基于多重共现的知识发现方法

庞弘燊 著

科学出版社

基于多重共现的知识发现方法

庞弘燊 著

北京师范大学图书馆
藏书

科学出版社
北京

内 容 简 介

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。但是，目前对共现现象的研究主要从两个特征项共现展开，本书基于多重共现的知识发现方法的研究致力于将三个或三个以上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、可视化技术、知识发现方法的基础上，拓展共现现象的研究范围。本书界定了多重共现的概念，构建了一套多重共现的基础理论体系，研究了可用于多重共现的可视化方式，设计并开发了三重共现的可视化分析工具，并进一步构建了基于多重共现的知识发现方法的分析体系，包括共现关联强度、被引关联强度、共现突发强度三个方面，最后通过实证研究验证了该套方法体系的分析效果及其可应用的研究范畴。

本书可供图书情报、数据挖掘、可视化分析、科技管理与评价等研究领域的科研人员与工作者参考。

图书在版编目(CIP)数据

基于多重共现的知识发现方法 / 庞弘燊著. —北京：科学出版社, 2017

ISBN 978-7-03-052943-5

Ⅰ. ①基… Ⅱ. ①庞… Ⅲ. ①知识工程 Ⅳ. ①TP182

中国版本图书馆 CIP 数据核字(2017)第 116093 号

责任编辑：裴 育 纪四稳 / 责任校对：桂伟利

责任印制：张 伟 / 封面设计：蓝 正

科学出版社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京教圖印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2017 年 6 月第 一 版 开本：720×1000 B5

2017 年 6 月第一次印刷 印张：9 插页：2

字数：181 000

定价：80.00 元

(如有印装质量问题，我社负责调换)

作 者 简 介

庞弘燊 1983 年生，中国科学院文献情报中心情报学博士，全国专利信息师资人才，兼职硕士生导师，现为深圳大学图书馆副研究馆员，曾工作于中国科学院广州生物医药与健康研究院信息情报中心。主要研究领域包括科学计量学、情报计量学、专利计量学以及科技政策管理研究、软科学研究等。主持科研项目 18 项，其中包括国家自然科学基金青年科学基金项目 1 项，中国科学院及省部级项目 10 项；组织撰写情报分析报告 20 余篇，其中专利与情报分析报告曾获国家部委采用及地方政府领导批示；在 SCI/SSCI/CSSCI 等收录的期刊中发表论文 40 余篇。

前　言

基于学科领域科技论文多重共现的知识发现方法是将各学科领域科技论文载体中的多特征项共现信息定量化、重点热点信息内容可视化的分析方法。目前国内外对特征项共现的研究方法以及工具软件多集中在两个特征项之间共现的研究，本书基于多重共现的知识发现方法的研究致力于将三个或三个以上特征项共现的现象作为研究主体，在总结现有相关的共现研究方法、数据挖掘技术、可视化技术、知识发现方法、情报计量分析方法的基础上，拓展共现现象的研究范围，以期设计出一套可应用于学科领域分析的多特征项共现情报计量分析方法，最后从应用研究的角度对所提出的理论方法进行验证。本书研究的情报计量分析方法在反映科学活动规律和科学知识领域方面可以增加多个分析角度和信息来源，并能为研究人员、科研管理部门多方位了解科学活动模式提供可靠依据。

本书的主要内容如下：

1) 共现相关理论的研究

对国内外共现相关领域的研究背景、发展现状与趋势、相关研究的理论与实践等进行研究；归纳目前不同共现研究的特点，并对不同特征项共现的知识发现与情报计量分析方法所能揭示的知识内容进行分析。

2) 多重共现基础理论体系的构建

构建起一套独特的多重共现基础理论体系，包括多重共现的定义及研究范畴、用于多重共现的变量符号、多重共现的矩阵定义、多重共现分析系数的计算公式、多重共现的数据组织形式等。

