="C2_目录学"><rdfs:subClassOf rdf:resource="#C1_文献学"/></owl:Class>
88 <owl:Class rdf:ID="C2_四库全书总目"><rdfs:subClassOf rdf:resource="#C1_文献学"/></owl:Class>
89 <owl:Class rdf:ID="C2_图书馆史"><rdfs:subClassOf rdf:resource="#C1_文献学"/></owl:Class>
90 <owl:Class rdf:ID="C2_文献学"><rdfs:subClassOf rdf:resource="#C1_文献学"/></owl:Class>
91 <owl:Class rdf:ID="C2_阅读"><rdfs:subClassOf rdf:resource="#C1_文献学"/></owl:Class>
92 <owl:Class rdf:ID="C2_电子政务"><rdfs:subClassOf rdf:resource="#C1_知识管理"/></owl:Class>
93 <owl:Class rdf:ID="C2_核心竞争力"><rdfs:subClassOf rdf:resource="#C1_知识管理"/></owl:Class>
94 <owl:Class rdf:ID="C2_以人为本"><rdfs:subClassOf rdf:resource="#C1_知识管理"/></owl:Class>
95 <owl:Class rdf:ID="C2_知识管理"><rdfs:subClassOf rdf:resource="#C1_知识管理"/></owl:Class>
```

图6 CLIS知识结构的部分OWL编码

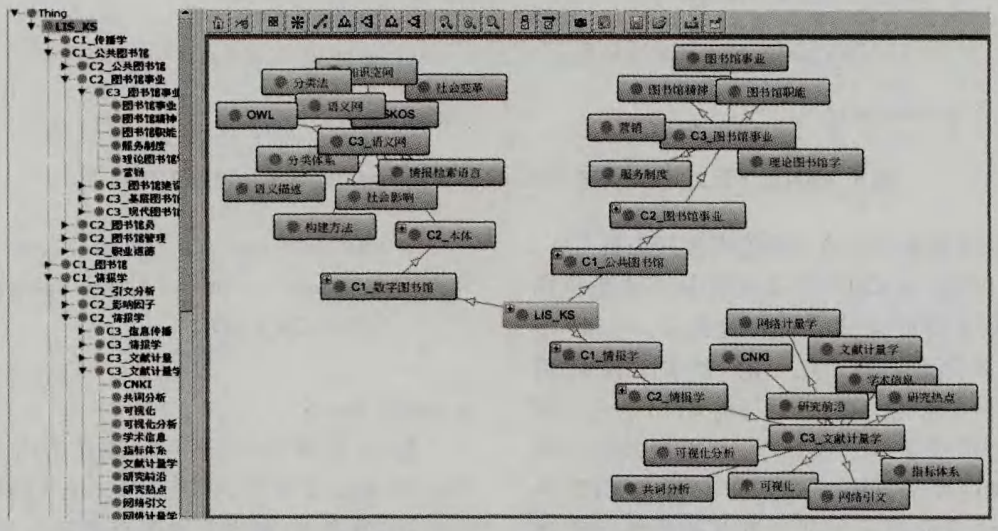

图7 CLIS中C3_语义网、C3_图书馆事业和C3_文献计量学等类目知识点的可视化展示

4.1 基于HC的知识类目内聚性分析

