利用引文内容进行主题级学科交叉类型分析*

徐庶睿[1,2] 章成志[1,2,3] 卢超[1,2]
[1] 南京理工大学信息管理系 南京 210094 [2]江苏省社会公共安全科技协同创新中心 南京 210094
[3] 江苏省数据工程与知识服务重点实验室(南京大学) 南京 210093

摘要: [目的/意义] 针对学科交叉宏观研究不能刻画学科交叉主题,以及学科交叉微观研究仍处于主题挖掘研究阶段的现状,从内容层面解决主题学科交叉度计算问题,并构建学科交叉分类的量化标准。[方法/过程] 首先,采集学术论文并解析引文内容;利用术语集获取术语和术语主题。然后,统计引文内容中的主题术语重复率。接着,计算学科间的主题学科交叉度。最后,基于主题学科交叉度分布熵,进行分类并分析。[结果/结论] 研究结果表明:①六个学科难以与医学在实践应用知识层面进行学科交叉;医学的理论基础与六个学科有明显的学科知识交叉。②学科交叉存在三种类型分别为:界内交叉、工具型交叉和界外交叉。综上,通过引文内容中的术语可以有效地计算主题学科交叉度,定量地研究学科交叉类型。

关键词: 主题 学科交叉类型 学科交叉性 引文内容 术语
分类号: G250
DOI: 10.13266/j.issn.0252-3116.2017.23.002

1 引言

学科交叉是创新的源泉。最负盛名的实例当属DNA双螺旋结构的发现。它的发现,是多学科知识交叉渗透、相互借鉴的结果[1]。沃森是生物学家,懂得基因的重要性。克里克是物理学家,熟悉X射线衍射晶体学理论。他们两人的知识范围正好可以互补。化学家富兰克林和物理学家威尔金斯积累的X射线衍射相关研究也为构建DNA分子模型做出了重要的贡献。

由此不难看出:面对类似的复杂问题,仅仅依靠某一学科、某一领域的知识是远远不够的。整合多个学科的理论与技术,有利于整合多方面的学科视角,为攻克难题提供更加坚实的基础。因此,学科交叉合作日益频繁,成绩斐然。这种学科合作研究势必会带来学科理论的碰撞和知识的交流,研究问题因而呈现出一定的主题属性。这种学科交叉主题的分布,反映了学科交流的深度和广度。因此,研究学科交叉主题,对深入认识学科间的交叉行为和现象、了解学科知识结构的形成和发展、以及新兴学科的产生,具有重要的研究意义。

然而,现有的学科交叉研究更多地关注于宏观层面的学科交叉态势研究,而微观层面的学科交叉主题研究还较少[2]。宏观学科交叉态势研究并不能具体地呈现学科交叉行为的内情况,即主题的交叉。"当两个学科发生交叉时,在哪些知识主题下发生了交叉,在哪些知识主题下没有交叉?"以及"当两个学科发生交叉时,在某些知识主题下的交叉程度是否相同?"等问题仅仅通过学科交叉态势研究是无法很好地予以回答的,需要深入微观主题层面探寻答案。不仅如此,研究学科交叉主题可以有效地探测研究热点和学科新的生长点,挖掘未来的研究发展动向和机会。但是目前对学科交叉主题的研究还处于主题挖掘阶段。研究者多通过获取文本特征词(限于论文标题等论文元数据)形成学科热点主题网络[2]。虽然元数据获取便利,但由于元数据是对论文全文内容的高度概括,无法直观展现学科间具体的交叉点和交叉主题。此外,目前的学科交叉类型研究主要以定性分析为主,缺少量化指标[3]。

H. Small认为当研究者倾向于使用相同的词或短语来描述同一个被引用对象时,这些词或短语就成为

* 本文系国家社会科学基金项目"面向知识创新服务的数据科学理论与方法研究"(项目编号: 16ZAD224)研究成果之一。
作者简介: 徐庶睿,硕士研究生;章成志(ORCID: 0000-0001-8121-4796),教授,博士,博士生导师,通讯作者,E-mail: zhangcz@njust.edu.cn;卢超,博士研究生。
收稿日期: 2017-06-16 修回日期: 2017-09-09 本文起止页码: 15-24 本文责任编辑: 王善军

这个被引对象的概念符号[4]。学术界进行交流的概念符号是术语[5]。术语具有稳定性、单义性和准确性等特征[6]。引文内容涵盖施引文献和被引文献的提及知识,是研究知识交流的平台[7]。引文内容中的术语,是引用中传递的信息内核[8],是施引文献和被引文献沟通的媒介。把术语和引文内容作为学科交叉问题分析的载体是有效的,可从内容微观层面具象化学科间的知识交叉。章成志等利用引文内容的术语探测学科交叉度,验证结合术语和引文内容的方法在学科交叉研究中的重要价值[9]。

综上所述,本文结合术语和引文内容进行主题级的学科交叉类型分析,从内容分析角度量化回答不同主题下的学科交叉度和学科交叉分类这两个问题。本文的"主题级"是指术语的主题,使用的术语集涵盖术语主题,从术语主题角度对学科交叉的知识集成现象进行分析。本文通过统计学科论文在引文内容中引用不同主题术语的重复率,从而确定不同主题下学科的交叉情况,并根据主题学科交叉度的分布进行交叉学科分类。

2 文献综述

本文从学科交叉相关研究和引文内容相关研究两个角度进行文献综述。

2.1 学科交叉相关研究

学科交叉目前还是一个模糊的术语概念,其定义主要涉及两个方面的问题:学科构成的界定以及学科间相互交叉的形式[10]。在相关的研究中,"学科交叉"经常与描述学科间不同作用形式的其他概念(例如:"多学科(multidisciplinary)""跨学科(transdisciplinary 或者 cross-disciplinary)")进行比较讨论[11-12]。

目前的学科交叉研究主要分为宏观学科交叉态势研究和微观学科交叉主题研究两个方面[2]。

(1)宏观学科交叉态势研究。截至目前,国内外对学科交叉态势的研究正在蓬勃发展。A. Stirling 和 A. Purvis 等指出学科多样性测度的三个维度分别是学科分类的数量(Variety),学科分类的平均分布程度(Balance)以及学科分类间的差异性(Disparity)[13-15]。如何将三个维度聚合成单一的指标成为实现学科交叉测度的关键。