3) 多重共现可视化分析工具的研究与开发设计

通过对比不同共现可视化方式的特点，分析可用于多重共现的可视化效果，并针对多重共现知识发现分析过程的前期步骤(数据处理和多重共现可视化图生成)，设计和开发多重共现知识发现可视化分析工具(MOVT)。

4) 基于学科领域科技论文多重共现的知识发现方法的理论与实证研究

通过应用多重共现情报计量分析方法的理论，挖掘不同学科领域的研究热点、研究重点、主要科研机构、发文场所、重要科学家等，以及它们之间的关联关系，验证分析方法的理论可行性和实际分析效果。

此为试读，需要完整PDF请访问：www.ertongbook.com

本书内容主要源于作者自2009年起在中国科学院攻读博士学位期间的研究，以及自2012年参加工作至今持续对相关研究内容以及实证案例分析部分的补充和完善。书中部分研究内容在2015年获得国家自然科学基金项目(71403261)的资助，并在后续不断深入研究的过程中相继获得深圳大学人文社会科学青年教师扶持项目(17QNFC30)、ISTIC-THOMSON REUTERS科学计量学联合实验室开放基金项目、中国科学技术信息研究所情报工程实验室开放基金等的资助。书中部分研究成果已经发表在《情报学报》、《图书情报工作》等刊物上。

本书凝聚了众人的智慧和努力，包括作者的博士生导师方曙老师，以及中国科学院文献情报中心、中国科学院成都文献情报中心的老师和同学等都给予了相关的指导和帮助；同时，中国科学院广州生物医药与健康研究院和深圳大学图书馆的领导和同事也给予了相关的支持，在此向他们表示诚挚的谢意。感谢科学出版社对本书出版给予的大力支持和帮助。在撰写本书过程中，除调查检索分析所获取的相关数据，作者还参考了许多相关的中外文文献，但由于研究分析的数据量较大、参考文献较多，难免会有所疏漏，在此对所有文献作者表示衷心的感谢，同时也欢迎广大同行和读者就相关问题进行交流和讨论。

庞弘燊

2016年12月于深圳

目　录

前言
第 1 章　绪论 …… 1
1.1　研究背景 …… 1
1.2　国内外相关研究领域的发展现状与趋势 …… 2
1.2.1　知识发现的理论与实践研究 …… 2
1.2.2　共现的相关研究 …… 2
1.3　研究问题的提出 …… 9
1.4　研究目标与研究意义 …… 10
1.5　研究思路与框架 …… 11
1.6　研究方法 …… 12
1.7　本书组织结构 …… 13
第 2 章　多重共现的基础理论研究 …… 15
2.1　相关概念内涵 …… 15
2.1.1　共现的分析范畴 …… 15
2.1.2　多重共现的定义与研究范畴 …… 16
2.2　多重共现特征项的变量符号 …… 17
2.3　多重共现的矩阵定义与数据组织形式 …… 19
2.4　多重共现的延展系数 …… 23
2.5　小结 …… 25
第 3 章　多重共现的可视化方法研究 …… 26
3.1　可视化概念简介 …… 26
3.2　知识图谱的可视化软件工具简介 …… 28
3.3　可应用于多重共现的可视化方式研究 …… 33
3.3.1　多重共现的社会网络可视化方式 …… 33
3.3.2　多重共现的交叉图技术可视化方式 …… 38
3.3.3　多重共现的可视化方式对比研究 …… 43
3.4　小结 …… 43