在进行多层聚类时，笔者设置了两个参数SumD=3和MaxNum=15用于控制聚类是否继续。然而在最终生成的CLIS知识结构中，笔者发现最底层簇内对象最多达到了25，远远超过了MaxNum阈值的限制，那该簇聚类结束的原因是簇内对象与簇中心的距离总和(记为Sum_dis)小于SumD阈值；于此相反，在最底层簇中也存在簇内对象数小于MaxNum，但Sum_dis远远大于SumD的知识类。笔者将前者记为A，后者记为B，那么可以进行合理假设：对于A，其对象必然聚集在其中心点附近，内聚性较好，簇内对象之间的差异并不明显；而B内对象则相对中心点较为分散，直接导致了其距离总和较大，内聚性相对较差。

对于上述两种较为极端的知识类，笔者各选择了2个案例，并采用HC方法对类目中关键词的分布情况和内聚特征进行细化分析，从而实现对假设的验证。

(1)A：簇内对象最多的知识类目分析

A1：LIS_KS > C1_公共图书馆 > C2_图书馆管理 > C3_读者服务 > C4_图书馆建筑 > C5_图书馆建筑 > C6_图书馆建筑，Sum_dis=2.787157

A2：LIS_KS > C1_情报学 > C2_情报学 > C3_文献计量 > C4_文献计量 > C5_知识交流，Sum_dis=2.833931

分别对A1和A2中25个关键词所构成的KSM进行HC，结果如图8中(A1)和(A2)所示。①在A1中，有11个对象完全相似，占了总量的一半左右；簇内对象之间的最大差异MaxDiff略大于0.7。②A2的中知识点的分布情况与A1非常相似，也有一组共11个对象之间的距离为0；簇内对象间的最大距离也略大于0.7。③由此可知，A1和A2中知识点间的最大差异都很小，甚至存在大量结点之间无差异，两者的聚类层次性都不明显，这说明对簇内知识点再进行分类的意义并不大，簇内知识点已经保持了较大的内聚性。

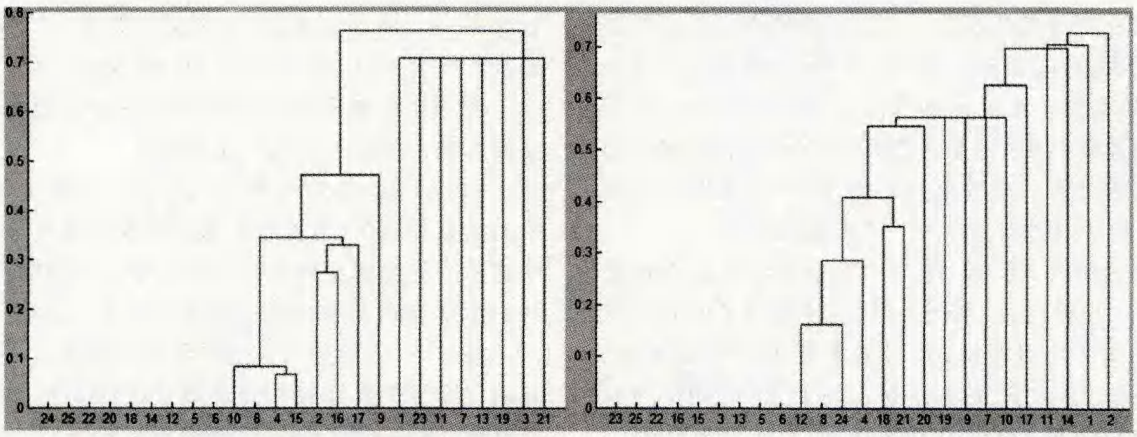

图 8　含有最多基本知识点的知识类目的 HC 结果

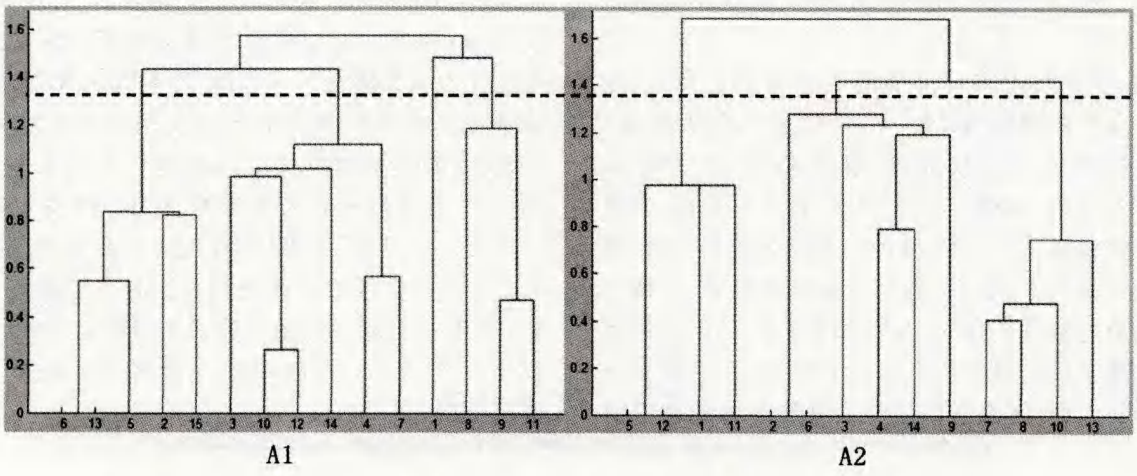

图 9　具有最大 Sum_dis 的知识类目的 HC 结果

(2)B：Sum_dis 最大的知识类目分析

B1：LIS_KS > C1_图书馆 > C2_信息共享 > C3_知识整合，Sum_dis = 8.502761

B2：LIS_KS > C1_文献学 > C2_四库全书总目 > C3_非物质文化遗产，Sum_dis = 7.464072

