基于多重共现的可视化分析工具设计

及其知识发现方法研究

庞弘燊[1,2]，方曙[1]

（1.中国科学院国家科学图书馆成都分馆 四川成都 610041；2.中国科学院研究生院 北京 100190 ）

摘　要： 本文在多重共现概念界定基础上，阐述了它与一般共现的区别，同时对 Morris 交叉图技术进行改进，开发了多重共现的可视化分析工具，并对多重共现的知识发现方法进行了研究。最后还综合运用多重共现可视化分析工具及多重共现的知识发现分析方法来对机构-期刊-关键词的样本数据进行了实证分析。通过分析发现，该套工具以及知识发现方法能较为有效地发现论文中三个特征项之间的多重共现关系，并能揭示出比一重共现现象更为广泛和深入的信息内容。

关键词： 共现；多重共现；可视化分析工具；共现交叉图；知识发现

Research on Knowledge Discovery Method and Design of Visual Analysis Tool Based on Multiple Occurrence of Entities in Papers

Pang Hongshen[1,2] , Fang Shu[1]

(1 The Chengdu Branch of the National Science Library, Chinese Academy of Sciences, Chengdu 610041; 2 Graduate University of Chinese Academy of Sciences, Beijing 100190)

Abstract: This paper discussed the concept of multiple occurrence and analysed the differences between multiple occurrence with the general occurrence. Meanwhile it improved the Morris' cross-mapping technology for the analysis of multiple occurrence. The authors developed a tool and designed a knowledge discovery method for a case study of

作者简介：庞弘燊（1983-），男，广东佛山，情报学博士研究生，email:winsunpang@126.com

institutions - journals – keywords multiple occurrence. By the case study, this paper found that the tool and the knowledge discovery method of multiple occurrence analysis can analysis the occurrence relationship among three entities in papers more effectively, and reveal more and deeper information than the general occurrence methods.

key words: occurrence; multiple occurrence; visual analysis tools; cross-map; knowledge discovery

1 引言

科技发展日新月异，在信息迅速膨胀与高度开放的今天，随着科学知识的普及、科学思想的传播、科学理论的研究、科学成果的应用和推广等使得信息源越来越庞大。在激增的信息当中，包含着许多科学活动规律的重要知识。

文献计量学者很早就注意到论文共现（co-occurrence and occurrence）现象，通过分析共现现象可以从多个角度解释、挖掘隐含在论文中的各类信息，揭示论文与论文之间的内容关联和逻辑关联。由于共现现象可以转换为形式化的表述方式（如共现矩阵，共现关系图）加以定量测度，尤其是在计算机技术的辅助下，共现分析以其方法的简明性和分析结果的可靠性，成为支撑信息内容分析研究过程的重要手段和工具，受到了研究者的关注并进行了大量理论探讨与应用研究。而学术期刊上单独成篇发表的论文数据看似孤立，实则有着千丝万缕的关联。每一篇论文都由若干个特征项(entities)组成，包括关键词、作者、机构、发表期刊等[1]。这些特征项结合在一起构成了一篇论文的重要特征，也是论文之间相互区别的重要特质。在文献计量研究中，通常用分析特征项之间关联的方法去探索论文的关联，进而映射科学领域在不同方面的关联结构，揭示科学活动的发展规律。

2 多重共现的概念

目前，对共现现象的研究主要从两个不同或相同的特征项共现展开，本研究致力于将三个特征项共现的现象作为研究主体，在总结现有的共现研究方法、数据挖掘技术、知识发现方法的基础上，拓展共现现象的研究范围，并设计出一套可应用于分析多重共现现象的工具及知识发现方法。

笔者把期刊论文中的多重共现定义为三个（含）以上相同类型或不同类型特征项共同出现的现象，如作者-关键词-发表期刊三个特征项同时在多篇论文中出现，作者-关键词-被引作者、作者-引文作者-关键词-引文关键词等三个或以上特征项的共现都属于多重共现研究的范畴。多重共现与一般的两个特征项共现（以下称作一重共现）相比，如作者-关键词-发表期刊的多重共现比作者-关键词、作者-发表期刊等的一重共现能够揭示更为深入的知识。分析作者-关键词-发表期刊的多重共现就相当于分析作者-关键词、作者-发表期刊、关键词-发表期刊的三个共现及其之间的关系。如下表 1 所示，可以看出多重共现现象对于揭示深度知识方面有着独特的优势。而图 1 中则形象地表示了多重共现与一重共现现象在研究对象上的区别：

表 1 多种共现特征项所能揭示的知识内容

Tab.1 The revealing knowledge contents of variety entities

特征项	例子	分析的视角	所能揭示的知识内容
一个特征项	作者	高产作者	高发文量的作者
	关键词	高频关键词	热门研究主题词
两个特征项共现（一重共现）	关键词-关键词	共词分析	关键词聚类揭示研究主题
	作者-关键词	作者与关键词关系分析	作者的研究领域
三个（含）以上特	作者-关键词-发表期刊	作者、关键词与发表期刊之	作者偏好在某期刊上所发表的主题类型、某期