A. Stirling 对传统多样性指标(如:信息熵指标)进行了总结,提出结合了学科分类数量(Variety)和学科分类平均分布程度(Balance)的指标[14],但是没有考虑学科之间的相似性。A. L. Porter 等提出 Integration score(整合度)作为学科交叉的综合性测度指标[16]。该指标不仅度量被引参考文献在不同学科的分布,还度量了学科间的相关性。但是该指标依赖于预先定义的学科分类,没有考虑学科中的动态变化[17]。在 A. L. Porter 等工作的基础上,I. Rafols 等引入基于网络分析的 Coherence(学科聚合性)指标[17]。通过评价参考文献多样性来对交叉学科性进行间接测定,加强对学科一致性的测度。目前宏观学科交叉态势的研究主要借鉴生物多样性指标进行测度,但各种测度指标缺少有效的对比分析,没有统一的测度标准以及测度数据,而且由于使用论文的引证关系为研究对象,学科间的交叉点,交叉主题,知识结构都无法具体呈现。

(2)微观学科交叉主题研究。目前交叉学科微观主题还处于挖掘研究阶段,少数研究者对此进行研究。当前多使用网络分析的方法进行研究。通过获取文本特征词(限于论文标题和摘要)形成学科热点主题网络[2]。

P. Vugteveen 等从引证关系的角度对交叉学科——河流学进行了研究[18]。作者使用网络分析的方法首次将学科主题和主要来源的学科领域进行了初步的关联,但该研究未考虑学科主题的学科交叉程度。X. Haiyun 等从内容分析的角度,将主题术语学科交叉性测量 TI 指标用于交叉学科的主题挖掘,对主题术语的学科交叉度进行比较,并通过实例进行论证[19]。TI 指标对学科交叉程度进行度量,但其主要借鉴特征词选择公式 TF-IDF,未考虑学科交叉相关测度指标。

2.2 引文内容相关研究

随着数据挖掘、自然语言处理等技术的发展,基于引文内容的引文分析研究逐渐得到研究者的重视。在文献中,引文内容是围绕在引文标记附近的文本内容[20]。

引文内容是对被引用文献主题最好的反映[4]。越来越多的学者聚焦引文内容,并论证引文内容相较于摘要、引证关系等的优势。祝清松等以高被引论文为研究对象,结果显示基于引文内容分析的核心主题能够较好地揭示高被引论文的被引原因(引用动机),而且与论文的研究内容相符合[21]。L. Shengbo 等比较了引文内容的主题和施引文献摘要的主题,结果显示摘要和引文内容有主题上的重合,但源自摘要的主题更加宽泛,来自引文内容的主题更加明确[22]。

综上所述,学科交叉研究在宏观态势研究上以 A. Stirling 的学科多样性分析框架[14]为基础,吸引了众多学者的关注和深入研究。但由于研究方法和研究数据

的局限性，该研究不能对具体的学科交叉细节进行研究，需要深入微观主题层面寻找方法。然而学科交叉微观主题的研究还处于起步探测阶段，对主题探测的范围仅限于论文的元数据层面，尚未深入全文内容层面进行主题分析。此外，当前并没有关于学科交叉分类的固定形式，研究者根据自身对学科交叉特征的认识给出了多种描述性分类方法。K. Huutoniemi 等在总结大量学科交叉分类研究的基础上，从定性分析的角度，给出学科交叉分类的可操作化定义[3]，仍缺乏量化指标。

目前，随着全文数据库的开放，深入文本内容进行知识挖掘的条件逐渐成熟。因此，本文从微观主题层面，结合术语和引文内容，定量地对学科交叉进行分类。徐庶睿等提出从内容层面计算总体学科交叉度的方法[23]，本文在此基础上，进一步从主题层面测量学科交叉度。

3 研究思路与实现方法

3.1 研究思路

学科间的交叉，是通过带有明确研究问题的科学研究产生的。这些研究反映出交叉的学科间各自具备的主题或者共有的主题。如图 1 所示，学科 A 和学科 B 有共同的研究主题，这些研究主题由具体的学科交叉点，即学科术语构成。本文利用术语和引文内容，计算不同主题下的学科交叉度，并根据主题学科交叉度的主题分布熵进行分类并分析。

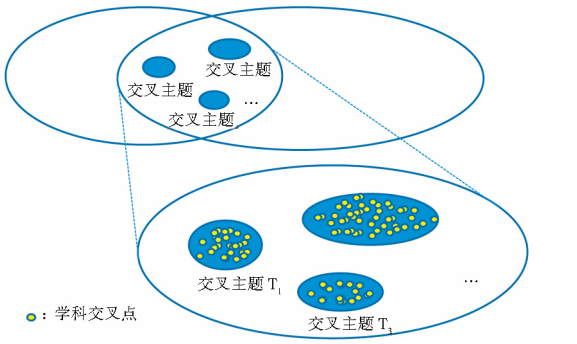

图 1 微观层面的学科交叉主题研究示意图

本文设计研究框架图如图 2 所示。首先，将全文数据中的引文内容和学科信息解析出来，本文仅将引文内容的范围限定为包含引用标记的引用句，以及引用句的前两句和引用句的后两句，最多共为五句话[24]。同时，获取术语集中的术语和术语主题。术语主题下的术语存在部分重合。其次，根据最大正向匹配的方法查找学科引文内容中的主题术语，按照主题布局统计术语重复率。接着，根据学科交叉度公式计算学科间的主题学科交叉度。然后，根据学科交叉的主题分布熵公式进行聚类分析。最后，进行不同类型下的主题分析。

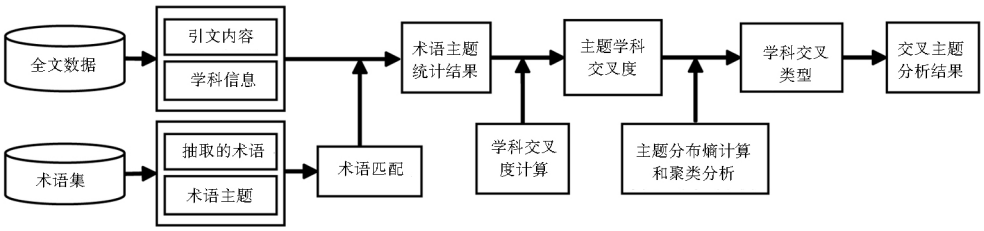

图 2 主题级学科交叉类型分析框架图

3.2 实现方法

3.2.1 学术论文引文内容抽取 随着开放获取运动的兴起，以 XML 格式为代表的结构化全文数据开始出现。世界著名科学期刊发行商 Springer 和 PLOS 等都提供或部分提供 XML 格式的全文阅读或下载。利用文本挖掘和自然语言处理技术，对全文信息中的引文内容信息进行抽取和保存。