对 B1 和 B2 中的关键词进行 HCA，结果如图 9 中(B1)和(B2)所示。①在 B1 中，知识点间的 MaxDiff 接近 1.6，几乎是 A1 和 A2 的一倍左右，可见相对于后者，B1 中的知识点被分布在一个较大的空间范围内；15 个知识点被明显的分成了 3 组，可见聚类层次性非常清晰。②B2 内知识点的分布情况类似于 B1，其中 5 和 12 以及 2 和 6 完全相似；14 个对象也被明显的分成 3 个类；MaxDiff 超过了 1.6，簇内对象之间的离散度比 B1 更大。

基于对 4 个底层类目的分析，笔者认为本节的假设基本成立：A 类目虽然含有较多的基本知识点，但由于内部对象相似度较大，相互聚集在一个较小的空间范围内，内聚性较强，不具备继续分类的可能；而 B 类目中基本知识点相互之间离散度较大，并且对象的层次性较为清晰，可见簇内对象内聚性相对较差，但由于对象数量少，可以根据实际应用的需要采用更精确的小规模聚类算法如 HC 等对类目进行细化。综上所述，通过对特殊类目内部分布的微观验证，基于 K-means 聚类获得的 CLIS 知识类目或内聚性强或内含知识点较少，其结构具有一定的合理性。

4.2　基于 MDSA 的知识类目耦合性分析

通过 HC，笔者了解了簇内对象之间的相对位置及其内聚程度。那么使用 K-means 聚类获得的 CLIS 知识类目之间是否是相对独立的，即知识类目的耦合性又如何？为此，笔者采用了 DMSA 方法，将不同知识类目中的结点散布于二维平面上，进而根据结点的平面位置分布以及类目之间的耦合性来

验证类目划分的合理性。本次实验依然以具有最多对象的类目A1、A2和具有最大Sum_dis的类目B1、B2作为案例。在具体操作之前，笔者假设：4个类目的对象彼此独立，类目之间的耦合性较低，聚类具有较好的类目划分效果；A1和A2中对象层次性较差，而B1和B2则存在进一步聚类的可能。

笔者对由A1、A2、B1和B2中基本知识点构成的KSM(7976)执行MDSA操作，计算出了所有关键词间的两两相对距离，并根据距离进行降维压缩。最终获得：①信度Stress = 0.10893(大于10%为一般可信)，效度RSQ = 0.98240(大于0.6为有效)，可见降维效果较好，但可信度一般；②从76维(关键词对象的描述特征)降至2维后关键词对象的位置坐标，据此可绘制出基本知识点的相对位置图如图10所示。