征项的共现（多重共现）		间的关系分析	刊的固定作者群及主题研究领域与变化等
	作者-关键词-引文关键词	作者、关键词与引文关键词之间的关系分析	通过关键词聚类和引文关键词聚类共同反映作者的研究领域

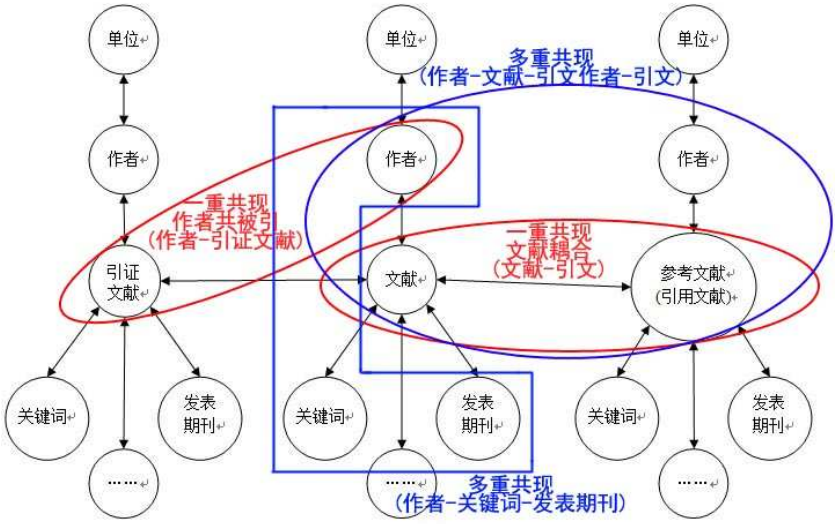


图 1 多重共现与一重共现研究对象的区别

Fig.1 The differences of research object between multiple occurrence with the general occurrence.

3 基于多重共现可视化分析工具的设计

要实现对海量论文数据的量化分析，就必须对文献数据进行特征提炼，抽取可以定量分析的结构化数据。揭示某种特征项内部的关联结构，是目前大部分的可视化技术所能实现的，并在科学计量研究中被广泛应用。比如为了揭示两种特征项之间的关联，美国科学计量专家 Morris[2,3]借助于两个共现矩阵相同特征项之间的关联，开发了交叉图和时间线技术并进行了应用研究，两种技术可以很好地弥补目前可视化技术不能揭示两种特征项关联的缺陷。Leydesdorff[4]把“异质网络”的思想进一步扩展到了 3-mode 网络，他把作者-期刊-关键词的特征项联系起来，通过不同类型节点在同一网络中的展现，不仅有利于分析同一类型节点间以及不同类型节点间的关系，而且也是研究网络更加真实的反映。还有一些分析工具也能对文献特征项的共现进行辅助分析，如 Thomson Data Analyzer (简称 TDA)，是一个具有强大分析功能的文本挖掘软件，可以对文本数据进行多角度的数据挖掘和可视化的全景分析[5]，并能对文献的各种特征项生成共现矩阵，以分析其共现关系。CiteSpace 则通过分析论文、引文、关键词、年份等特征项之间的共现关系来识别学科的发展趋势和前沿领域。

3.1 多重共现的可视化方式

基于多个不同特征项共现关系的可视化方法与基于两个特征项关联的科学知识图谱可视化方法(包括主成分分析、聚类分析和多维尺度分析等多元统计分析方法)相比，在反映科学活动规律和科学知识领域方面增加了多个分析角度和信息来源，因此，其中蕴含的信息量也大幅度增加，有很大的挖掘和探索价值。

本研究为了能更有效地显示出多重共现中各特征项的关联关系，笔者借鉴了 Morris 的交叉图显示方式，并对其作出改进，如图 2 所示是 Morris 的交叉图，而图 3 则是笔者对其作出改进后的交叉图显示样例。Morris 的交叉图技术分别用 x 轴、y 轴表示两种不同类型的特征项，首先利用 x 轴、y 轴轴线揭示每种特征项内部的关联结构，然后在 x 轴、y 轴的交叉点用节点（颜色、大小）显示两种特征项关联强度。其交叉图中不仅包括了两个不同特征项的关联关系，还包含了这两个特征项自身各自的聚类关系。而为了能够显示出三个特征项之间的关联关系，笔者在 x 轴、y 轴的交界处加