3.2.2 术语和术语主题获取 由于语言资源的不断丰富与自然语言处理技术的不断发展，术语和术语主题的获取主要有两种途径：基于机器的术语和术语主题自动抽取、基于术语词典的术语和术语主题获取[23]。基于机器的术语和术语主题自动抽取方法主要分为：基于语言学的方法、基于统计的方法以及混合方法[23]。术语词典主要包括《美国国会图书馆标题表》《医学主题词表》等。利用文本挖掘和自然语言处理技术，对术语词典中的术语和术语主题进行抽取和保存。

3.2.3 学科交叉度计算 学科交叉度的计算方法有很多，如余弦相似度、杰卡德相似系数等。本文根据余弦相似度计算学科间引用术语的相似度，将学科引用术语的相似度从内容分析角度量化表示为学科交叉度[23]。设学科 D_1 和 D_2 表示向量空间模型中的两个向量：

$$D_1 = D_1(w_{11}, w_{12}, \cdots, w_{1n})$$

$$D_2 = D_2(w_{21}, w_{22}, \cdots, w_{2n})$$

于是，可以利用 n 维空间中两个向量之间的距离，即两个向量夹角的余弦值来表示学科间引用术语的相似系数。计算方法如公式 1 所示。

$$Sim(D_1, D_2) = cos\theta = \frac{\sum_{k=1}^{n} w_{1k} \times w_{2k}}{\sqrt{\sum_{k=1}^{n} w_{1k}^2 \sum_{k=1}^{n} w_{2k}^2}} \quad (1)$$

本文将学科交叉度分为总体学科交叉度和主题学科交叉度。其中，总体学科交叉度是指某两个学科间基于术语库中的全部术语计算得到的学科交叉度。主题学科交叉度是指某两个学科间基于某一主题下的全部术语计算得到的学科交叉度。

3.2.4 学科交叉的主题分布熵计算 本文通过统计学科交叉的主题分布熵分析学科交叉在主题下的不同分布。学科交叉主题分布的熵值计算公式如公式 2：

$$ITE(I_n) = \sum_j P_{ij} \log_2 P_{ij} \quad (2)$$

其中，I_n 为第 n 对学科交叉，$ITE(I_n)$ 为交叉学科 I_n 的主题分布熵，P_{ij} 为学科交叉 d_i 在主题 j 上出现的概率，通过公式 3 计算得到。

$$P_{ij} = \frac{Sim_{ij}}{\sum_j Sim_{ij}} \quad (3)$$

其中，Sim_{ij} 为学科交叉 d_i 在主题 j 上的学科交叉度。

3.2.5 聚类分析 常见的聚类算法分为基于划分的聚类、基于层次的聚类、基于网络的聚类、基于密度的聚类和基于模型的聚类等[26]。本文使用基于划分的 *AP* 聚类算法对统计结果进行聚类分析。

AP 聚类算法[27]是 2007 年由 *B. J. Frey* 等提出的一种聚类算法。该算法无需事先定义类数，在迭代过程中不断搜索合适的聚类中心，自动从数据点间识别类中心的位置及个数，使所有的数据点到最近的类代表点的相似度之和最大。

4 结果分析

4.1 实证数据分析

选取医学术语集为本文的实证术语集，选取 *PLOS ONE* 和《医学主题词表》为实证数据来源。

4.1.1 学术论文引文内容抽取 以 *PLOS ONE* 为实证数据来源。*PLOS ONE*（http://journals.plos.org/plosone/）是目前学术界非常有影响力的开放存取期刊，载文涉及从自然科学到社会科学等 10 多种学科。该期刊对所发表论文提供 XML 格式的全文下载，是本文理想的数据来源。本文以“生物学（Biology，BI）”“化学（Chemistry，CH）”“计算机科学（Computer Science，CS）”“数学（Mathematics，MA）”“物理学（Physics，PH）”和“心理学（Psychology，PS）”这 6 个学科领域为样本，选取 2006 年 12 月 20 日到 2014 年 12 月 18 日这段时间内所发表的论文。为保持各学科间样本的均衡性，论文数高于 300 篇的学科均随机采样 300 篇。总共获取 1 725 篇论文[23]。

对全文数据进行解析，抽取出每条参考文献的基本信息以及参考文献在全文中所对应的引用内容，从 1 725 篇源文献中共获得 53 869 条引文内容[23]，如表 1 所示。其中，本文共抽取了 214 187 个句子，引文内容的平均句子数约为 4 句。

表 1 学科文献和引文内容数

学科	论文数	引用内容	句子总数	平均句子数
生物学（BI）	300	9 927	40 076	4.04
化学（CH）	300	10 347	41 923	4.05
计算机科学（CS）	243	7657	29 401	3.84
数学（MA）	300	6 342	24 647	3.89
物理学（PH）	282	9 604	38 501	4.01
心理学（PS）	300	9 992	39 639	3.97
总量统计	1 725	53 869	214 187	3.98

4.1.2 术语和术语主题获取 通过解析《医学主题词表》获取医学术语和术语主题。《医学主题词表》（Medical Subject Headings，简称 MeSH，https://www.nlm.nih.gov/mesh/）是美国国立医学图书馆编制的权威性主题词表，应用广泛[28]。MeSH 由主题词变更表、字顺表和树状结构表组成。字顺表是医学主题词表的主表。树状结构表将字顺表中互不联系的主题词组成树状等级结构。

本文从 2017 版本的 MeSH 中共获取医学术语共 142 968 个[23]，分布在 115 个主题下，主题名称如表 2 所示，其中词语量表示医学主题下的医学术语数量。

4.2 学科引用主题统计分析

对 6 个学科引文内容引用不同主题术语的重复率进行统计，使用 HemI 软件（Heatmap Illustrator，version 1.0）绘制热谱图，结果见图 3。

首先，生物学、化学和物理学与医学在 A01（身体各部位）至 A19（真菌结构），以及 B05（生物形态）这两个主题范围下有较多的学科知识交叉。主要原因可能是由于生物学、化学和物理学与医学，在学科研究对象上的重叠，导致交叉密集。生物学、化学和物理学与医学，还在 G01（物理现象）至 G07（生理现象）、G15（植物生理现象）至 H01（自然科学学科）、V01（出版组件）至 Z01（地理位置）这些主题范围下有学科知识交叉。