(1)从总体上来看，4类对象被分成了3组，处于第二象限的圆圈为A1，小点A2处于第四象限，而B1(黑色叉)和B2(菱形)则没有被完全分开，全部堆积在了第一象限，但是B1对象多处于B2的间隙中。这说明A1和A2表现出了较强的独立性，相互之间以及与B1和B2都存在较大的差异，总体上看各类目的耦合性较低；由于降维操作，使得对象间关联出现了一定程度的失真，B1和B2在二维平面上合成了1类，区别它们的特征可能在降维过程中丢失了；B1和B2之间的关联较之与A1、A2更为紧密，B1来自“C1_图书馆”，而B2来自“C1_文献学”，图书馆向来都是文献资源的主要收藏地，两者之间在某个维度上存在一定关联。

(2)A1和A2中均有25个对象，但图中显示出来的结点却较少，可见有大量的对象被相互覆盖，说明这两个集合的对象相对比较集中，内聚性较高；当然也不排除有个别对象远离集合中心，表现出了较大的差异性，可能是与其他对象相似度较低，也可能是由于降维失真导致差异性被放大；A1和A2中相异对象之间的相互位置呈现线性增长趋势，对象间层次性不明显，即进一步对其进行划分存在较大困难，也可以说当前分类基本上已经达到最优。

(3)B1和B2中的对象也较为集中，说明基于K-means聚类的知识类目划分达到了一定效果，相似对象确被聚集在了一起；但是不同于A1和A2中对象的线性分布，B中对象则表现出了较为明显的层次性，即其中对象又可以被划分为若干个团体，B1很明显可被分成3个团体，而B2则被分成了4个子类。可见B1和B2均存在进一步聚类的空间，可根据应用需要进一步细化。但是需要注意的是，MDSA是一种通过降维压缩来平面展示对象之间差异的方法，存在一定的失真，不能将其作为进一步分类的依据。

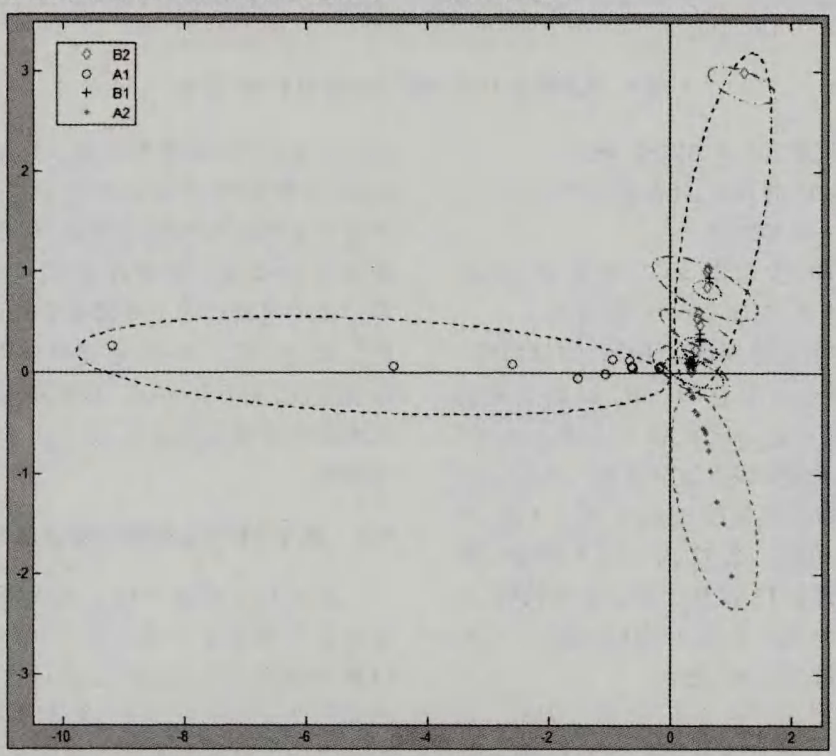

图10 CLIS中4个知识类目的MDSA结果

基于上述分析，可见4个类目从总体上来看彼此独立，反映出CLIS知识类目间的耦合性较差，CLIS知识结构具有较强的合理性，但是B类目相互间的区分度相对较差；A1和A2中对象集中，内聚性强而层次性差，B1和B2正好相反，存在进一步聚类的可能，这与上节的结论基本一致。可见本节之前的假设成立，CLIS知识结构的耦合性较低。