入了另一特征项来显示其与 x 轴、y 轴上两种特征项的多重共现关系（如图 3 所示），同时在坐标的外围处，笔者也加入该另一特征项来分别显示其与 x 轴、y 轴特征项的共现关系。从图 2 中可以看出，Morris 的交叉图着重于显示两个特征项之间的关联关系，如图 2 中机构合作与研究主题交叉图可以考察哪些机构合作研究了哪些相关的研究主题；而笔者改进后的交叉图除了可用于显示两个特征项之间的关联关系之外，还可以显示三个特征项之间的共现关系，如图 3 中机构-期刊-关键词多重共现交叉图可以考察哪些机构在哪些期刊中发表某类研究主题（关键词）的论文。多重共现的交叉图可以揭示更多的信息，挖掘出原有交叉图可视化技术所无法揭示的信息。

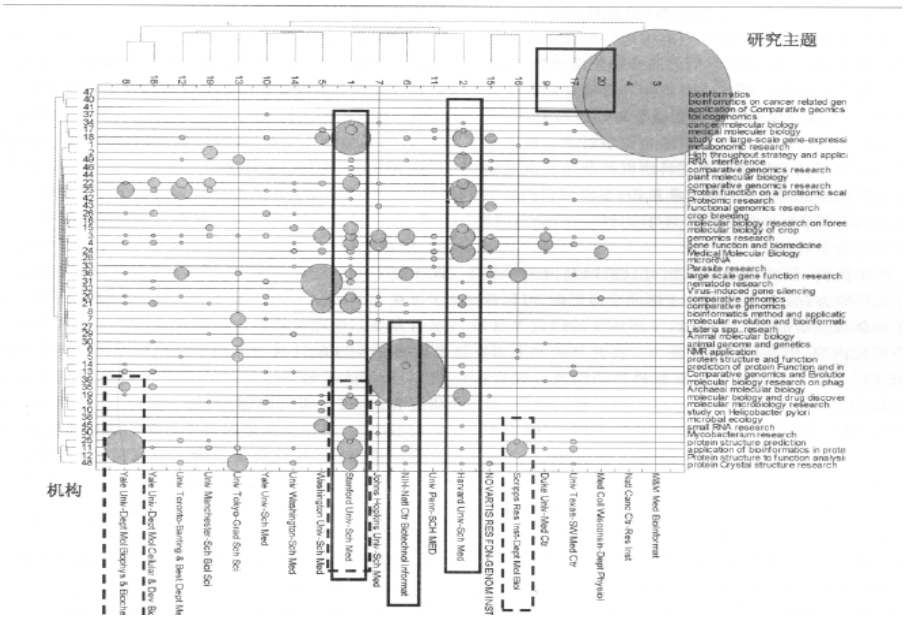


图 2 Morris 机构-研究主题交叉图[6]

Fig.2 Morris' cross-map[6]