第 4 章　多重共现知识发现方法的理论研究……45
4.1　知识发现的概念、模式及一般过程……45
4.1.1　数据、信息与知识的定义……45
4.1.2　知识发现的概念……46
4.1.3　知识发现的模式……47
4.1.4　知识发现的一般过程……48
4.2　多重共现的知识发现方法体系设计……51
4.2.1　共现关联强度的分析方法设计……52
4.2.2　被引关联强度的分析方法设计……59
4.2.3　共现突发强度的分析方法设计……65
4.2.4　多重共现的知识发现方法与一般共现分析效果对比……70
4.3　多重共现知识发现可视化分析工具的设计与开发……71
4.3.1　MOVT 数据处理及可视化绘图流程……72
4.3.2　MOVT 模块构成……73
4.3.3　MOVT 与 DIVA 对比……74
4.4　小结……74
第 5 章　三重共现知识发现方法的实证研究……76
5.1　共现关联强度的实证分析……76
5.1.1　研究领域的分析一……76
5.1.2　研究领域的分析二……82
5.1.3　研究机构的分析……86
5.1.4　机构间的对比分析……95
5.1.5　研究学者的分析……102
5.2　被引关联强度的实证分析……105
5.2.1　分析模型……105
5.2.2　数据来源……106
5.2.3　样例分析……107
5.2.4　分析效果……109
5.3　共现突发强度的实证分析……110
5.3.1　分析模型……110
5.3.2　数据来源……112
5.3.3　样例分析……112
5.3.4　分析效果……113

5.4　各方法联立的实证分析……114
5.5　小结……115
第 6 章　总结与展望……117
6.1　研究总结……117
6.1.1　相关理论的研究……117
6.1.2　多重共现基础理论体系的构建……117
6.1.3　多重共现的可视化方法研究……119
6.1.4　多重共现知识发现方法的理论研究……120
6.1.5　三重共现知识发现方法的实证研究……121
6.2　研究的创新之处……122
6.3　研究展望……122
参考文献……128

第1章 绪 论

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，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。

1.1 研究背景

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)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，受到研究者的关注并得到了大量理论探讨与应用研究。

在学术期刊上公开发表的论文，都有着比较严格的著录规范，包括正文以及一系列对论文相关信息进行描述的特征项。因此，期刊论文作为科学知识、科研成果的有形载体，除了直接反映成果的研究内容，还蕴藏着大量表征科学活动基本性状的信息。例如，题名是对论文主题的扼要表达；关键词是反映论文主题的学术词汇；摘要是对论文内容的高度概括；而其他特征项如作者、机构、引文、发表期刊、出版年份等则是对论文外部特征全方面、多角度的补充。

这些在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。

而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

部分可视化技术所能实现的，并在科学计量研究中被广泛应用，如聚类分析和多维尺度分析等。这些可视化技术揭示的是一种特征项之间的关系，如关键词共现、作者合作、文献共被引等，然而，这些关系揭示的信息存在一定的局限性。

众所周知，作为反映科学论文特征的不同特征项之间，也存在千丝万缕的联系。例如，施引论文与被引论文两种不同特征项之间的相互引用关系，作者与关键词之间的使用关系，论文与论文作者之间的隶属关系等。到目前为止，许多研究者已经对论文中特征项之间的联系进行了多方面的研究[2-25]。

1.2　国内外相关研究领域的发展现状与趋势

1.2.1　知识发现的理论与实践研究

知识发现又称数据库中的知识发现(knowledge discovery in database, KDD)，是指从大量数据集合中识别出有效的、新颖的、潜在有用的以及最终可理解模式的高级处理过程，数据挖掘(data mining)是知识发现过程中的一个主要步骤[26]。目前国际上知识发现的研究方向主要以知识发现的任务描述、知识评价与知识表示为主线，以有效的知识发现算法为中心，以知识发现模型为重点，研究知识发现自身的运行机制和内在机理及其在各领域中的实际应用。

知识发现的基本任务包括[27]数据分类、数据聚类、衰退和预报、关联和相关性、顺序发现、描述和辨别、时间序列分析等。知识发现的过程如图 1-1 所示，可以概括为三部曲，即数据准备(数据筛选、数据预处理、数据变换)、数据挖掘以及结果的解释评价。

目前国内外基于数据库本身的知识发现研究，除了借鉴数据库知识发现的理论，主要是运用聚类分析、神经网络方法、决策树方法、粗糙集技术、遗传算法、关联规则挖掘等数据挖掘算法分析发现数据集中数据之间的关联关系。