综上所述，基于HC和MDS分析可知，基于本文方法获得的CLIS知识类目具有高内聚性低耦合的特点，说明CLIS知识结构具有较强的可信度和合理性。

5 CLIS学科知识结构的演化分析

本文构建的CLIS知识结构建立在CSSCI（2003～2012年）基础上。那么在这10年间，CLIS的知识结构是如何调整的？本节试图通过CLIS知识结构在学科发展演化分析中的具体应用进一步论证其有效性。

笔者仅从知识点的频次变化来探测学科研究的发展[36]。①知识点的频次变化可以从两个角度衡量，一是统计关键词所属知识类目的年度频次变化；二则是以文献为单位，统计其所属知识类目的年度频次状况。关键词和文献数量并不一致，所得结果将会形成偏差，本文以前者作为考察角度探测知识点年度发展轨迹。②采用的方法是将关键词映射到各级知识类目中，进而统计关键词所属知识类目的年度频次，描绘出其10年间的变化轨迹以分析CLIS知识结构的动态演化。

经过计算，本文筛选出的3081个关键词分布在50 493篇CSSCI论文中，而CLIS一级类目的年度分布如图11所示，10个一级类目的发展轨迹大体呈现出3种趋势，①研究规模最大的3个类“C1_高校图书馆”、“C1_图书馆”和“C1_数字图书馆”在经历了快速发展（03－05），一段时间的稳定（05－09）后，出现了下降态势（09－12），说明这些类目的研究基本上已经过了研究高峰；②规模次之的3个类“C1_公共图书馆”、“C1_竞争情报”和“C1_知识管理”在10年间则总体处于稳定的发展状态，但是增长的幅度不大，而且在2010年之后均出现了下调趋势，其发展势头开始明显减弱；③研究规模较小的其他4个类则处于明显的上升趋势，特别是“C1_情报学”、“C1_搜索引擎”和“C1_传播学”类，增长幅度较大，发展势头良好，属于CLIS的新兴研究热点，特别是后两个交叉研究方向的快速发展说明CLIS与其他学科的知识交流趋于频繁，“C1_文献学”作为传统研究方向稳中带升。

那么，CLIS一级类目的年度变化又是如何造成的呢？可以借助CLIS知识结构进一步探索更小类目甚至基本知识点的演化规律。但鉴于篇幅，笔者仅就“C1_数字图书馆”和“C1_情报学”进行了下钻分析，其二级类目的年度发展轨迹如图12所示。