表 2　医学主题的中英文对照和主题词语量

主题序号	英文名	中文名	词语量
A01	Body Regions	身体各部分	296
A02	Musculoskeletal System	肌肉骨骼系统	929
A03	Digestive System	消化系统	285
A04	Respiratory System	呼吸系统	201
A05	Urogenital System	泌尿系统	274
A06	Endocrine System	内分泌系统	327
A07	Cardiovascular System	心血管系统	421
A08	Nervous System	神经系统	2519
A09	Sense Organs	感觉器官	427
A10	Tissues	组织	679
A11	Cells	细胞	2499
A12	Fluids and Secretions	体液和分泌物	191
A13	Animal Structures	动物结构	203
A14	Stomatognathic System	口颌系统	291
A15	Hemic and Immune Systems	血液和免疫系统	550
A16	Embryonic Structures	胚胎结构	257
A17	Integumentary System	皮肤系统	40
A18	Plant Structures	植物结构	179
A19	Fungal Structures	真菌结构	30
A20	Bacterial Structures	细菌结构	29
B01	Eukaryota	真核生物	11 429
B02	Archaea	古生菌	112
B03	Bacteria	细菌	1 275
B04	Viruses	病毒	2 362
B05	Organism Forms	生物形态	113
C01	Bacterial Infections and Mycoses	细菌感染和真菌病	1 506
C02	Virus Diseases	病毒疾病	944
C03	Parasitic Diseases	寄生虫病	538
C04	Neoplasms	肿瘤	4 195
C05	Musculoskeletal Diseases	肌肉骨骼系统疾病	2 450
C06	Digestive System Diseases	消化系统疾病	1 372
C07	Stomatognathic Diseases	口颌疾病	1 065
C08	Respiratory Tract Diseases	呼吸道疾病	1 091
C09	Otorhinolaryngologic Diseases	耳鼻喉疾病	744
C10	Nervous System Diseases	神经系统疾病	9 770
C11	Eye Diseases	眼疾病	1 458
C12	Male Urogenital Diseases	男性生殖器疾病	1 130
C13	Female Urogenital Diseases and Pregnancy Complications	女性生殖器疾病和妊娠并发症	1 611
C14	Cardiovascular Diseases	心血管系统疾病	2 372
C15	Hemic and Lymphatic Diseases	血液和淋巴系统疾病	1 873
C16	Congenital，Hereditary，and Neonatal Diseases and Abnormalities	先天性、遗传性和新生儿疾病和畸形	6 412
C17	Skin and Connective Tissue Diseases	皮肤和结缔组织疾病	2 222
C18	Nutritional and Metabolic Diseases	营养和代谢性疾病	3 010
C19	Endocrine System Diseases	内分泌系统疾病	1 042
C20	Immune System Diseases	免疫系统疾病	1 743
C21	Disorders of Environmental Origin	环境因素诱发疾病	12
C22	Animal Diseases	动物疾病	432
C23	Pathological Conditions，Signs and Symptoms	病理条件、体征和症状	4 673
C24	Occupational Diseases	职业病	105
C25	Chemically-Induced Disorders	化学诱导的疾病	564
C26	Wounds and Injuries	创伤与损伤	1 331
D01	Inorganic Chemicals	无机化合物	1 059
D02	Organic Chemicals	有机化合物	12 529
D03	Heterocyclic Compounds	杂环化合物	10 346
D04	Polycyclic Compounds	多环碳氢化合物	3 507
D05	Macromolecular Substances	大分子物质	1 628
D06	Hormones，Hormone Substitutes，and Hormone Antagonists	激素、激素代用品和激素拮抗剂	1 555
D08	Enzymes and Coenzymes	酶与辅酶	11 368
D09	Carbohydrates	碳水化合物	1 816
D10	Lipids	脂类	1 590

（续表 2）

主题序号	英文名	中文名	词语量
D12	Amino Acids，Peptides，and Proteins	氨基酸、肽、蛋白质	23 657
D13	Nucleic Acids，Nucleotides，and Nucleosides	核酸、核苷类和核苷酸类	1 443
D20	Complex Mixtures	复杂混合物	837
D23	Biological Factors	生物因素	5 601
D25	Biomedical and Dental Materials	生物医学和牙科科学	789
D26	Pharmaceutical Preparations	药物制剂	423
D27	Chemical Actions and Uses	化学反应和用途	2 735
E01	Diagnosis	诊断	3 427
E02	Therapeutics	治疗	2 822
E03	Anesthesia and Analgesia	麻醉和镇痛	130
E04	Surgical Procedures，Operative	外科操作，手术	2 415
E05	Investigative Techniques	包埋技术	5 123
E06	Dentistry	牙科	684
E07	Equipment and Supplies	设备和供应	1 684
F01	Behavior and Behavior Mechanisms	行为和行为机制	1 982
F02	Psychological Phenomena and Processes	心理现象和过程	990
F03	Mental Disorders	精神疾病	1 294
F04	Behavioral Disciplines and Activities	行为训练和活动	652
G01	Physical Phenomena	物理现象	833
G02	Chemical Phenomena	化学现象	1 799
G03	Metabolism	代谢	649
G04	Cell Physiological Phenomena	细胞生理现象	704
G05	Genetic Phenomena	遗传现象	2 764
G06	Microbiological Phenomena	微生物学现象	238
G07	Physiological Phenomena	生理现象	1 806
G08	Reproductive and Urinary Physiological Phenomena	生殖和泌尿生理现象	559
G09	Circulatory and Respiratory Physiological Phenomena	循环和呼吸生理现象	628
G10	Digestive System and Oral Physiological Phenomena	消化系统和口腔生理现象	128
G11	Musculoskeletal and Neural Physiological Phenomena	肌肉骨骼和神经生理现象	849
G12	Immune System Phenomena	免疫系统现象	436
G13	Integumentary System Physiological Phenomena	外皮系统的生理现象	52
G14	Ocular Physiological Phenomena	眼生理现象	110
G15	Plant Physiological Phenomena	植物生理现象	97
G16	Biological Phenomena	生物学现象	650
H01	Natural Science Disciplines	自然科学学科	621
H02	Health Occupations	卫生职业	587
I01	Social Sciences	社会科学	1 931
I02	Education	教育	296
I03	Human Activities	人类活动	339
J01	Technology，Industry，and Agriculture	工艺学、工业和农业	1 984
J02	Food and Beverages	食物和饮料	408
J03	Non-Medical Public and Private Facilities	非医疗公共和私人设施	169
K01	Humanities	人文科学	679
L01	Information Science	情报科学	1 906
M01	Persons	人群	1 315
N01	Population Characteristics	人口特征	634
N02	Health Care Facilities，Manpower，and Services	卫生保健设施，人力和服务	2 049
N03	Health Care Economics and Organizations	卫生保健经济和组织	1 709
N04	Health Services Administration	卫生服务行政管理	1 442
N05	Health Care Quality，Access，and Evaluation	卫生保健质量，实施，评估	1 693
N06	Environment and Public Health	环境与公共卫生	2 324
V01	Publication Components	出版组件	60
V02	Publication Formats	出版格式	184
V03	Study Characteristics	研究特征	25
Z01	Geographic Locations	地理位置	742

