基于 CSSCI 本体的学术期刊关联分析*

邓三鸿　王　昊　苏新宁
（南京大学信息管理系　南京 210093）

【摘要】试图改变学术资源评价的传统分析模式，提出先建立完整的知识库，继而在知识库基础上获得分析结论的“知识驱动型”分析思路。为此，采用本体机制来实现 CSSCI 学术资源的知识组织，以面向对象的方式来描述期刊及其相关知识，建立基于本体的 CSSCI 学术资源网络模型（CSSCI_Onto）；通过本体中揭示的学术资源知识，不仅可以综合地了解期刊之间的两两关系及其关联程度，而且能够挖掘期刊之间隐含的双向和多元关联模式，发现一定时期内学科中的核心期刊集，为学科内期刊之间的合作和分工提供可参考的事实依据，促进期刊研究内容呈现合理分布。重点探讨 CSSCI 来源期刊之间关联的构建及分析，是基于 CSSCI 知识本体实现学术资源关联分析的一个组成部分。

【关键词】CSSCI 本体　学术期刊　关联分析　学术资源网络模型　学术评价
【分类号】G250

Association Analysis of Academic Periodicals Based on CSSCI_Onto

Deng Sanhong　Wang Hao　Su Xinning
（Department of Information Management, Nanjing University, Nanjing 210093, China）

【Abstract】The paper tries to change the traditional analysis mode of academic resources evaluation, and puts forward the “knowledge - driven” analysis idea which establishes a complete knowledge base firstly, then obtains conclusions on the basis of this knowledge base. For this, the paper adopts the Ontology mechanism to achieve the knowledge organization for CSSCI academic resources, which describes the periodicals and related knowledge by object - oriented approach, so that to establish CSSCI Academic Resource Networks Model (CSSCI_Onto). On this basis of academic knowledge annotated in the Ontology, not only the relationship and association degree between periodicals can be learned comprehensively, but also the bidirectional and multiple association patterns that are implied in original knowledge can be mined, and further the set of core periodicals in the disciplinary during certain period of time can be discovered, which provides factual basis for cooperation and division between periodicals, so that to promote a reasonable distribution of research content of periodicals. The paper focuses on the construction and analysis for relationship of periodicals, which is a part of academic resources association analysis based on CSSCI_Onto.

【Keywords】CSSCI Ontology　Academic periodicals　Association analysis　Academic Resources Networks Model　Academic evaluation

1　引　言

中国社会科学引文索引（Chinese Social Science Citation Index, CSSCI）自诞生以来，每年从全国所有人文社科

收稿日期：2011 - 01 - 30
收修改稿日期：2011 - 02 - 26
*本文系国家社会科学基金项目“面向语义网本体的知识管理研究”（项目编号：09CTQ010）的研究成果之一。

类期刊中精选出400－500种出版质量较好、学术水平较高、具有一定学科影响的学术期刊作为来源刊，收录其刊载论文及其引文信息[1]。各来源刊之间由于主题共现、相互引用、同被引等存在内容关联。随着期刊的不断成熟，这种关联也越来越明显地展示出期刊的发展方向及期刊间的相互作用，甚至形成一定时期内学科的核心期刊集。因此，对人文社科学术期刊关联进行分析，能够了解期刊之间的交叉关系及其交融程度，发现学术期刊的发展规律，促使学科内期刊的研究内容呈现合理分布。既可以促进期刊之间的相互合作，为更大规模、更高质量精品期刊的出现奠定理论基础；也可以使期刊了解自身在学科中的研究内容定位，为其研究方向的进一步确定和发展提供事实依据。

学术期刊关联分析是科学评价和知识服务的重要组成部分，其最终目的是为了促使期刊这一重要学术载体的良性发展，创造更好的学术环境。因此，期刊关联分析已经开始引起了学术界的关注，但是更多的还是聚焦在基于期刊被引进行期刊评价[2]，也有学者尝试分别从期刊引用和主题共现探讨人文社科类期刊间的内容关联并建立关系云图[3-5]。文献[4]虽然综合上述两个角度来探讨期刊关联，但是在数据源选择上具有较大局限，关联标准的权重选择具有随意性。造成分析局限的主要原因在于这些研究是基于传统的CSSCI数据结构来实现，都是为了单一目的而聚集数据，在数据的处理上存在不一致性和不规范性。为此，本文试图将具有语义描述能力的本体（Ontology）机制引入到CSSCI的知识组织中，建立基于本体的CSSCI学术资源知识地图（简称CSSCI_Onto）[6]，以面向对象的方式来组织期刊及其相关概念，以统一和规范期刊与其他学术资源间的关联模式，实现期刊与其他学术资源间丰富语义关联的深度揭示。在此基础上，通过知识挖掘发现隐含在原有知识下用户感兴趣的期刊关联模式，探讨可提供具体决策支持的分析结论，以增强期刊间合作，促进期刊综合性和专业性的分流，实现学术期刊的良性发展。