①明显的，引起“C1_数字图书馆”发展趋势下降的主要原因是“C2_数字图书馆”和“C2_信息检索”研究总量的收缩，而且两者均是在2005年开始出现了迅速下滑，特别是后者，很可能是在2005年前后出现了从“纸质信息检索”到“网络信息检索”的内容调整，导致了其研究规模的震动；此外除“C2_网络信息资源”发展较为稳定外，“C2_本体”和“C2_图书馆学”实际上还是处于稳中上升的趋势，尽管幅度不大，说明“C1_数字图书馆”中的一些研究方向的热度并没有完全减退。②在“C1_情报学”中，除了“C2_科学评价”的轨迹比较稳定外，其他4个类的研究均处于上升趋势，直接导致了该大类的迅速发展；除了最大的“C2_情报学”之外，其他4个二级类目实际上存在很大的交叉，均可归纳为“科学评价与引文分析”研究方向，由此可见该方向目前已经成为情报学领域的研究重点，而且研究热度正处于逐年上升趋势。

6 结 语

本文以本体思想作为理论指导，将狭义的学科知识结构理解为学科知识点的层次体系，进而从CSSCI期刊论文集合中衍生出CLIS的学科知识结构，并对其在2003～2010年10年间的发展轨迹进行了初步探测。论文构建的CLIS知识体系为后期该学科广义知识结构的研究奠定了知识基础。

本文虽然通过对知识类目的微观检测和学科发展趋势的宏观分析对CLIS_KS进行了合理性和有效性的验证，但在过程中也发现一些问题有待于今后进一步探讨。①采用的K-means聚类适合文中大规模数据处理的特点，但是其聚类结果的不稳定性也对CLIS知识类目生成造成了一定的影响；②以类目中频次最高的关键词作为类目名称，导致存在大小类同名的现象，这种处理方法值得商榷；③鉴于篇幅限制，CLIS知识结构的应用在文中涉及较少，知识类目演化分析仅采用了传统的频次统计方法，有待于进一步改进和完善。

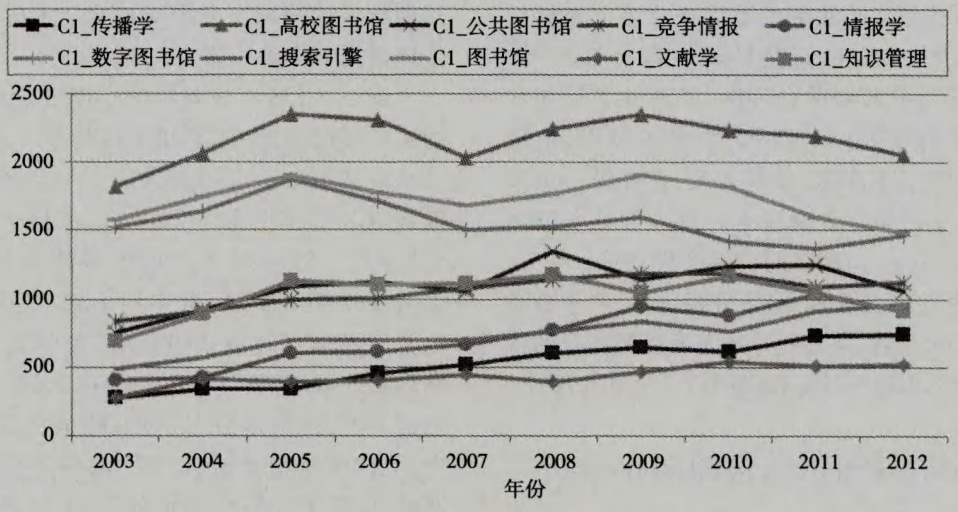

图 11 CLIS 一级类目的年度分布

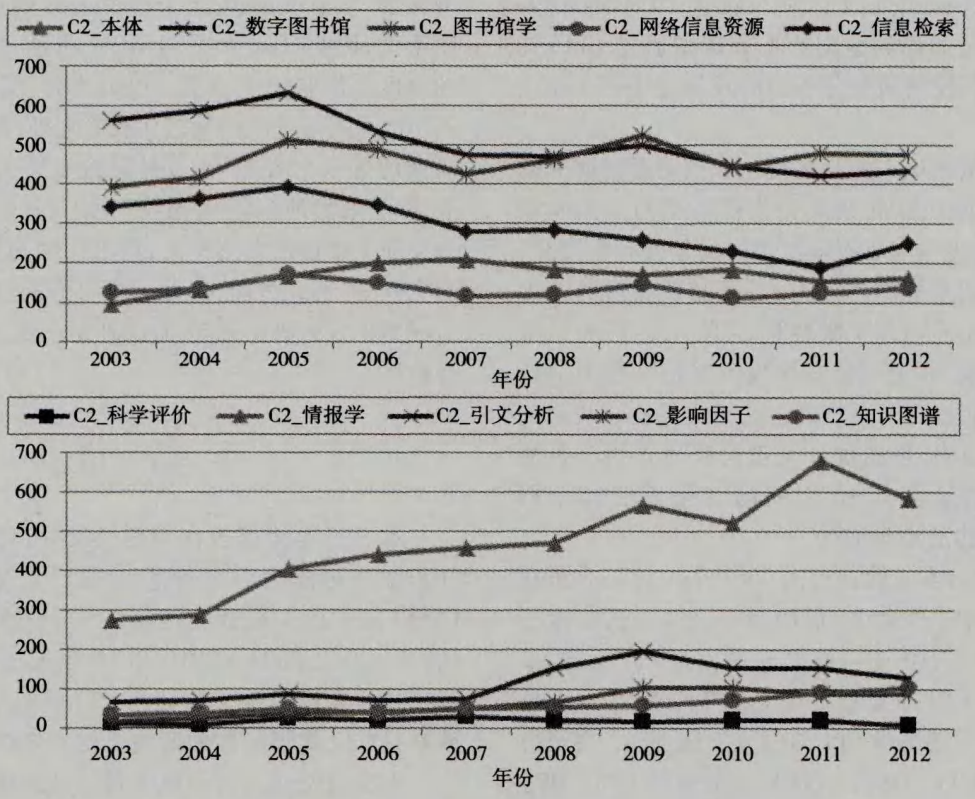

图 12 “C1_数字图书馆”和“C1_情报学”二级类目的年度发展轨迹

参 考 文 献

[1] Song M, Kim S Y. Detecting the knowledge structure of bioinformatics by mining full-text collections [J]. Scientometrics, 2013, 96(1): 183-201.