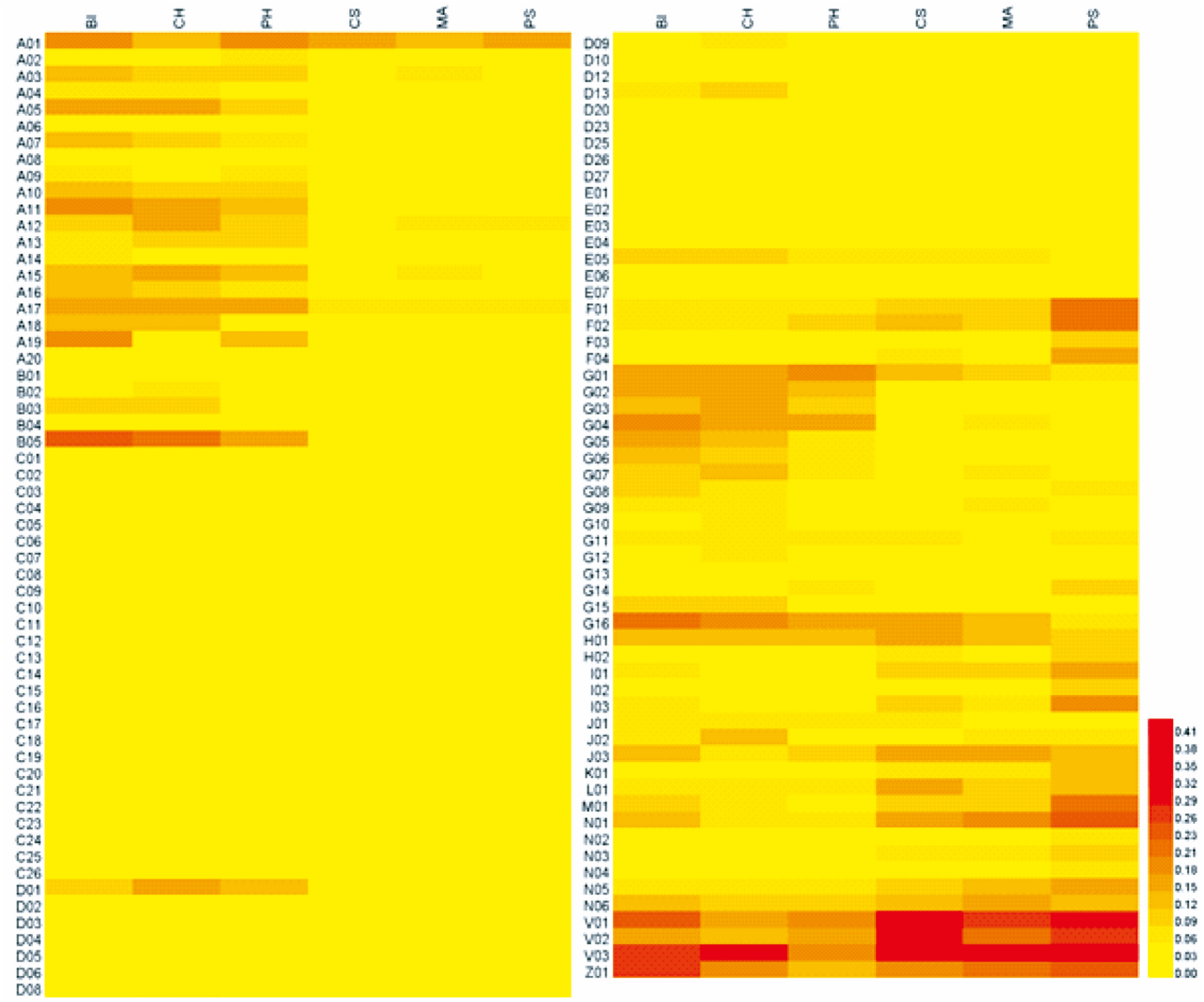

图 3 医学术语引用主题分布图

其次，计算机科学和数学与医学的学科知识交叉结果，主要分布于主题分类的后半部分（图 3 右），主要为 V01（出版组件）至 V03（研究特征）。主要原因可能是计算机科学，数学和医学的学科差异大，学科交叉较多地集中在补充材料（supplementary material），数据集（dataset），统计（statistics）等词语上。

最后，心理学和医学的学科知识交叉结果，也主要分布于主题分类的下半部分（图 3 右），主要范围为 F01（行为和行为机制）至 F04（行为训练和活动），以及 M01（人群）、N01（人口特征）、V01（出版组件）至 V03（研究特征）。F01（行为和行为机制）至 F04（行为训练和活动）这个主题范围会突显的主要原因可能是当前社会的心理疾病，其医学问题、心理问题都密切结合在一起。心理学和医学在发展中相互交叉，产生了如心理医学、医学心理学等新的交叉学科。

图 3 中的 C01（细菌感染和真菌病）至 C26（创伤与损伤）、D01（无机化合物）至 D27（化学反应和用途）、E01（诊断）至 E07（设备和供应）中的医学术语被本文所选的六个学科引用少。可能原因是这些知识具有医学的自身特色，属于医学在实践过程中的具体应用，六个学科与医学在这些方面的知识交叉少。

4.3 基于主题学科交叉度分布的学科交叉类型分析

本文根据公式（1），计算总体学科交叉度。本文根据公式（1）（2）和（3），计算学科交叉的主题分布熵。对 15 组学科交叉使用 AP 聚类的方法进行聚类分析，结果如图 4 所示。

根据 AP 聚类的结果，将 15 组学科交叉分为 3 类。根据 3 个中心点（生物学和物理学的交叉、数学和物理学的交叉以及生物和心理学的交叉）的涵盖范围对聚类结果进行人工调整，将计算机科学和数学的交叉以及计算机科学和心理学的交叉视为异常点，将化学和计算机科学的交叉归入以数学和物理学的交叉为中心