2 构建人文社科期刊知识本体

当前CSSCI主要关注来源文献和被引文献，在信息组织上以文献作为主要考察对象，对其他学术资源之间内在关联的描述明显不足。因此，笔者从CSSCI（2000－2006）共7年的数据中提取知识元及其关系，构建了CSSCI知识地图（即CSSCI_Onto）[7]，借助本体机制以面向对象的方式来统一、明确、规范地确立学术资源之间的语义关联。本文则试图以该知识库为基础，揭示人文社科类学术期刊之间的显性和隐性关联。

2.1 学术期刊知识本体概念模型的构建

CSSCI_Onto由概念库、实例库和规则库等构成，其中概念库是对CSSCI学术资源的抽象描述，包括了3层概念层次结构，共39个本体类，336个属性[8]。期刊是顶层主要概念，含有来源期刊和被引期刊两个子概念，而来源期刊又可以分为来源单刊（以“期”为单位）和来源种刊（以“种”为单位），本文重点讨论种刊之间的关联，下文提到的期刊均指种刊。除了分类关系之外，期刊与其他学术资源之间以属性的形式建立非分类语义关系。

期刊概念的属性集合如表1所示：

表1 来源期刊概念的属性分类

属性分类	举例
数值属性	期刊代码、期刊名称、年册数、创刊年代、主办单位、出版地区、出版单位、ISSN、通讯地址、备注
同类对象属性	关联期刊
异类对象属性	所属学科、关联主题、来源文献、来源单刊、来源作者、基金类别、基金项目、载文类型、作者部门、机构、机构类型、省、市、关联引文、引文语种、引文类型、引用期刊、引用出版社、引文作者、引文年代、年度指标统计

这些属性被分为：

（1）数值属性（Datatype Property），用于描述期刊自身的性质和状态，例如“期刊代码”、“期刊名称”、“创刊年代”等；

（2）对象属性（Object Property），以某一概念的实例作为属性值，描述的是概念实例之间的关系。对象属性又可以分为同类对象属性和异类对象属性，前者以同类对象作为概念属性值，描述同类概念实例之间的关系，属性值主要来自与中心实例并列的期刊元组[9]，如“关联期刊”，这是期刊关联分析的语义基础；后者以其他类型对象作为属性值，例如“来源文献”、“关联主题”等，揭示的是期刊概念和文献、主题等概念之间的关系，借此可以从微观角度考察期刊间的关联。

2.2 CSSCI学术期刊的语义标注

概念模型组成了CSSCI_Onto的概念库，定义了本体的“元结构”；但CSSCI中具有核心作用的是各类学术资源的实例及其关系，将CSSCI学术知识以网络形

式关联在一起以提供完善的知识服务是构建 CSSCI_Onto 的主要目的。为此,需要根据“元结构”对 CSSCI 数据进行语义标注,即根据概念模型中定义的概念和概念属性模板,抽取实例并设置实例属性值。本文仅探讨期刊实例的语义标注过程和结果。

在 CSSCI(2000 - 2006)中,一共存在 558 种人文社科类学术期刊。根据期刊概念属性的类型,可以采用不同的方式设置这些实例的属性值。

(1)数值属性主要来自关系的字段,可以根据不同字段“同一元组”这一依赖关系获得期刊实例的数值属性值,例如根据“期刊名称”获得与其“同一元组”的“期刊代码”、“ISSN”等属性值;

(2)期刊概念的异类对象属性大部分是多值属性,需要统计属性实例相对于中心实例的关联次数以区分关联的强度,为了提高关联强度的合理性,甚至可以基于 TF - IDF 算法来计算属性实例的影响程度,例如对于“关联主题”,可以统计期刊实例各个“关联主题”的出现次数,计算关联度;

(3)对于同类概念属性,则可采用标准加权方式设置属性值。

期刊与期刊之间由于内容交叉而存在关联,内容交叉则主要表现在主题共现(期刊间的关联在一定程度上取决于其关联主题的交叉程度)、引用(引用率越高表明期刊关联度越大)、同被引(频繁的同被引能够在一定程度反映出期刊内容的交叉度)和同学科(相同学科内的期刊之间具有一定的内容相关性)。因此,本文选择上述 4 个标准作为期刊关联依据,分别计算期刊实例间的关联,结果如表 2 至表 5 所示:

表 2　基于主题共现的图书情报档案领域期刊关联情况示例(局部 top5)