1.2.2　共现的相关研究

1. 共现的基本理论研究

目前国内外已有很多关于各种类型共现现象的研究，包括相同类型特征

此为试读，需要完整PDF请访问：www.ertongbook.com

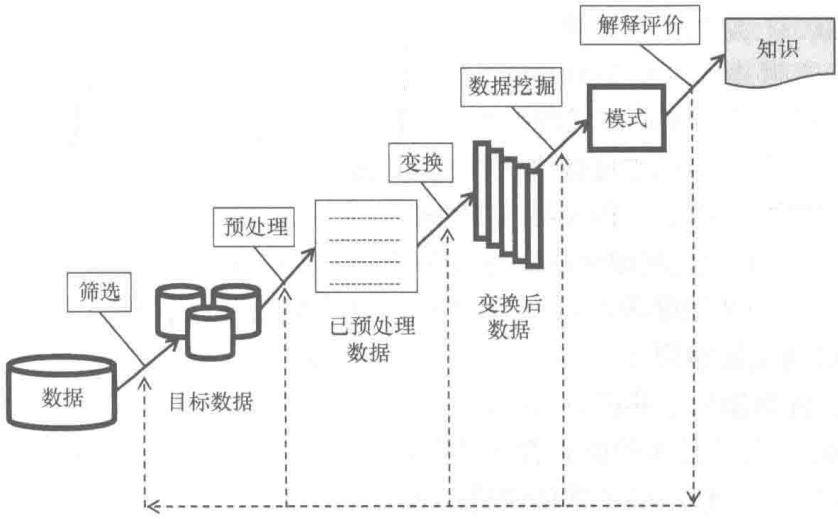

图 1-1　知识发现的过程

项的共现研究(co-occurrence)和不同类型特征项的共现研究(occurrence)。相同类型特征项的共现包括论文共现、关键词共现、作者共现等，其中研究最早、影响最大的是论文共现。不同类型特征项的共现包括作者文献耦合、作者关键词耦合等。以下是目前各种共现类型的研究现状。

1) 文献耦合(bibliographic coupling)

文献耦合是指文献通过参考文献进行的耦合。具体地，就是当两篇文献共同引用了一篇或多篇文献时，这两篇文献之间的关系就称为文献耦合。耦合的强度取决于共同参考文献(被引文献)的数量。Fano 在 1956 年注意到这种现象，首次提出文献耦合的概念和思路，但之后并没有引起人们广泛的关注[2]。1963 年，Kessler 在对 *Physical Review* 期刊进行研究时注意到越是学科专业内容相近的论文，它们参考文献中的相同文献的数量就越多，于是他把两篇(或多篇)同时引用一篇论文的论文称为耦合论文，并把它们之间的这种关系称为文献耦合，相同参考文献的数量即耦合强度[3]。两篇文献的耦合强度越高，说明这两篇文献之间的研究主题越相似。在美国科学情报研究所(ISI)的 Web of Science 数据库中，就是通过文献耦合为用户提供相关文献的信息。但是文献耦合的分析方法有一定的制约性，因为对于选定的论文数据，耦合关系不会随着时间流逝而发生改变，而是保持固定的，从这个意义上讲，耦合分析的结论是静态的。

2) 共被引(co-citation)

共被引是目前广受关注、研究成果最多的共现研究，包括文献共被引、

作者共被引、期刊共被引等。

(1) 文献共被引分析(document co-citation analysis, DCA)。文献共被引，又称文献共引，是指两篇文献同时被后来的其他文献所引用。具有共被引关系的文献之间借共被引强度体现彼此之间的关联度和内容的相似性，其实质是将一组具有共被引关系的文献作为分析对象，综合利用数学、统计学和逻辑分析方法，把对象之间错综复杂的共被引关系量化、抽象并简单表达的过程[4]。1973 年，美国情报学家 Small 在对“粒子物理学专业”进行知识结构描述时，发现两篇论文被相同文献引用的强度可以用来测度其内容相似程度，在此基础上创造性地提出了共被引的概念[5]。文献共被引分析方法经过 40 多年的演进，以 Small 等为代表的研究者从引文数据的选择、共被引矩阵的标准化处理，到不同层次及等级聚类方法的改进、可视化方法的引入等方面进行了大量的研究，使得文献共被引分析的理论和技术日臻完善。而近年来，利用共被引聚类来挖掘科学的热点领域、前沿领域以及发展领域正成为研究的焦点。