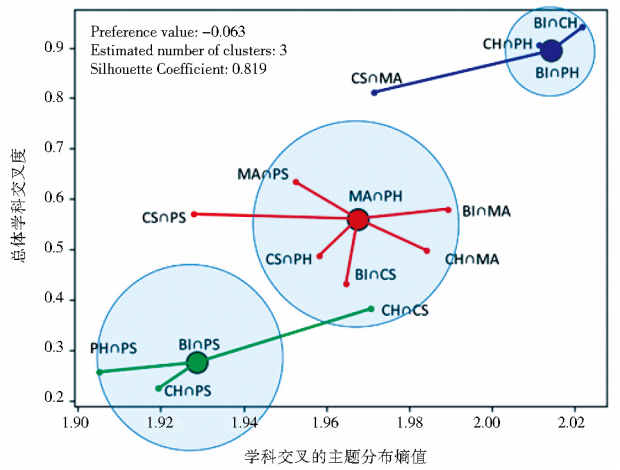

图4　基于AP聚类的15组学科交叉聚类结果

的分类中。

第一类学科交叉主要分布在图4的右上角，以生物学和物理学的交叉为中心，主要包括化学和物理学的交叉，生物学和化学的交叉。这三组学科交叉，不仅学科间的交叉程度高，而且主题下学科交叉度的熵值高，即在主题下学科交叉度分布均匀。首先，学科交叉程度高。物理学、化学、生物学是现代自然科学的基础性学科[29]，许多新兴学科源于这三门学科的融合，如物理化学、化学物理学等。其次，主题学科交叉度分布均匀。说明物理学、化学和生物学之间容易产生交叉，并且交叉面广，在许多主题下都有较高程度的学科交叉。由于物理学，化学，生物学都属于自然科学，所以将这类学科交叉视为“界内交叉”。

第二类学科交叉主要分布在图4的中间区域，以数学和物理学的交叉为中心，主要包括数学和心理学的交叉，计算机科学和物理学的交叉，生物学和计算机科学的交叉，化学和数学的交叉，生物学和数学的交叉以及化学和计算机科学的交叉，共有七组学科交叉。这七组学科交叉的学科交叉度和主题学科交叉度分布都适中，且都是和计算机科学或者数学产生的学科交叉。这类数据说明计算机科学和数学的特殊地位。计算机科学和数学在学科交叉过程中的基础辅助作用十分突出，所以将这类学科交叉视为“工具型交叉”。

第三类学科交叉主要分布在图4的左下角，以生物学和心理学的交叉为中心，主要包括物理学和心理学的交叉，化学和心理学的交叉。这三组学科交叉，不仅学科间的交叉程度低，而且主题下学科交叉度的熵值也低，即在主题下的学科交叉度分布集中。心理学是一门研究人类的心理现象、精神功能和行为的科学。心理学具有社会科学性，与以物理学、化学、生物学为代表的自然学科，难以形成学科交叉，并且难以形成广泛的学科交叉。因为研究对象已经涉及自然科学类和社会科学类，所以将这类学科交叉视为“跨界交叉”。

此外，数学和计算机科学的学科交叉属于异常点。主题下学科交叉度的熵值适中，但是学科交叉度较突出，高于所有和计算机科学或者数学交叉的学科交叉组合。可能原因是由于计算机科学和数学的发展是相辅相成，密不可分的。计算机科学的各种程序在应用数学的思想和算法[30]。同时，数学的应用层面要依靠计算机科学技术实现，例如计算机是数学建模的一个重要工具，计算机可以模拟出建模所需的“理想状态”，为模型求解提供直观的背景。考虑到其余学科与计算机或数学的交叉情况，从属于第二类“工具型交叉”，为了便于后文分析，本文将数学和计算机科学的学科交叉归入第二类“工具型交叉”。

类似地，心理学和计算机科学的学科交叉也属于异常点。与心理学发生学科交叉的分布基本都位于图4的左下角，但是当心理学和计算机科学发生学科交叉时，学科交叉度和主题分布熵都增加，跳出了第三类“跨界交叉”的范围，介于第二类“工具型交叉”和第三类“跨界交叉”之间。计算机科学和心理学从上世纪开始相互交叉，相互渗透。在心理学研究领域，作为人类行为统计研究的心理实验往往产生海量的数据，计算机技术在海量数据的存储、分析、信息挖掘等方面，发挥着巨大的作用。同时，心理学理论在计算机领域的应用也得到了长足的发展，例如感性工学、人机交互系统、数据头套等技术的发展。同样地，为了便于分析，本文将心理学和计算机的学科交叉归入第三类“跨界交叉”。

4.4　不同交叉类型下的主题分析

根据公式(1)，计算115个主题下的主题学科交叉度。使用HemI软件(Heatmap Illustrator，version 1.0)绘制热谱图，将数值范围为0至1的主题学科叉度平均分为五段，渲染成五种颜色进行可视化展示。对15组学科交叉数据按照上节分类结果进行排列，如图5所示。其中，第一行组块为第一类“界内交叉”，第二行和第三行组块为第二类“工具型交叉”，第四行组块为第三类“跨界交叉”。

首先，根据整体分布可得，从上到下，红色、黄色色块数量逐渐减少，蓝色、深绿色色块数量逐渐增加。说明从第一类“界内交叉”至第三类“跨界交叉”的过程中，高主题学科交叉度的主题数量逐渐减少，低主题学科交叉度的主题数量逐渐增加。第一类“界内交叉”