中心期刊	关联期刊	关联度	中心期刊	关联期刊	关联度	中心期刊	关联期刊	关联度
…	…	…	…	…	…	…	…	…
情报学报	情报杂志	0.18347	情报理论与实践	情报杂志	0.22078	现代图书情报技术	情报杂志	0.17431
	情报科学	0.18282		情报科学	0.21939		情报科学	0.1589
	情报理论与实践	0.16481		图书情报工作	0.20046		图书情报工作	0.14196
	图书情报工作	0.15953		情报资料工作	0.15099		情报理论与实践	0.11479
	现代图书情报技术	0.14447		图书馆论坛	0.13749		图书馆论坛	0.11152
中国图书馆学报	图书情报工作	0.19963	情报资料工作	图书情报工作	0.22435	图书情报工作	情报杂志	0.17266
	图书馆论坛	0.18542		情报杂志	0.21247		情报科学	0.16194
	情报杂志	0.18186		情报科学	0.2113		图书馆论坛	0.14321
	图书馆理论与实践	0.17826		图书馆论坛	0.19822		图书馆理论与实践	0.12398
	情报科学	0.17541		图书馆理论与实践	0.18914		图书馆杂志	0.12058
…	…	…	…	…	…	…	…	…

(注:取主题共现阈值 I = 50(50 个以上主题共现才能保证期刊间具有关联),共获得 555 种期刊之间 141 602 个关联对)

表 3　基于期刊引用的图书情报档案领域期刊关联情况示例(局部 top5)

中心期刊	关联期刊	关联度	中心期刊	关联期刊	关联度	中心期刊	关联期刊	关联度
…	…	…	…	…	…	…	…	…
情报学报	情报理论与实践	0.06003	情报理论与实践	情报学报	0.05546	现代图书情报技术	大学图书馆学报	0.06166
	图书情报工作	0.04775		图书情报工作	0.05052		情报学报	0.06034
	现代图书情报技术	0.04412		情报科学	0.03649		图书情报工作	0.04904
	中国图书馆学报	0.04027		中国图书馆学报	0.03614		中国图书馆学报	0.04380
	情报科学	0.03292		现代图书情报技术	0.03170		情报理论与实践	0.03061
中国图书馆学报	图书情报工作	0.07077	情报资料工作	中国图书馆学报	0.05895	图书情报工作	中国图书馆学报	0.06291
	图书馆	0.04120		图书情报工作	0.05528		大学图书馆学报	0.04011
	图书馆杂志	0.03684		大学图书馆学报	0.03406		情报学报	0.03268
	大学图书馆学报	0.03431		情报理论与实践	0.03195		情报理论与实践	0.03071
	情报学报	0.03112		情报学报	0.02612		现代图书情报技术	0.02614
…	…	…	…	…	…	…	…	…

(注:取期刊互引阈值 Y = 3(期刊间引用达到 3 次,相互间则具有关联),得到 557 种期刊间共 24 461 个关联对)

表 4　基于期刊同被引的图书情报档案领域期刊关联情况示例(局部 top5)

中心期刊	关联期刊	关联度	中心期刊	关联期刊	关联度	中心期刊	关联期刊	关联度
…	…	…	…	…	…	…	…	…
情报学报	大学图书馆学报	0.24879	情报理论与实践	现代图书情报技术	0.28975	现代图书情报技术	大学图书馆学报	0.29117
	图书情报工作	0.24863		大学图书馆学报	0.28212		图书情报工作	0.24673
	现代图书情报技术	0.2427		图书情报工作	0.26015		图书馆杂志	0.24023
	中国图书馆学报	0.2268		中国图书馆学报	0.24307		中国图书馆学报	0.23603
	情报理论与实践	0.221		情报学报	0.23401		图书情报知识	0.23472
中国图书馆学报	图书情报工作	0.23471	情报资料工作	现代图书情报技术	0.31737	图书情报工作	中国图书馆学报	0.24143
	大学图书馆学报	0.20744		大学图书馆学报	0.30672		大学图书馆学报	0.22045
	图书馆杂志	0.15907		图书情报知识	0.29816		图书馆杂志	0.16929
	现代图书情报技术	0.14472		图书馆杂志	0.29008		现代图书情报技术	0.15559
	图书馆论坛	0.13982		图书情报工作	0.27131		图书馆	0.14299
…	…	…	…	…	…	…	…	…

(注:取同被引阈值 T = 30(期刊同时被 30 种以上刊物引用时,相互间具有关联),共得 471 种期刊间 117 264 个关联对)

表 5　CSSCI 各学科内期刊间关联系数

学科名称	来源刊量	刊间关联度	学科名称	来源刊量	刊间关联度	学科名称	来源刊量	刊间关联度
630 管理学	31	0.11014	780 考古学	7	0.11854	890 体育	9	0.11621
710 马列毛泽东思想	14	0.11332	790 经济	85	0.10781	910 统计学	4	0.12789
720 哲学	12	0.11420	810 政治学	45	0.10913	920 心理学	8	0.11722
730 宗教	6	0.12038	820 法学	23	0.11113	930 综合性社会科学	48	0.10898
740 语言学	23	0.11113	840 社会学	9	0.11621			
751 外国文学	6	0.12038	850 民族学	16	0.11265	960 人文、经济地理	10	0.11540
752 中国文学	16	0.11265	860 新闻出版广播	17	0.11236			
760 艺术	20	0.11167	870 图书情报档案	19	0.11188	970 环境科学	10	0.11540
770 历史学	27	0.11058	880 教育学	39	0.10949	980 高校社科学报	54	0.10871