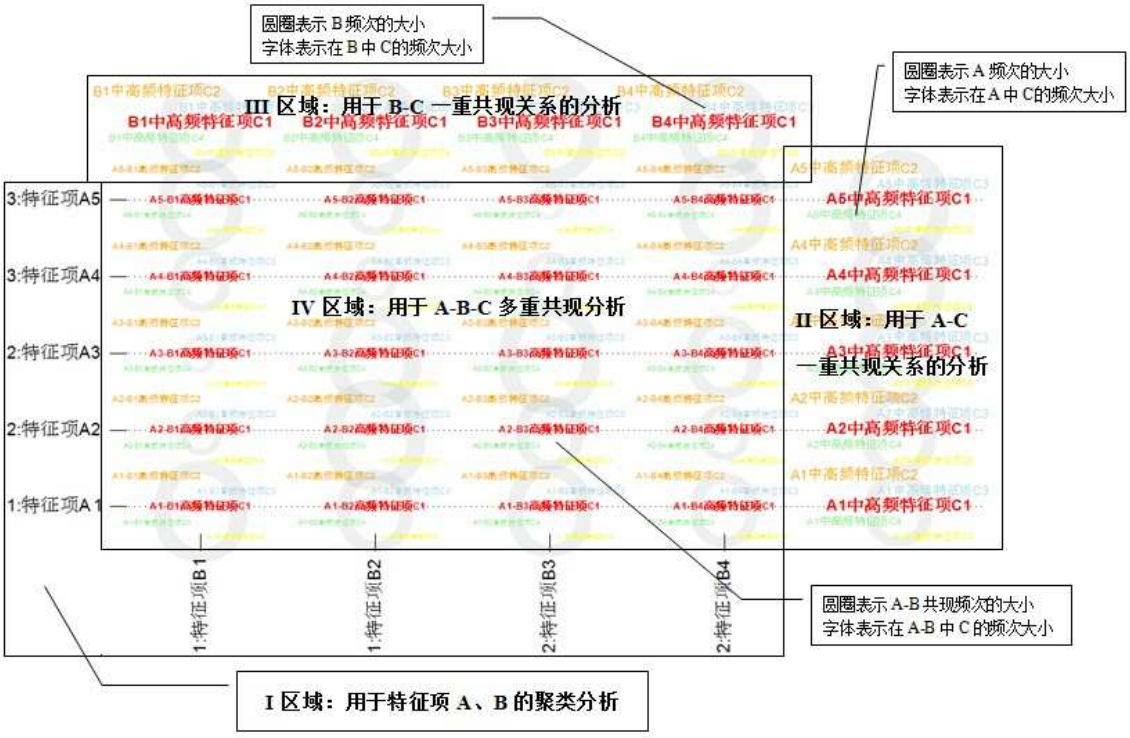


图 3 多重共现交叉图

Fig.3 Cross-map of multiple occurrence

3.2 基于多重共现可视化分析工具的框架

基于多重共现可视化分析的要求，笔者设计了适用于分析多重共现的可视化分析工具框架，如下图所示：

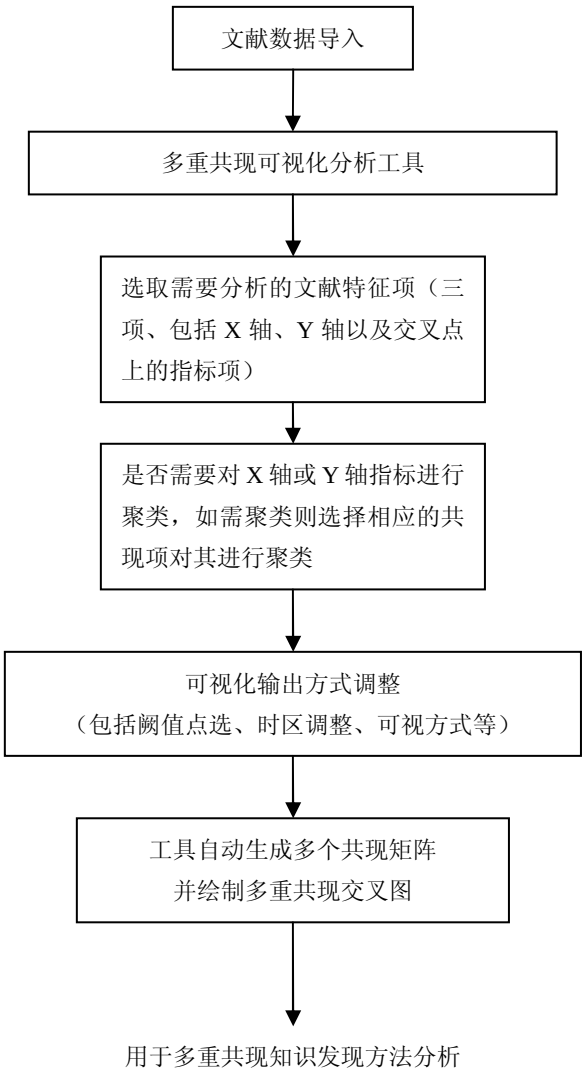


图 4 基于多重共现可视化分析工具的框架

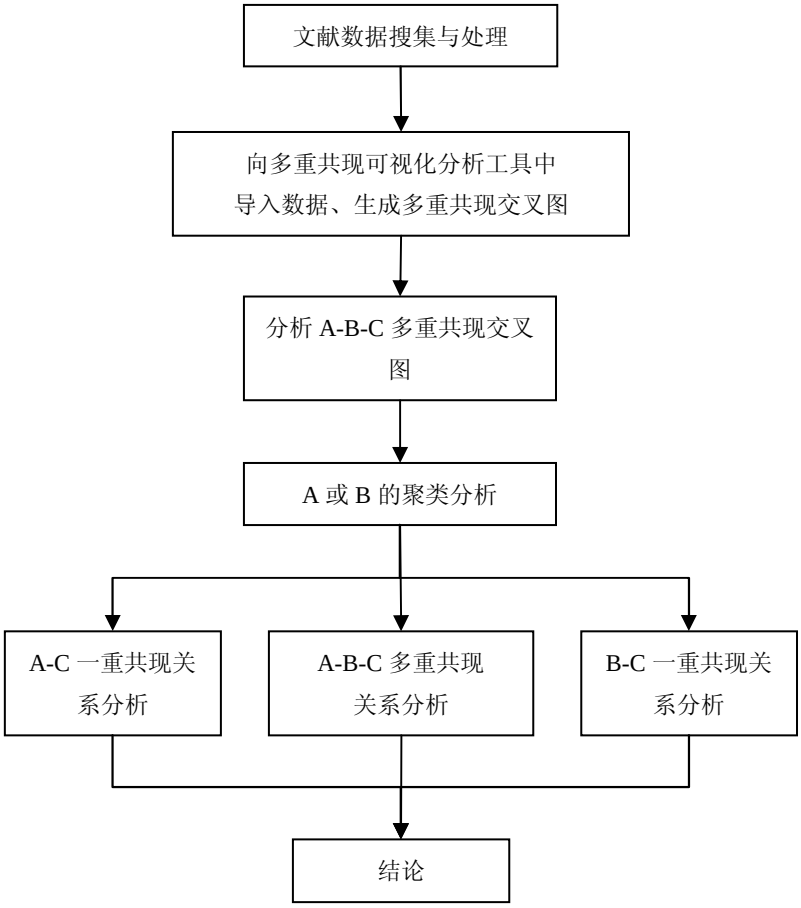
Fig.4 The framework of visual analysis tool based on multiple occurrence