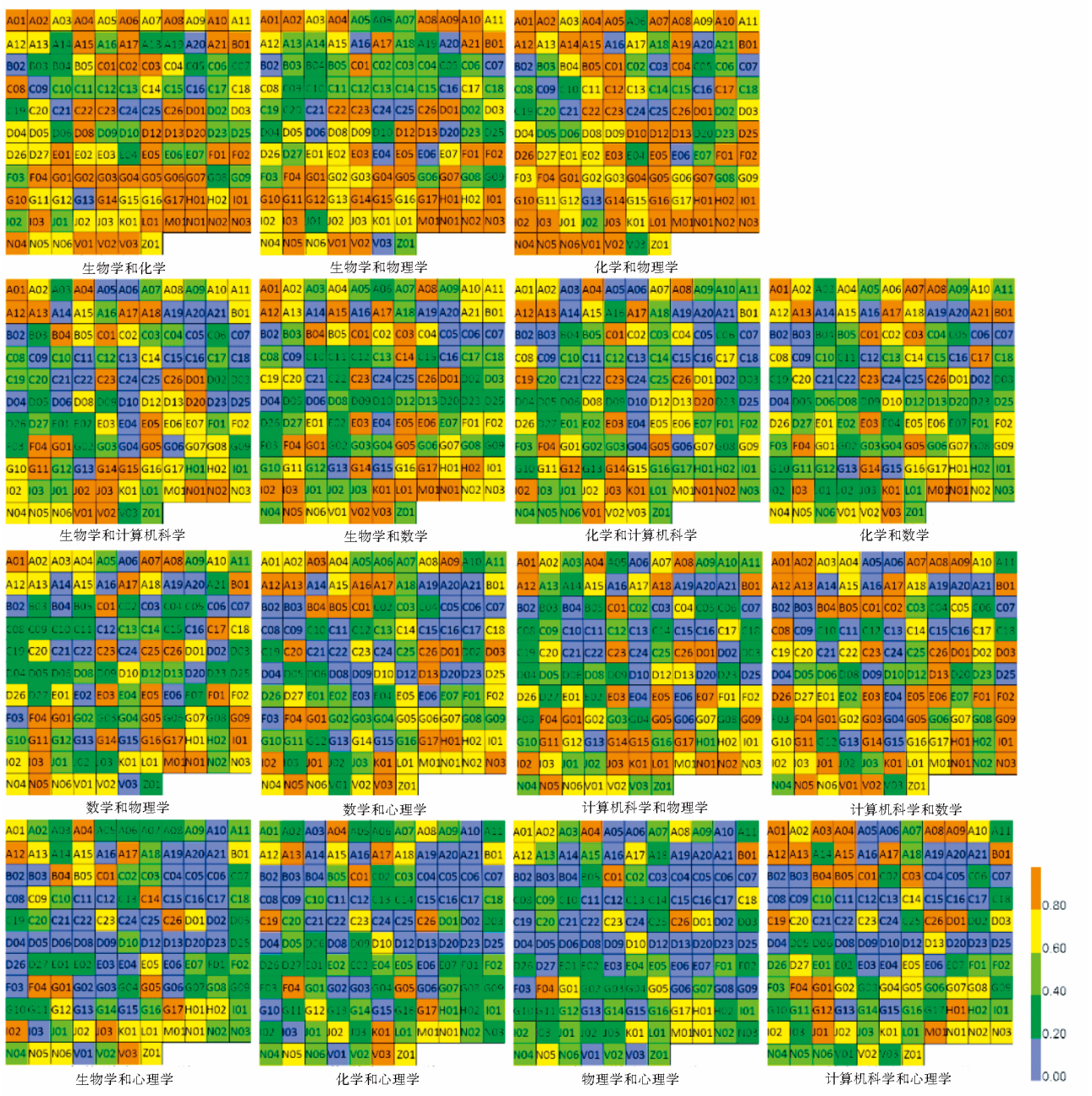

图 5 学科间交叉主题可视化

的学科交叉面广，主要集中在 A01（身体各部分）、A02（肌肉骨骼系统）、A04（呼吸系统）等 32 个主题下，而且 C 类（疾病知识）主题下，低主题学科交叉度的主题较多。第二类“工具型交叉”高主题学科交叉度的主题数量，和低主题学科交叉度的主题数量，分布相当。第三类“跨界交叉”在 C 类（疾病知识）和 D 类（药品知识）主题下的学科交叉度，普遍较低。

其次，学科交叉度高的学科间，会存在低主题学科交叉度的主题，反之亦然。举例来说，生物学和化学，生物学和物理学，化学和物理学的学科交叉度高，但仍存在低主题学科交叉度的主题。其中，三者都出现的低主题学科交叉度的主题为 A20（细菌结构）、B02（古生菌）、C16（先天性、遗传性和新生儿疾病和畸形）等六个主题。再比如，虽然心理学和化学，心理学和物理学，心理学和生物学的学科交叉度低，但仍存在高主题学科交叉度的主题。其中，三者都出现的高主题学科交叉度的主题为 A04（呼吸系统）、C01（细菌感染和真菌病）、C26（创伤与损伤）等四个主题。

5 结论

学科交叉研究是图书情报学科领域的研究热点。目前主题下的学科交叉度计算少有学者关注。本文提

出结合术语和引文内容，从内容层面计算主题学科交叉度的方法，并利用总体学科交叉度和学科交叉的主题分布熵值，对学科交叉类型进行分析。研究结果显示，该方法有效地将学科交叉分为各具特点的三种学科交叉类型：分别为“界内交叉”“工具型交叉”和“跨界交叉”。结合术语和引文内容对学科交叉现象进行分析的方法，弥补了元数据的不足，不但测量了学科间的总体交叉程度[23]，还具体展示了学科间的交叉点[23]和交叉主题，并从微观层面定量研究学科交叉类型。

通过本文的研究，我们认为：首先，学科交叉具有倾向性。学科交叉度越高，学科在更多的主题下也会更容易产生高交叉(如图5所示)。化学和物理学之间的高主题学科交叉度的主题个数，远远多于，物理学和心理学之间的高主题学科交叉度的主题个数。该结论可以为学科交叉研究提供新的研究维度。

其次，学科交叉具有多样性。一方面，学科交叉类型多样。本文将学科交叉分为三类，这三类学科交叉类型有各自的特性。第一类“界内交叉”主要为自然科学(以生物学、化学、物理学为代表)领域内部的学科交叉。百年来诺贝尔自然科学奖获得者所属的研究领域的相关研究[31]可以证明自然科学类学科易发生交叉，影响广。第二类“工具型交叉”主要是工具型学科(以数学和计算机科学为代表)与其他学科的学科交叉。工具型学科对于其他学科辅助作用十分突出。第三类“跨界交叉”主要是自然科学领域和社会科学领域(以心理学为代表)的学科交叉。随着学科“跨界”程度的增加，其学科交叉性逐渐减弱。原因可能是社会科学领域在知识体系上通常更加自给自足，知识从其他学科领域流入社会科学领域较为困难[32]。另一方面，学科交叉度和主题学科交叉度具有多样性。学科交叉度存在高低不同程度，高学科交叉度的学科交叉下，既有高主题学科交叉度，也有低主题学科交叉度。同样地，低交叉度的学科之间，也会存在较高程度的交叉主题。

最后，学科交叉具有特异性。如数学在六个学科中的学科总体交叉度变化范围最小(0.52－0.8)。可能原因是由于数学是一门工具性学科，是计算机科学、物理学、化学等各学科的基础，更容易与其他学科发生学科交叉。在发生学科交叉的同时，数学因其学科的辅助性，与其他学科的交叉性却又不强。在学科主题交叉度的分布中，数学与其他学科的学科主题交叉存在一致性，交叉主题的分布上具有相似性。

本研究也存在着较多的不足和局限性以期在下一步工作中解决。比如本研究的实验数据规模较小，学科选取存在一定的局限性。下一步工作需要获取更多的实验数据，引入更多的学科，以期获得更有效和更一般性的结论[23]。另外，由于学科知识体系的构建较为困难，而医学主题词表的使用广泛并且具有权威性，数据存储格式为XML，易处理，是本文理想的主题术语来源。下一步工作将扩大主题术语的涵盖范围，选取美国国会主题词表(LCSH)进行主题术语的对比性分析实验。

参考文献：

[1] WATSON J D. The double helix: apersonal account of the discovery of the structure of DNA[M]. New York: Weidenfeld and Nicolson, 1968.