(注:根据“多期刊学科期刊间的关系相对于少期刊学科期刊间的关系弱”思想,同学科期刊间关联度设为 $0.1\times(1+1/\log_2(n\times(n-1)))$,其中 n 为学科中期刊数量,n > 1)

根据不同标准权重各异(分别为 0.4,0.25,0.3,0.05)的思想计算加权平均值,以获得期刊实例间的综合关联度,每一种期刊取综合关联度≥0.1 且综合关联度最高的 20 种期刊作为其关联期刊,由此可获得 558 种人文社科类期刊之间共 9 627 对关联。其中,图书情报档案、管理和经济等 3 学科 12 种期刊之间的关联云图,如图 1 所示:

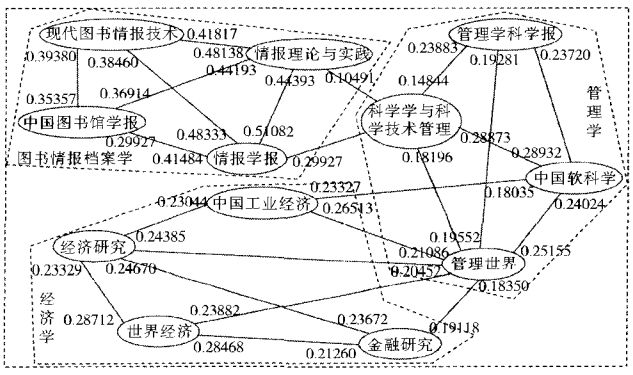

图 1　图书情报档案、管理和经济 3 学科 12 种学术期刊间关联云图

3　基于 CSSCI_Onto 的期刊关联分析

经过概念模型的构建及其对 CSSCI 数据的语义标注后,与期刊相关的所有知识以面向对象的方式被组织在 CSSCI 知识库中,成为 CSSCI_Onto 的重要组成部分。基于这些知识,笔者对期刊关联进行了系统分析,得到了一些准确、合理且具有一定实用价值的结论。

3.1　基于中心期刊的关联分析

从“来源期刊”概念的“关联期刊”属性值中可以获得 CSSCI 所有来源期刊的关联期刊及其关联度。本文以图书情报档案类的重要期刊《情报学报》为例进行分析。

(1)同一学科的期刊在讨论的内容上存在较大的相似性,直接导致其关联期刊多为同一学科领域内的期刊,特别是与期刊保持高关联度的期刊,例如《情报学报》的 20 种关联期刊中,图情档领域的 18 种期刊占据了其前 18 位,如图 2 所示。

(2)由于研究主题的相似(包括主题共现、引用和同被引等),有时也会导致不同学科的期刊之间发生关

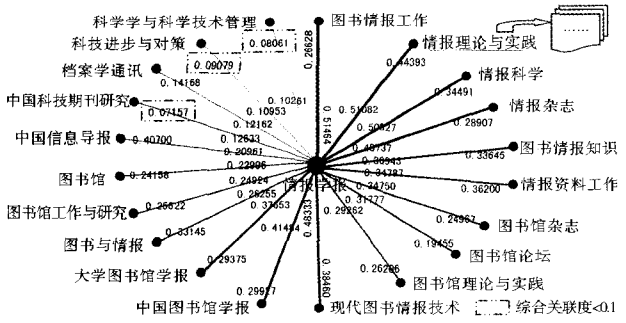

图2 “情报学报”的关联期刊地图

联，例如管理学领域的《科技进步与对策》和《科学学和科学管理技术》由于在探讨内容上与《情报学报》存在更多的交叉性，因此这两种期刊取代《档案学研究》成为《情报学报》的另外两种关联期刊，由此也可认为《情报学报》刊载文章甚少涉及档案学的内容，仅与《档案学通讯》保持微小的关联性。

（3）与《情报学报》关联的期刊首先是情报学类和图书情报综合类期刊，其次是图书馆学类，最后是档案学类和其他相关学科的期刊，其中最依赖的期刊是《图书情报工作》、《情报理论与实践》和《情报科学》，这3种期刊对图书情报学的理论和实践都比较重视，具有一定的综合性，导致它们对《情报学报》学术研究的影响较大，而《现代图书情报技术》、《情报杂志》相对侧重于情报技术和应用研究，对《情报学报》的影响力就相对居后。