[2] Chang Y W. Tracking scientometric research in Taiwan using bibliometric and content analysis [J]. Journal of Library and Information Studies, 2012, 10(2): 1-20.

[3] Powers J H. On the intellectual structure of the Human-Communication Discipline[J]. Communication Education, 1995, 44(3): 191-222.

[4] 苏新宁,邹志仁. 中国人文社会科学学术影响力报告(2011版)[M]. 北京:高等教育出版社, 2011.12.

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

[6] Galvagno M. The intellectual structure of the anti-consumption

and consumer resistance field: An author co-citation analysis[J]. European Journal of Marketing, 2011, 45 (11-12): 1688-1701.

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

[8] Torres-Salinas D, Moed H F. Library Catalog Analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics [J]. Journal of Informetrics, 2009, 3(1): 9-26.

[9] Ma R M. Discovering and analyzing the intellectual structure and its evolution of LIS in China, 1998-2007 [J]. Scientometrics, 2012, 93(3): 645-659.

[10] Prebor G. Analysis of the interdisciplinary nature of library and information science [J]. Journal of Librarianship and Information Science, 2010, 42(4): 256-267.

[11] Kurihara T, Tomari N, Aratani T. Trend of EASTS research in the past 20 years[C]. In: Proceedings of the Eastern Asia Society for Transportation Studies, 2013, 9.

[12] Jeong S, Kim H G. Intellectual structure of biomedical informatics reflected in scholarly events [J]. Scientometrics, 2010, 85(2): 541-551.

[13] Hu C P, Hu J M, Deng Sh L. A co-word analysis of library and information science in China [J]. Scientometrics, 2013, 97(2): 369-382.

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

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

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

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

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

[19] Calabretta G, Durisin B, Ogliengo M. Uncovering the intellectual structure of research in Business Ethics: A journey through the history, the classics, and the pillars of journal of Business Ethics[J]. Journal of Business Ethics, 2011, 104(4): 499-524.

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

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

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

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

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

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

[26] Samiee S, Chabowski B R. Knowledge structure in international marketing: a multi-method bibliometric analysis[J]. Journal of the Academy Of Marketing Science, 2012, 40(2): 364-386.

[27] Persson O, Danell R, Wiborg Schneider J. How to use Bibexcel for various types of bibliometric analysis [G]. // F. Åström, R. Danell, B. Larsen, J. Schneider (eds.). International Society for Scientometrics and Informetrics, Leuven, Belgium, 2009: 9-24.

[28] SPSS[OL]. [2014-07-14]. http://www-01.ibm.com/software/analytics/spss/

[29] SAS[OL]. [2014-07-14]. http://www.sas.com/en_us/home.html

[30] Borgatti S P, Everett M G, Freeman L C. Ucinet for Windows: Software for Social Network Analysis[C]. Analytic Technologies, Harvard, MA, 2002.

[31] de Nooy W, Mrvar A, Batagelj V. Exploratory Social Network Analysis with Pajek[M]. UK: Cambridge University Press, 2005.01.

[32] Meyer M, Zaggl M A, Carley K M. Measuring CMOT's

intellectual structure and its development [J]. Computational and Mathematical Organization Theory, 2011, 17(1): 1-34.

[33] Chen C M, Paul R J. Visualizing a knowledge domain's intellectual structure[J]. Computer, 2001, 34(3): 65-71.

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

[35] 马瑞敏，倪超群. 基于作者同被引分析的我国图书情报学知识结构及其演变研究[J]. 中国图书馆学报，2011(6)：17-26.

[36] 吕红，邱均平. 国际图书情报学二十年研究热点变化与研究前沿分析[J]. 图书馆杂志，2013，32(9)：14-20.

（责任编辑 马 兰）