(2) 作者共被引分析(author co-citation analysis, ACA)。作者共被引，又称作者共引，是在文献共被引的基础上提出的，ACA 假定两位作者的文章若被后来的文献同时引用，那么表明这两位作者之间有联系，同时被引用的次数越多，就说明他们之间关系越紧密。ACA 起源于美国的 Drexel 大学，在 1981 年该校的 White 和 Griffith 合作发表的“Author co-citation: A literature measure of intellectual structure”一文开创了 ACA 的先河，文章对 39 位信息科学作者进行了共被引分析，划分出了情报学五大体系的核心作者，为之后的 ACA 研究提供了良好的范例[6]。1990 年，McCain[7]将 ACA 的分析步骤归纳为选择作者、检索共被引频次、生成共被引矩阵、转化为 Pearson 相关系数矩阵、多元统计分析、解释结果及效度分析六个步骤，人们将其称为 ACA 传统法模式或 Drexel 模式。现在，作者共被引分析已成为一种高效和多产的分析方法，不仅可以用来揭示科学结构的发展现状乃至变化情况，还可以用来进行前沿分析、领域分析、科研评价等。但是目前对于 ACA 的相关分析方法还有待优化，如在 Pearson 相关系数的适应性、对角线值的确定、矩阵的标准化等问题上还存有争议[8]。

(3) 期刊共被引分析(journal co-citation analysis, JCA)。期刊共被引，又称期刊共引，是指以期刊为基本单元而建立的共被引关系。具体来说，就是 n 种($n\geqslant 2$)期刊的论文被其他期刊同时引用，称这 n 种期刊具有共被引关系。其共被引程度以引用它们的期刊种数(或次数)来衡量，这个测度称为期刊共被引

强度或频次。期刊共被引分析把数量众多的期刊按被引证关系联系起来，从而从利用的角度揭示了各学科期刊之间的相互关系和结构特征[9]。1991 年，美国得克萨斯大学的 McCain 将文献共被引、作者共被引的思路和技术应用到期刊共现研究上，对经济学领域期刊进行共被引分析，以此为例考察在期刊水平上得出的聚合情况[10]；1998 年和 2000 年 McCain 和 Ding 等分别对神经网络领域、信息检索领域主流期刊进行了多维尺度(MDS)分析，考察了在不同时间段期刊的交流结构，通过透视期刊共被引结构来发现学科研究的变迁[11,12]。期刊共被引分析也可以用于挑选与评价期刊，运用社会网络分析方法如 MDS 分析来发现处于被引中心圈的期刊，即核心期刊[13]。

3) 共篇(co-text)

共篇分析属于论文共现研究。2002 年，中国学者崔雷和郑华川注意到论文之间基于相同关键词会产生关联，提出了“共篇”的概念，认为两篇论文共同出现相同关键词的数量越多，则两篇论文的内容相关性越强；并通过对胃癌前病变的研究现状和热点进行了探索，比较了共被引分析与共篇分析结果，发现两者的聚类分析的结论大致相同[14]。

4) 共词(co-word、co-term)

共词分析方法属于内容分析方法的一种，其原理主要是对一组词两两统计它们在同一篇文献中出现的次数，对这些词进行聚类分析，进而分析这些词所代表的学科和主题的结构变化。共词分析的思路最初是在 20 世纪 70 年代由法国文献计量学家提出的。1986 年，法国国家科研中心(CNRS)的 Callon 等出版了 *Mapping the Dynamics of Science and Technology* 一书，当时被称为“LEXIMAPPE”[15]。由于在结果分析方面关键词具有得天独厚的直接性，很快引起了研究者的高度关注。共词分析方法发展至今，主要经历了三个阶段，即第一代基于包容指数和临近指数的共词分析方法，第二代基于战略坐标的共词分析方法以及新一代基于数据库内容结构分析的共词分析方法[16]。