3.2 基于期刊间双向关联的综合分析

将CSSCI_Onto中所有期刊实例的关联期刊及其关联度列出，设置一定的约束条件过滤掉期刊间的弱关联，可以获得指定领域中关系最密切的期刊关联，并据此在二维平面中做出这些期刊的关联地图。图情档领域20种学术期刊在综合关联度≥0.4且双方均存在关联时的期刊关联云图，如图3所示：

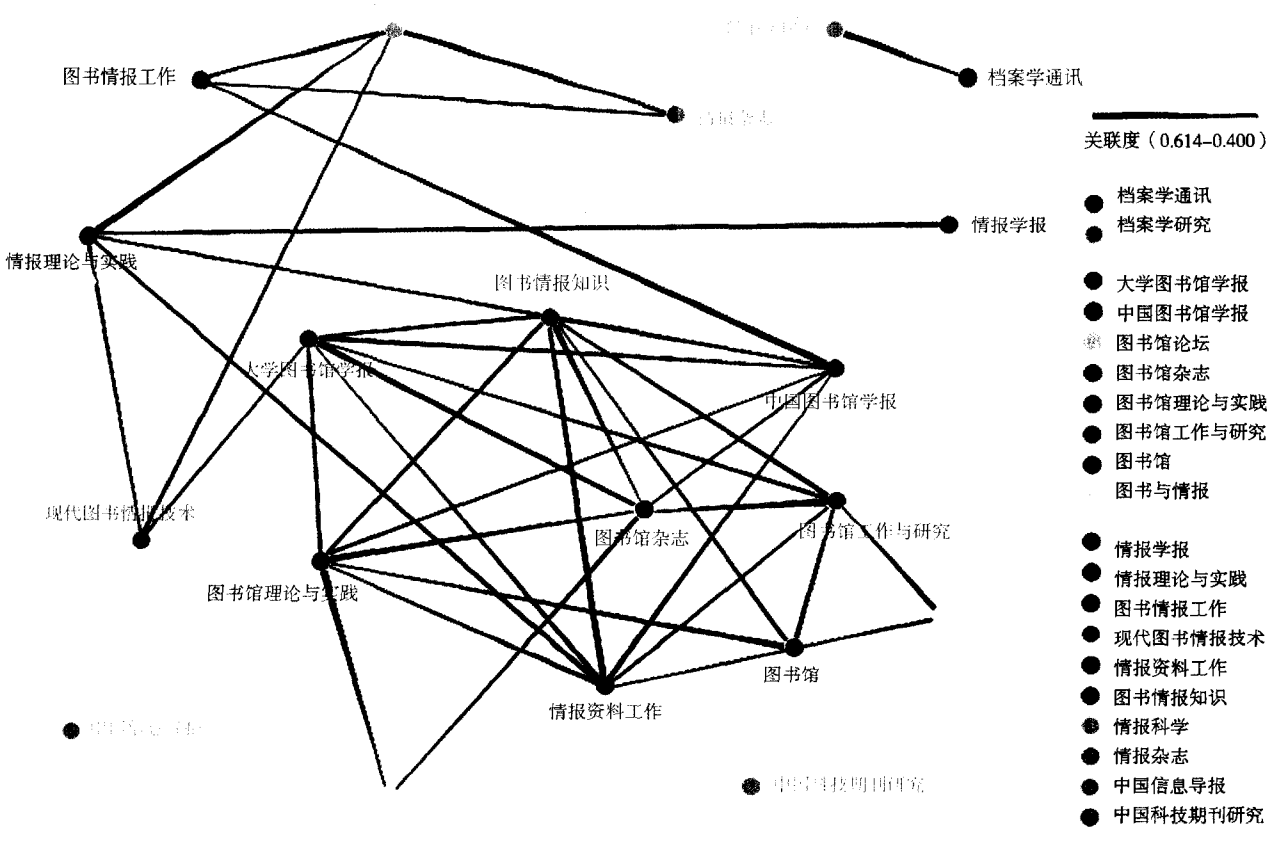

图3 图书情报档案领域综合关联度≥0.4且具有双向关联的期刊关联云图

图 3 中，期刊之间存在连线，表明期刊之间存在双向关联，即双方相互依赖的程度均较高；期刊的连线越多，表明这种期刊在该领域中不仅对其他期刊具有影响，而且自身也颇受领域内其他期刊的影响，是领域内比较传统的学术期刊，例如《图书情报知识》、《大学图书馆学报》、《情报资料工作》等。

从图 3 中笔者发现：

（1）档案学类两种核心期刊《档案学通讯》和《档案学研究》自成体系，且保持了较高的关联度，表明档案学目前的研究内容与图情学存在较大差异，鉴于两者属于同一学科的事实，可以进一步挖掘两者的结合点，利于产生新的研究方向以实现创新；