[2] 许海云，尹春晓，郭婷，等. 学科交叉研究综述[J]. 图书情报工作，2015，59(5)：119－127.

[3] HUUTONIEMI K, KLEIN J T, BRUUN H, et al. Analyzing interdisciplinarity: typology and indicators[J]. Research policy, 2010, 39(1)：79－88.

[4] SMALL H. Cited documents as concept symbols[J]. Social studies of science, 1978, 8(3)：327－340.

[5] 郑述谱. 俄罗斯当代术语学[M]. 北京：商务印书馆，2005.

[6] 张榕. 术语学与术语信息处理[M]. 北京：中国社会科学出版社，2015.

[7] 马晓雷. 被引内容分析：探究领域知识结构的新方法尝试[M]. 北京：外语教学与研究出版社，2011.

[8] 胡志刚. 全文引文分析：理论、方法与应用[M]. 北京：科学出版社，2016.

[9] 章成志，徐庶睿，卢超. 利用引文内容监测多学科交叉现象的方法与实证[J]. 图书情报工作，2016，60(19)：108－115.

[10] MUGABUSHAKA A, KYRIAKOU A, PAPAZOGLOU T. Bibliometric indicators of interdisciplinarity: the potential of the Leinster-Cobbold diversity indices to study disciplinary diversity[J]. Scientometrics, 2016, 107(2)：593－607.

[11] WAGNER C S, ROESSNER J D, BOBB K, et al. Approaches to understanding and measuring interdisciplinary scientific research (IDR): a review of the literature[J]. Journal of informetrics, 2011, 5(1)：14－26.

[12] ABOELELA S W, LARSON E, BAKKEN S, et al. Defining interdisciplinary research: conclusions from a critical review of the literature[J]. Health services research, 2007, 42(1, part 1)：329－346.

[13] STIRLING A. On the economics and analysis of diversity[J]. Science policy research unit, 1998, 28：1－156.

[14] STIRLING A. A general framework for analysing diversity in science, technology and society[J]. Journal of the Royal Society interface, 2007, 4(15)：707－719.

[15] PURVIS A, HECTOR A. Getting the measure of biodiversity[J].

Nature, 2000, 405(6783): 212.
[16] PORTER A L, COHEN A S, ROESSNER J D, et al. Measuring researcher interdisciplinarity [J]. Scientometrics, 2007, 72(1): 117-147.
[17] RAFOLS I, MEYER M. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience [J]. Scientometrics, 2010, 82(2): 263-287.
[18] VUGTEVEEN P, LENDERS R, VAN DEN BESSELAAR P. The dynamics of interdisciplinary research fields: the case of river research [J]. Scientometrics, 2014, 100(1): 73-96.
[19] XU H, GUO T, YUE Z, et al. Interdisciplinary topics of information science: a study based on the terms interdisciplinarity index series [J]. Scientometrics, 2016, 106(2): 583-601.
[20] ZHANG G, DING Y, MILOJEVIC S. Citation content analysis (CCA): a framework for syntactic and semantic analysis of citation content [J]. Journal of the American Society for Information Science & Technology, 2012, 64(7): 1490-1503.
[21] 祝清松, 冷伏海. 基于引文内容分析的高被引论文主题识别研究 [J]. 中国图书馆学报, 2014, 40(1): 39-49.
[22] LIU S, CHEN C. The differences between latent topics in abstracts and citation contexts of citing papers [J]. Journal of the American Society for Information Science & Technology, 2013, 64(3): 627-639.
[23] 徐庶睿, 卢超, 章成志. 术语引用视角下的学科交叉测度——以PLOS ONE上六个学科为例 [J]. 情报学报, 2017, 36(8): 809-820.
[24] MEI Q, ZHAI C. Generating impact-based summaries for scientific literature [C] // Proceedings of the Association for Computational Linguistics. Columbus: ACL, 2008: 816-824.
[25] 章成志. 多语言领域本体学习研究 [M]. 南京: 南京大学出版社, 2012.
[26] 金建国. 聚类方法综述 [J]. 计算机科学, 2014(S2): 288-293.
[27] FREY B J, DUECK D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972-976.
[28] 孙海霞, 钱庆, 吴英杰, 等. MeSH词表的语义相似度计算研究 [J]. 现代图书情报技术, 2010(6): 12-16.
[29] 钱学森. 现代科学技术 [N]. 人民日报, 1977-12-09(2).
[30] 周经野, 刘任任. 计算机科学的数学基础 [M]. 湖南: 湘潭大学出版社, 2007.
[31] 张春美, 郝凤霞, 闫宏秀. 学科交叉研究的神韵——百年诺贝尔自然科学奖探析 [J]. 科学技术与辩证法, 2001, 18(6): 63-67.
[32] YAN E. Finding knowledge paths among scientific disciplines [J]. Journal of the Association for Information Science and Technology, 2014, 65(11): 2331-2347.

作者贡献说明:
徐庶睿: 实施实验过程, 进行数据分析, 起草论文;
章成志: 数据采集, 提出论文研究思路, 修改论文;
卢超: 讨论研究思路, 修改论文。

Using Citation Contents for the Interdisciplinary Type Analysis at a Topical Level

Xu Shurui[1,2] Zhang Chengzhi[1,2,3] Lu Chao[1,2]

[1] Department of Information Management, Nanjing University of Science and Technology, Nanjing 210094
[2] Jiangsu Science and Technology Collaborative Innovation Center of Social Public Safety, Nanjing 210094
[3] Jiangsu Key Laboratory of Data Engineering and Knowledge Service (Nanjing University), Nanjing 210093

Abstract: **[Purpose/significance]** This paper calculates the interdisciplinary degree at a topical level and establishes quantitative standards for the interdisciplinary classification in terms of the contents, because the current issues of interdisciplinary macro-research lack the description of interdisciplinary topics, and interdisciplinary micro-research is still in the topic detection stage. **[Method/process]** Firstly, full-text articles were collected, and terminologies were extracted. Secondly, the repetition rate of subject terminologies in citation contents were calculated. Then, the degrees of topical interdisciplinarity were calculated. Finally, the disciplines were categorized according to the topical distribution entropy of interdisciplinarity. **[Result/conclusion]** The results are as follows: (i) All six disciplines share abundant interdisciplinary knowledge with much medical theoretical basis knowledge but rare medical practical knowledge. (ii) Three types of interdiscipline are observed: internal interdiscipline, instrumental interdiscipline, external interdiscipline. In conclusion, the interdisciplinary types can be quantitatively studied at the micro level through the terminologies in the citation content.

Keywords: topic interdisciplinary type interdisciplinarity citation content terminology