经过 30 多年的发展，共词分析方法从原理到使用都有了大幅度改进。利用共词分析方法基本原理可以概述研究领域的研究热点，横向和纵向分析领域学科的发展过程、特点以及领域或学科之间的关系，反映某个专业的科学研究水平及其发展历史的动态和静态结构，以及基础研究和技术研究之间的关系等。但是这种方法也存在着一些弊端，例如，共词分析对于词的选择非常敏感，作者取词的习惯、未经规范的关键词在表征论文内容的完整性等都会

造成结论的模糊、晦涩。此外，还有些研究对共词结论的可解释性提出质疑，认为其具有随意性较大和存在不确定性的缺陷，因而关于这一研究仍需不断地完善和改进[17,18]。

5) 作者合作(co-authorship)及国家、机构合作

在全球化趋势日益明显的今天，科研人员相互之间进行科研合作是非常普遍的现象，很多一流的研究成果需要通过不同科研人员的紧密协作完成。在文献计量研究中，作者共同署名即作者合作而在论文中产生不同作者姓名的共现，成为对合作研究定量测度的指标，类似地，机构名称共现、国家共现也是科学合作在不同层次和规模水平上的表现形式。科学计量学之父普赖斯是最早关注并对作者合作进行计量研究的学者，他在 20 世纪 60 年代初的研究表明，从 20 世纪开始，多作者合著论文一直呈快速增长的趋势[19]。美国的 Beaver 等在 1978 年对作者共现从社会学和历史学的角度进行了分析[20]。在近 40 年的研究中，学者们研究了不同国家、机构、学科之间、学科内部等作者合作的影响和规律[21]。此外，在研究机构合作方面，美国学者 Börner 等对机构合作数据处理技术进行了研究[22]。2006 年，杨立英等分析了化学领域国际 Top20 机构合作的规律，在分析化学领域机构合作规律的基础上，通过比较研究内容相似但合作较少的机构，提出了可以作为发现潜在合作机构的依据[23]。

6) 作者文献耦合分析(author bibliographic-coupling analysis, ABCA)

作者文献耦合分析是将文献耦合的方法扩展到作者层次，通过作者所有作品中参考文献的耦合强度来建立作者之间的关系。Zhao 等[24]在 2008 年提出了作者文献耦合分析，描述活跃作者自身的研究活动以便获得研究领域内更加真实的情况。作者文献耦合将文献耦合扩展到作者单元，可以反映当前的研究活动结构及其随时间的变化，不仅包括第一作者，而且也包括其他作者。通过对该方法进行实证研究，发现这种方法与作者共被引分析所揭示的知识结构是相互补充的，两种方法的结合可以更加全面地了解领域的知识结构。

7) 作者关键词耦合分析(author keywords coupling analysis, AKCA)

利用关键词的耦合强度分析作者之间关系的方法称为作者关键词耦合分析。同 ABCA 一样，从本质上讲，AKCA 也是一种耦合方法的扩展应用，而且这个名称也能很好地说明研究的层次(作者)以及所利用方法的本质(关键词耦合)。刘志辉等[25]在 2010 年提出作者关键词耦合分析方法，该方法利用作者作品的关键词耦合强度建立作者之间的关系，进而以这种关系生成的邻接矩阵

为基础进行分析。他们认为，AKCA 的首要问题是如何计算两位作者之间的相似度(*S*)，并将作者之间关系的强度(即相似度 *S*)定义为在所研究的时间段内，作者所发表论文的关键词的耦合强度，即两位作者拥有相同关键词的数量；并通过对科学计量学研究现状的分析对该方法进行了实证研究。作者关键词耦合网络与作者合著网络的二次指派程序(quadratic assignment procedure, QAP)分析表明，两种网络之间具有相关关系，但与后者相比，前者更能揭示出作者间潜在的关系。