（2）《中国信息导报》和《中国科技期刊研究》与图情档类期刊关系并不密切：或者充分依赖于图情档类期刊但反过来对其影响较小，属于单方面依赖，如《情报杂志》、《情报科学》、《情报理论与实践》对《中国信息导报》的关联度（影响）均超过了 0.5，但是反过来，《中国信息导报》对这 3 种期刊的影响较小，关联度列于倒数；或者与图情档类期刊研究内容存在差异，双方关联程度都不大，如《中国科技期刊研究》，其更多依赖于编辑出版类期刊，对图情档类期刊的影响也较小；

（3）其他 16 种期刊相互关联形成一个整体：图书馆学类期刊相互之间普遍存在较高的关联度，在图 3 中表现为相互之间的连接较多，而相反情报学类期刊的高关联则相对较少，图书情报综合类则介于两者之间，可以看出图书馆学的研究范围相对较大，而情报学起源于图书馆学，是对图书馆学某些领域研究内容的深化，专业性较强；

（4）作为情报学重要刊物的《情报学报》似乎并没有成为其他期刊的依赖，究其原因笔者认为一方面是由于其作为双月刊文章数量较少，研究内容多注重学科前沿且比较专深，与其他刊物拥有交叉内容的机会较少；另一方面图 3 中显示的均为双向强关联，《情报学报》对其他刊物可能影响甚大，但是并不保证其受其他刊物影响也甚大；

（5）与《情报学报》保持高关联度的期刊为《情报理论与实践》，说明这两种期刊在研究内容上具有较大的相似性，相互依赖较多，而《情报理论与实践》也是学科内为数不多的、能够对《情报学报》产生较大影响的重要刊物。

3.3　基于期刊间平均关联度的多维尺度分析

图 3 揭示了期刊之间的双向关联情况，以期刊间连线的粗细来描述关系的密切程度，由于 A 对 B 和 B 对 A 的关联程度不同，在图中则表现为连线的两头粗细不同，甚至因为某些单向关联度较小导致期刊间的关联在图中没有显示。如果能够在二维平面中以期刊间距离来描述期刊间的平均关联，则可以根据期刊在平面中的分布情况进行大致的分类，得到聚类结果，这一过程可以通过多维尺度分析和层次聚类分析来实现，即建立期刊 × 期刊的相似（或距离）矩阵，以期刊间的平均关联度作为矩阵相似值，然后通过降维操作将期刊间相似情况转化到二维平面中，用点间距离表示期刊之间的关联度。

图情档领域 20 种期刊的多维尺度分析结果如图 4 所示：

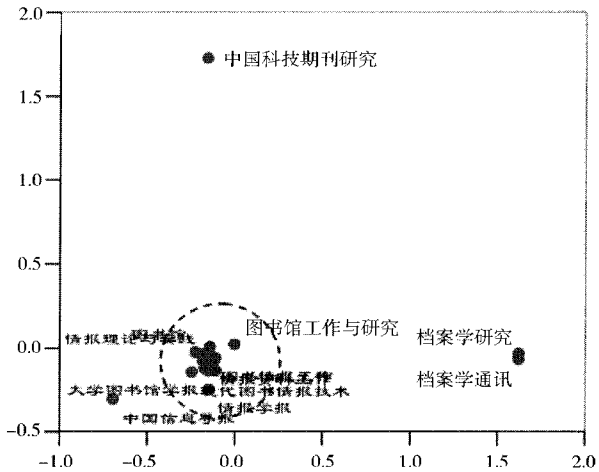

图 4　图书情报档案领域 20 种期刊间关系的多维尺度分析结果

其中，两个分析指标 Stress = 0.00861，DAF = 0.99139，表明模型的拟合效果很好，具有较高的置信度。

（1）期刊聚类分析结果，如图 5 所示。

从图 5 可以发现，图情档领域的 20 种期刊被明确划分为 4 个类别：

①《中国科技期刊研究》自成一类，考察该期刊的关联期刊，不难发现其最依赖的多为《编辑学报》、《编辑之友》、《中国出版》等新闻出版广播领域的期刊，而与图情档领域期刊的关联度较小；