该多重共现可视化分析工具可以分析的范畴包括三个特征项之间的共现关系，以及其特征项两两之间的共现分析，还能根据用户的需求对 X 轴或者 Y 轴上的特征项进行聚类，以下将基于该可视化分析工具详细阐述关于多重共现知识发现方法的分析流程。

4 基于多重共现的知识发现方法研究

4.1 分析流程

在基于一般共现分析方法以及多重共现可视化分析工具设计的基础上，本研究设计了一套可用于多重共现分析的知识发现方法，如下图所示：



注：A 代表 X 轴上的特征项、B 代表 Y 轴上的特征项、C 代表交叉点上的特征项

图 5 基于多重共现知识发现方法的分析模型

Fig.5 An analysis model of knowledge discovery method based on based on multiple occurrence

首先对需要分析内容的文献数据进行搜集，然后把搜集到的样本数据导入到多重共现可视化分析工具中生成多重共现交叉图。在分析多重共现交叉图的时候，可以根据分析需求分有五方面的分析内容让用户自主选择或组合：（1）聚类数据的分析（如图 3 中 I 区域所示），X 轴或 Y 轴上特征项可以根据任意特征项的共现来进行聚类，并以前面标识的数字显示出分类后的结果；（2）A-C 一重共现关系的分析（如图 3 中 II 区域所示），可对 A-C 共现的频次、特点等进行分析；（3）B-C 一重共现关系分析（如图 3 中 III 区域所示），可对 B-C 共现的频次、特点等进行分析；（4）A-B-C 多重共现关系分析（如图 3 中 IV 区域所示），可对 A-B-C 共现的频次、特点等进行分析；（5）整体结论分析，可对以上四个区域的聚类和共现特点进行总体性的分析或概括等。

4.2 实证分析

为了能更有效说明上述的基于多重共现知识发现方法，本研究选取了机构-期刊-关键词的多重共现作为实证分析，如图 6 所示是机构-期刊-关键词多重共现的具体分析流程图，

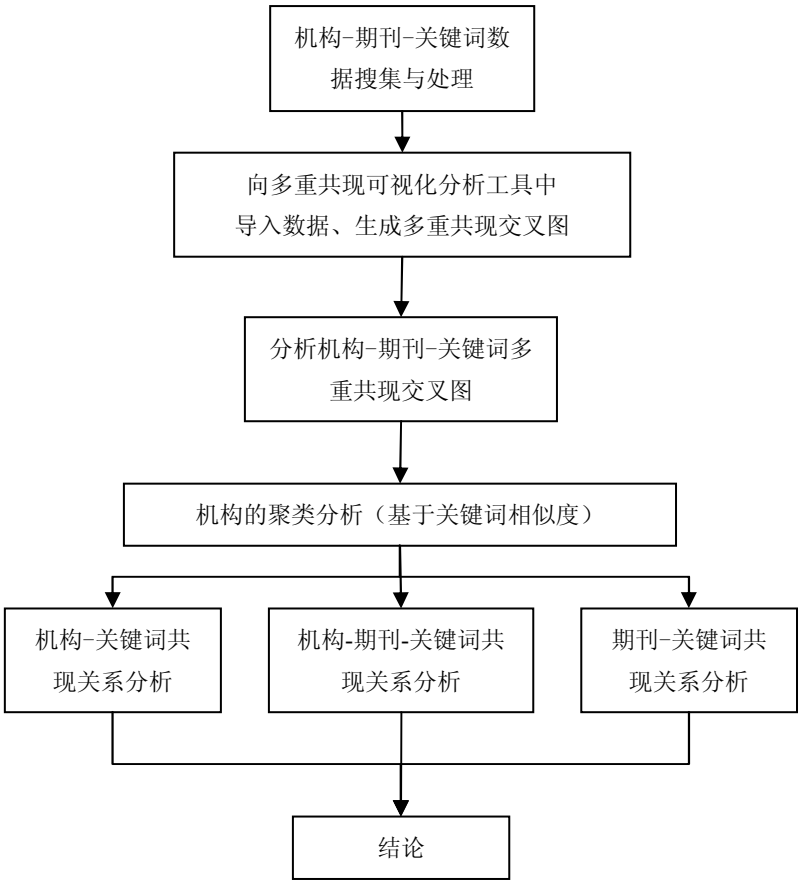


图 6 机构-期刊-关键词多重共现分析流程图

Fig.6 The analysis process of institutions - journals - keywords multiple occurrence