8) 机构关键词共现

杨立英等在对化学领域机构的共现研究中，为了考察国际 Top20 机构在研究主题上的聚合情况，分析了机构与关键词共现现象，展示了机构在研究主题上的相似性与相异性[23]。

表 1-1 是对目前研究的各种共现现象的总结，对其所用到的矩阵数据、抽取特征项的信息以及分析内容等进行了归纳。

表 1-1　各类共现分析的特点

共现分析方法	所使用到的矩阵数据	需要从论文中抽取的信息	分析内容
文献耦合	论文-参考文献	参考文献	分析文献研究主题的相似性
文献共被引	论文-引证文献	引证文献	分析文献关联度和内容的相似性
作者共被引	作者-引证文献	作者、引证文献	分析作者研究领域的相似性
期刊共被引	期刊-引证期刊	期刊、引证期刊	揭示各学科期刊之间的相互关系和结构特征
共篇	论文-关键词	关键词	分析文献内容的相关性
共词	关键词-关键词	关键词	分析研究热点及其之间的相关性
作者合作	作者-作者	作者	分析作者合作情况
国家合作	国家-国家	国家	分析国家合作情况
机构合作	机构-机构	机构	分析机构合作情况
作者文献耦合	作者-参考文献	作者、参考文献	通过作者所有作品中参考文献的耦合强度来建立作者之间的关系
作者关键词耦合	作者-关键词	作者、关键词	利用关键词的耦合强度分析作者之间的关系
机构关键词共现	机构-关键词	机构、关键词	分析机构在研究主题上的相似性与相异性

2. 多特征项共现的相关研究

目前对特征项共现的研究多集中在两个特征项之间共现的研究，而对三个或三个以上特征项之间的共现关系研究并不多，此外，对多特征项共现的分析方法及可视化方式的研究也不多见。

1991 年，Braam、Moed 和 van Raan 把共被引和共词分析联系起来[28,29]，通过使用参考文献的聚类以及它们与关键词的关系来分析期刊论文的数据集。该方法首先基于共被引来聚类参考文献，进而形成基础参考文献组。在数据集中的论文根据它们所引用的基础参考文献组分派到指定的重叠组中。然后，由重叠组中的关键词组成的词组就可以基于论文组中关键词的频次进一步形成。这使得关键词可以与基础参考文献的聚类关联起来，并且有助于标识基础参考文献集合和搜索论文集合。

为了揭示两种特征项之间的关联，美国科学计量专家 Morris 等[30,31]借助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，这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内部与外部关联的缺陷。

胡琼芳和曾建勋[32]从文献共被引、耦合、共篇三个维度出发，提出并实现了一种综合三个特征项的文献相关度判定方法。其研究相当于使用论文中的参考文献-引证文献-关键词三个共同出现的特征项进行匹配对比，挖掘各论文之间的相关性。该方法认为经局部相似度计算，任何一篇文献及其所有相关文献都可以表示成三维的相似空间向量形式。每篇文献均可看成三维空间中的一点。两文献的相关度可近似由点与点之间的距离表示。采用式(1-1)的夹角余弦公式来计算点之间的距离。距离越近，两向量夹角越小，相应余弦值越大，相关度越大。

$$\text{Relevence}(A, B) = \cos(X, Y) \frac{\sum_{i=1}^{m} (x_i \cdot y_i)}{\sqrt{\left(\sum_{i=1}^{m} x_i^2\right)\left(\sum_{i=1}^{m} y_i^2\right)}} \tag{1-1}$$

其中，A 与 B 为两篇文献；X、Y 分别为 A、B 的相似空间向量；x_i 与 y_i 为分量值；m 是向量维数，在综合三个特征项的文献相关度判定研究当中 m 为 3。若将 A 看成源文献，经归一化处理后，A 的相似空间向量则变为 $X(1,1,1)$，相关文献 B 的向量表示为 $Y(y_1, y_2, y_3)$，其中 y_1、y_2、y_3 分别为 B 与 A 基于共被引、耦合和

此为试读，需要完整PDF请访问：www.ertongbook.com