②档案学期刊《档案学通讯》和《档案学研究》为一类，说明档案学研究和图情研究在内容上存在较大差异，这与图

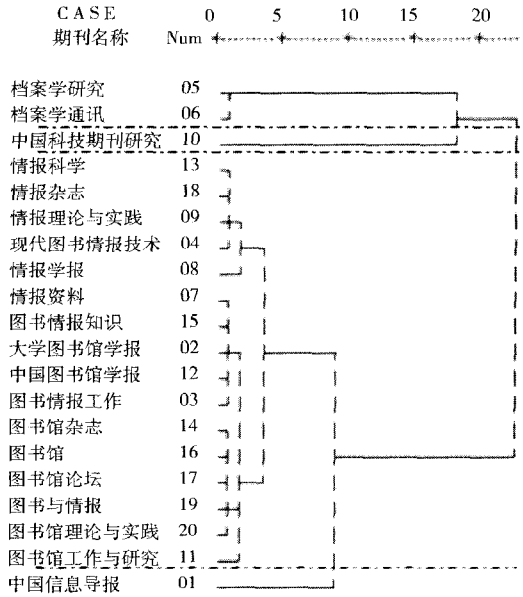

图5　图情档领域期刊的聚类分析结果

3中所示一致；

③《中国信息导报》自成一类，该刊虽然依赖于图情各刊，然而其通常不是图情各刊多依赖的期刊，对其他期刊的影响较小；

④其他16种图书情报类期刊分为一大类，如图4中虚线圆围成的区域。

(2)将图4中虚线圆围成的区域放大，如图6所示：

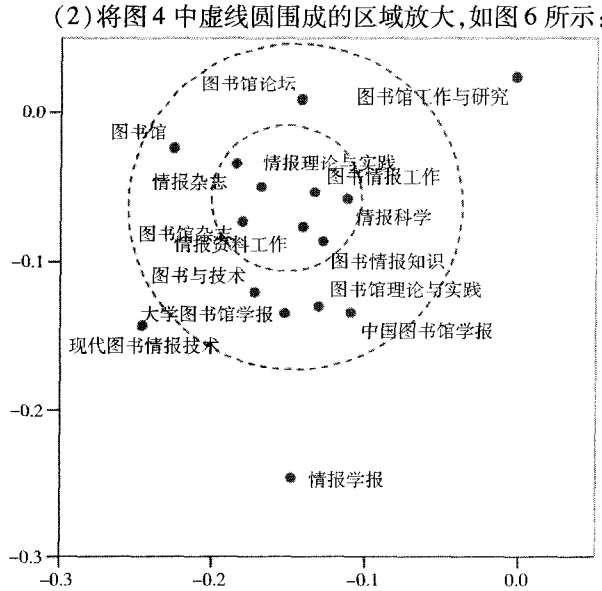

图6　图书情报类16种期刊间关系的多维尺度分析结果(中心放大)

结合聚类的系统树状图可以发现：

①其中《情报学报》由于文章少而且研究主题专深，导致其与其他各刊之间关联不是很紧密，印证了基于期刊间双向关联的综合分析中的分析结果；

②《图书馆工作与研究》是双月刊，偏重于图情理论和方法在图书馆工作和事业中的具体实施研究，《现代图书情报技术》偏重于计算机技术在图情研究中的具体应用，这两种刊的专业性较强，研究范围相对专业，与各刊也存在较大差异；

③居于图形中心的多为情报类和图书情报综合类期刊，而图书馆学类期刊则居于外围，说明总体而言前者相比于后者更是图情档领域学术论文所关注的重心，具有较大的学术影响，小圈内的期刊可以认为是2000－2006年间图书情报档案领域的核心期刊集。

4　结　语

本文重点探讨了CSSCI来源期刊之间关联的构建及分析，这是基于CSSCI知识本体实现学术资源关联分析的一个组成部分。在CSSCI_Onto中揭示的学术资源知识基础上，不仅可以了解期刊之间的两两关系及其关联程度，而且能够挖掘期刊之间的双向和多元关联，发现一定时期内学科领域中的核心期刊集，为学科内期刊之间的合作和分工提供可参考的事实依据，实现期刊研究内容的合理分布。

笔者改变了学术资源评价惯有的分析思路，即根据分析目的聚集数据，再将数据升华为知识，得到分析结论；而是先建立完整的知识库，继而考察在该知识库中可以得到的分析结论。基于知识的分析模式完全改变了原来的基于数据的传统分析模式，期刊间的关联也不再建立于单一关联标准之上，在此基础上获得的分析结论也相对合理和有效。本文仅对学术资源中的期刊概念进行了关联分析，而知识本体中蕴含的丰富知识可以实现更广泛意义的学术资源关联分析，更多有价值的信息有待于今后进一步挖掘。

参考文献：

[1] 南京大学中国社会科学研究评价中心[EB/OL]. [2011－01－02]. http://cssci.nju.edu.cn.

[2] 苏新宁. 中国人文社会科学期刊学术影响力报告[M]. 北京：中国社会科学出版社，2009.