本研究的实证分析拟通过分析机构-期刊-关键词的这三个特征项的共现关系，来发掘它们之间存在的关联信息。比如在分析机构-期刊-关键词这三个特征项的共现关系中，如果从机构的视点出发，可以发掘出哪些机构偏向于在哪种期刊上发表某类研究主题的文章；从期刊的视角出发，可以揭示某类期刊上所具有的稳定机构作者群，还有该机构在期刊上所发表论文的主题方向；而从关键词的角度出发，可以找出发表关于某类主题论文的机构群体和期刊集合。

实证研究需要限定所选取机构和期刊的样本数量，以在 X 轴和 Y 轴上有效地显示。因此笔者选取中国校友会网 2011 中国大学排行榜 20 强高校[7]的图书馆作为机构分析样本，在期刊样本上则选择了 CSSCI（2010-2011 年）来源期刊 “图书馆、情报与文献学”中的 18 种图书馆学、情报学期刊[8]（以下简称 18 种期刊）。然后根据所选取的图书馆样本和期刊样本在 CNKI 的中国学术期刊网络出版总库数据库搜索相关的论文，论文发表年份限定为 2006-2010 年（检索日期：2011 年 3 月 12 日），共检索出 1114 篇论文（如表 1 所示），基本可以认为检索出的论文集合代表着目前高校图书馆的主要研究方向。

通过数据导入，形成如图 7 所示的多重共现交叉图，图 7 中间区域圆圈所示代表机构-期刊的共现频次大小（相当于某机构在某期刊上发文量的大小），发文越多，圆圈越大。高频关键词区域则可显示某机构在某期刊上发表论文所使用的高频关键词状况，按照其所使用的关键词频次高低，标以不同的颜色深浅和字号大小作为区别，关键词频次越高，其标识的颜色越深并且字号也越大。