[3] 宋唯娜，杨康. 中国文学期刊引用网络分析——基于CSSCI(2003－2007年度)数据[J]. 西南民族大学学报：人文社会科

学版,2011(1):236 - 240.
[4] 梁勇,章成志,王昊. 基于 CSSCI 的期刊知识地图的构建[J]. *现代图书情报技术*,2008(2):58 - 63.
[5] 金莹. 数据挖掘在 CSSCI 中的应用[D]. 南京:南京大学,2006.
[6] 王昊. 信息资源网络模型及引用[M]. 南京:南京大学出版社,2010.
[7] 王昊,苏新宁. 基于本体的 CSSCI 学术资源网络模型构建及其应用研究[J]. *情报学报*,2010,29(2):331 - 341.
[8] 王昊,苏新宁. 基于 CSSCI 本体的学科关联分析[J]. *现代图书情报技术*,2010(10):10 - 16.
[9] Astrova I. Reverse Engineering of Relational Database to Ontologies [C]. In: *Proceedings of the ESWC* 2004. 2004: 327 - 341.

(作者 E - mail:ywhaowang@nju.edu.cn)

雅虎线索:新的搜索数据源

雅虎目前使用微软的 Bing 来提供其网络搜索服务,但是该公司仍将在网络搜索上不断创新。雅虎不断为其各种搜索产品添加新的功能和接口。最近的一个例子就是 2010 年 11 月中旬推出雅虎线索(Yahoo! Clues)。虽然现在只是 Beta 版,而且是早期产品,但是雅虎线索提供的独特信息是不容易从其他来源获得的。

雅虎线索呈现了有关流行搜索词的信息,这些搜索词统计自"数以百万计的人们每天在雅虎上进行的搜索"。与 Google 趋势和 Google 搜索解析类似,雅虎线索既有类似的也有其独特的数据统计值。目前可使用的数据包括搜索量、搜索人口特征(年龄、性别和收入分布)、搜索者地理位置、搜索流量,以及相关搜索。广告商可以使用这个免费的工具来获取对有关搜索模式和搜索人口特征数据类型的感性认识。

雅虎线索的搜索屏幕上有两个搜索框。第二个搜索框用于比较两个查询的情况,因此可以只使用第一个搜索框来查看某个查询的情况。输入一个查询之后,跳到下一个框或者点击"发现"按钮查看结果。这时,可以看到三个时间段的统计数据:当天、过去 7 天、过去 30 天。Google 趋势有更长时间段的数据,可以追溯至 2004 年,但是搜索解析只能追溯到过去 30 天或者过去 7 天。因此,与 Google 趋势相比,雅虎线索提供了更多的最近搜索活动信息。

由于现在尚处于早期阶段,因此不是所有的查询词都能查到结果。对于那些有结果的查询,结果数据部分都有相应的图形化表示,而且大多显示有百分比值。

顶部的搜索量图是用百分制来衡量的,100 分表示过去一个月内搜索量居于最高。如果是比较两个查询,那么这两个查询中查询量更大的那个标以 100 分。接下来是人口统计特征,Google 趋势和 Google 搜索解析目前还没有提供这项特征。显示有年龄分组、性别分组、收入分组。雅虎解释说"雅虎线索聚集了整个雅虎搜索的年龄和性别信息",收入分层是在假设的基础上,基于雅虎搜索里通用邮政编码信息和美国人口普查局的各州收入数据的匹配。这些人口统计数据在浏览时是很吸引人的,图形表示使得人们很容易理解,而不需要进行复杂的数字计算。这些人口统计数据,特别是收入数据,可信度是一个值得商榷的问题。Erik Wagner 比较了 Google 趋势和雅虎线索之后在其博客上指出"收入数据可能是非常不精确的","我不相信这些人口统计信息"。

相对来说,人们更相信搜索地理位置信息,因为过去几年里地理定位技术有很大发展。在"按位置统计"部分,目前只专注于美国(不像 Google 趋势那样提供全球数据)。雅虎同时提供国家的地图并标明搜索活动最频繁的前 10 个州。雅虎线索在帮助页面上这样写道:"为了避免人口最多的几个州一直排在前几名,雅虎线索使用的是相对人口数量,并显示了前 10 名。我们使用百分制来显示数据,其中,100 分表示搜索最集中。"点击地图中前 10 名中的州将显示该州搜索量排名前十的城市。

雅虎将显示前续和后续查询的部分命名为搜索流。搜索流是雅虎独有的信息,Google 并没有提供。它显示在地理位置部分的下方,并且,当鼠标移过一些人口统计数据时也会显示。对大多数搜索来说,前续 5 个和后续 5 个搜索会显示出来,以帮助用户查看其他人的搜索过程序列。

雅虎线索提供的最后一个信息是在显示屏底部的"相关检索",雅虎是这样解释它的功能:这些是"整个雅虎搜索里最常用的相关的检索词,不限于搜索流里显示的用户搜索模式。"

虽然与 Google 趋势类似,雅虎线索仍提供了额外的信息。搜索流和人口统计特征能揭示某些方面的搜索者行为和差异,而这些是平常无法获取的信息。

(编译自:http://newsbreaks.infotoday.com/NewsBreaks/Yahoo - Clues - A - New - Source - for - Search - Data - 71982.asp)

(本刊讯)