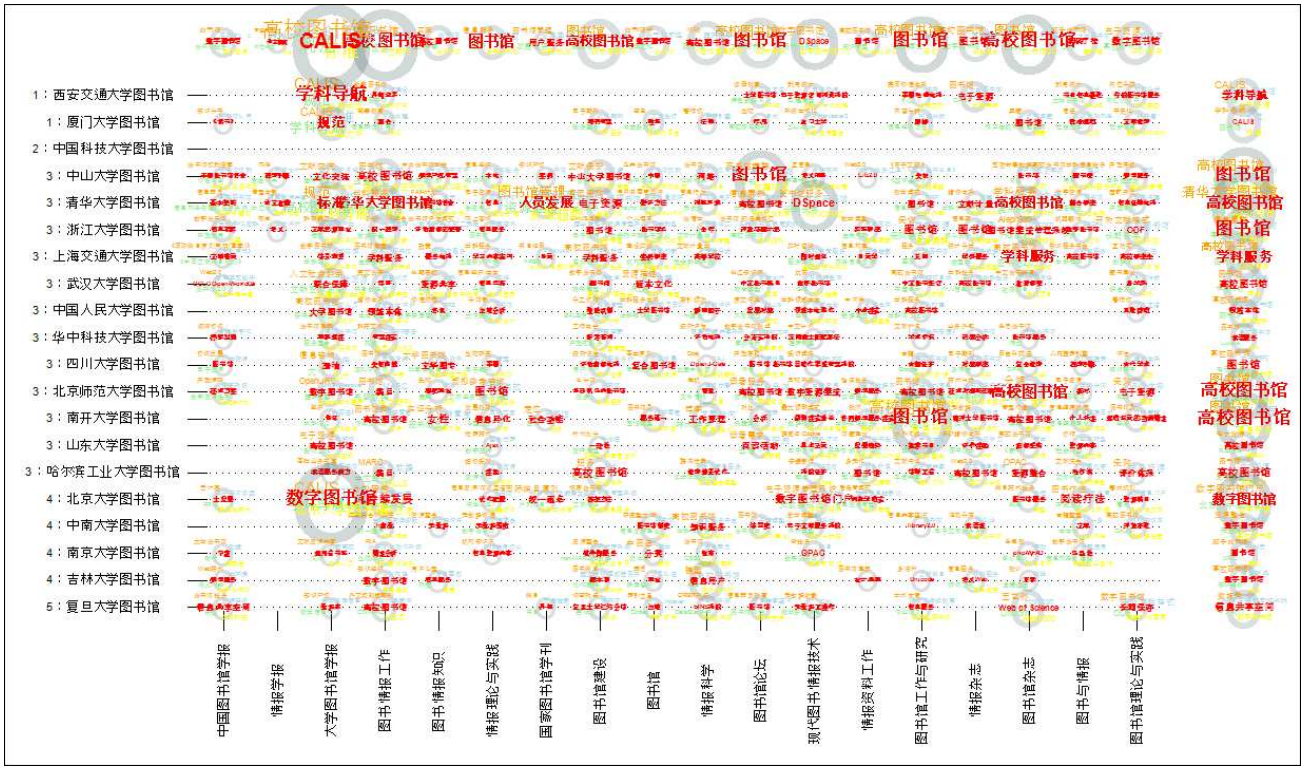


图 7 机构-期刊-关键词多重共现交叉图

Fig.7 Cross-map of institutions - journals - keywords multiple occurrence

根据图 7 所示的机构-期刊-关键词多重共现交叉图，笔者依据基于多重共现知识发现方法的分析流程，从以下五个方面进行分析：

（1）聚类数据的分析：从图 7 左侧机构的分类数字标识来看，多重共现可视化分析工具依据各高校图书馆在 18 种核心期刊上发表论文的机构-关键词共现矩阵，对机构分成了五类。1 类：包括西安交通大学图书馆、厦门大学图书馆；

2 类：包括中国科技大学图书馆（在数据库中没有查找出以中国科技大学图书馆为署名单位发表在 18 种期刊中的论文）；

3 类：中山大学图书馆、清华大学图书馆、浙江大学图书馆、上海交通大学图书馆、武汉大学图书馆、中国人民大学图书馆、华中科技大学图书馆、四川大学图书馆、北京师范大学图书馆、南开大学图书馆、山东大学图书馆、哈尔滨工业大学图书馆；

4 类：北京大学图书馆、中南大学图书馆、、南京大学图书馆、吉林大学图书馆；

5 类：复旦大学图书馆。

从分类当中可以认为，在同类别中的机构间在 18 种期刊中发表的论文主题较为相似，而不同类别的机构间发文主题则差异较大。

（2）机构-关键词一重共现关系的分析：

高校图书馆在核心期刊中发文的高频关键词，可以折射出高校图书馆的主要研究方向。根据图 7 右侧从中可以看出高校图书馆非常关注服务，许多研究都是围绕用户服务展开，比如信息服务、读者服务、图书馆服务、学科服务等的一些高频词。同时高校图书馆也很注重资源环境建设等方面的研究，如关键词：数字图书馆、电子资源、网络环境、数据库、数字资源、CALIS 等就有所反映。而从发文量上来看，清华大学图书馆、中山大学图书馆、北京大学图书馆、南开大学图书馆这四个机构发文量最多。

（3）期刊-关键词一重共现关系分析

通过对核心期刊刊载高校图书馆论文的关键词进行分析，可以找出各种期刊偏重于录用高校图书馆关于哪类主题的研究论文。从图 7 上侧可以看出，18 种期刊刊载了许多以高校图书馆、数字图书馆、大学图书馆为关键词的论文，其中发现某些期刊所刊载论文聚集在某几类的关键词，如《大学图书馆学报》刊载了许多以 CALIS、学科导航、规范、标准为关键词的论文，《图书馆杂志》刊载了较多以学科服务、OPAC、图书馆 2.0 为关键词的论文，《图书馆工作与研究》刊载了较多以信息服务、采访、采购模式为关键词的论文，《图书馆建设》刊载了较多以电子资源、集团采购、图书馆服务为关键词的论文。除此以外，有一些期刊偏重于录用关于技术类主题的论文，如《情报科学》刊载了较多以网络、搜索引擎、个性化信息服务为关键词的论文，《现代图书情报技术》刊载了较多以 DSpace、数字图书馆、OPAC、数字图书馆门户为关键词的论文，还有如《国家图书馆学刊》偏重于刊载研究图书馆实务和发展的论文，其高频关键词包括有用户服务、图书馆管理、资源建设、发展趋势、战略规划。由此可见，各核心期刊所刊载的高校图书馆论文的研究主题各具特色。而从载文量来看，《大学图书馆学报》、《图书情报工作》、《图书馆杂志》、《图书馆工作与研究》刊载了较多由 20 所高校图书馆所发表的论文。

（4）机构-期刊-关键词多重共现关系分析：

图 7 中间区域圆圈大的地方代表某机构在某期刊上发表了较多的文章，如北京大学图书馆在《大学图书馆学报》上发表了大量的文章，中山大学图书馆在《图书馆论坛》上发表了较多的文章，清华大学图书馆在《现代图书情报技术》上发表了较多的文章，南开大学图书馆在《图书馆工作与研究》上发表了较多的文章。从中可以看出高校图书馆与核心期刊的发文关系具有一定的地域性特点，即高校图书馆的研究人员会较为集中地在其所处地域（或邻近地域）的核心期刊上发表较多的文章。

而从高频关键词的标识上来看，北京大学图书馆在《大学图书馆学报》上发表了许多以数字图书馆、CALIS 为关键词的论文，西安交通大学图书馆在《大学图书馆学报》上发表了许多以 CALIS、学科导航为关键词的论文。除此以外，还发现有多所大学图书馆在多种核心期刊上发表了较多以高校图书馆、CALIS、图书馆为关键词的论文，这三个关键词在图 7 中的机构-期刊-关键词交叉区域频繁地出现，这标志着高校图书馆的研究以图书馆类研究为主，较少涉及情报类研究，并且都是以高校图书馆这一群体的视角出发，较多地关注于 CALIS 的研究。另外各个高校图书馆在各种期刊上也发表多种主题的研究论文，如厦门大学图书馆在《大学图书馆学报》上发表了较多以学科导航、标准、规范、CALIS 为关键词的论文，上海交通大学图书馆在《图书馆杂志》上发表了较多以学科服务为关键词的论文，北京大学图书馆在《图书与情报》上发表了较多以阅读疗法为关键词的论文、北京师范大学图书馆在《图书馆杂志》上发表了较多以高校图书馆为关键词的论文、南开大学图书馆在《图书馆工作与研究》上发表了较多以信息服务、图书馆员、美国为关键词的论文，清华大学图书馆在《图书馆杂志》上发表了较多以学科服务、学科馆员为关键词的论文。由此可见各个高校图书馆除了会在某些期刊上发表关于某几个特定主题的文章之外，在各类期刊中发文的主题可以说是百花齐放。

（5）整体结论分析：

此外从图 7 中笔者还发现某些高校图书馆发表的论文会采用本身机构的名称作为关键词，比如北京大学图书馆、清华大学图书馆、南开大学图书馆、中山大学图书馆等。由此可以看出这几个高校图书馆其研究多从本馆的工作出发，注重以本馆实践来做研究。

最后通过对上述几点分析的总结，笔者归纳出以下四点结论：（a）高校图书馆与核心期刊的发文关系具有一定的地域性特点；（b）高校图书馆较为注重用户服务和资源环境建设等方面的研究；（c）各核心期刊所刊载的高校图书馆论文的研究主题各具特色。（d）各高校图书馆除了会在某些核心期刊上发表关于某几个特定主题的文章之外，在各类期刊中发文的主题可以说是百花齐放。

5 小结

本研究提出了多重共现的概念，并阐述了其与一般共现的区别，同时对 Morris 交叉图技术进行改进，开发了多重共现的可视化分析工具，使其适用于多重共现的分析，并对多重共现的知识发现方法进行了研究。最后还综合运用多重共现可视化分析工具及多重共现的知识发现分析方法来对机构-期刊-关键词的样本数据进行了实证分析。通过分析发现，该套工具以及知识发现方法能较为有效地发现论文中三个特征项之间的多重共现关系，并能揭示出比一重共现现象更为广泛和深入的信息内容。

此外，本研究当中的多重共现可视化分析工具及多重共现的知识发现方法也可用于分析论文中其它不同类型的特征项关联关系，可依据不同的研究目的来分析不同特征项之间的多重共现关系。但是由于多重共现交叉图技术的局限性，目前只能分析三个或以下特征项的共现关系，并且对于其它不同类型特征项共现关系的分析效果，仍有待进一步的分析验证。

参考文献：

[1] Morris S.A. Unified mathematical treatment of complex cascaded bipartite networks: the case of collections of journal papers [D]. Oklahoma:OklahomaStateUniversity，2005.

[2] Morris S.A.etc.DIVA: a visualization system for exploring document databases for technology forecasting[J]. Computers & Industrial Engineering,2002(43):841–862

[3] Morris S.A., Gary G. Yen. Crossmaps: Visualization of overlapping relationships in collections of journal papers[EB/OL].[2011-3-21]. www.pnas.org/cgi/doi/10.1073/pnas.030760410

[4] Loet Leydesdorff. What Can Heterogeneity Add to the Scientometric Map? Steps towards algorithmic historiography[EB\OL]. [2011-3-21]. http://arxiv.org/abs/1002.0532

[5] Thomson Data Analyzer[EB\OL].[2011-7-7].

http://science.thomsonreuters.com.cn/productsservices/TDA/

[6] 杨良斌,杨立英,乔忠华.基因组学领域的学术机构科研活动分析[J].图书与情报,2010(1):93-98

[7] 2011 中国大学排行榜揭晓，北京大学实现 4 连冠[EB/OL].[2011-3-21].http://www.cuaa.net/cur/2011/

[8] 图书馆、情报与文献学（20 种）[EB/OL].[2011-3-21]. http://www.cssci.com.cn/lyk2010/tq.htm

作者简介：

第一作者：

庞弘燊，1983/12 生，男，中国科学院国家科学图书馆博士生，研究方向：情报计量学

第二作者：

方曙，中国科学院国家科学图书馆成都分馆馆长，博士生导师，电子邮箱：winsunpang@126.com