

生物医学领域多源文献数据学科映射方法 优化研究^{*}

李晓瑛 刘懿 李爱花 杨雪梅 唐小利
(中国医学科学院医学信息研究所, 北京 100020)

摘要: 针对生物医学领域多源文献数据分类体系不一致、学科归属判别困难的问题, 在分析国内外主流文献分类体系内容结构与已有映射方法优缺点的基础上, 制定生物医学领域分类体系映射规则, 提出集语义自动匹配、文献分类数据验证、人工审核于一体的融合映射方法。最终分别建立国际三大分类体系Web of Science、Scopus、FOR中生物医学类目与教育部《研究生教育学科专业目录(2022年)》相关类目的语义映射关系, 开展用于肿瘤专题文献篇级分类的国内外学科体系映射实践。研究成果能够促进国际学术成果的集中检索、分类组织、统计评价等科研管理与学术分析工作, 并为我国学科专业目录调整、人才培养、科研布局等提供决策依据与信息支撑。

关键词: 语义映射; 学科归属; 分类体系; 映射规则

中图分类号: G250 **DOI:** 10.3772/j.issn.1673-2286.2023.03.001

引文格式: 李晓瑛, 刘懿, 李爱花, 等. 生物医学领域多源文献数据学科映射方法优化研究[J]. 数字图书馆论坛, 2023 (3) : 1-9.

新医科建设背景下, 医学与理、工、农等学科交叉融合深化, 对高校科研管理与学术分析任务提出了新要求和挑战。目前各高校的科研管理和学术评价主要以多来源且分类体系各异的数字文献数据为基础, 采用知识组织方法、文献计量法和信息分析方法, 对多源跨库数据进行融合分析。其中, 来自Web of Science (WoS)、Scopus等国际权威引文数据库的外文期刊数据是统计分析的基础与对象。然而, 这些文献数据库基于相对独立的分类体系组织各自的资源, 与我国高校和科研院所普遍采用的教育部《研究生教育学科专业目录(2022年)》(以下简称《学科目录》)有所差异。将多来源的国际文献数据按照统一的分类体系进行组织, 解决资源的学科归属问题, 是我国高校开展科研管理

与学术分析任务的一个重要前提。

一般而言, 可通过两种不同的途径来解决跨库文献数据的学科归属问题。一种是基于不同分类体系的映射与互操作^[1], 即在理解源分类与目标分类内涵的基础上, 建立等同或相近类目之间的映射关联, 将文献数据的源分类信息匹配到目标分类体系中。另一种则是摒弃文献原有分类数据, 将其视为大数据分类问题^[2-3], 应用神经网络和深度学习模型, 按照目标分类体系进行统一归类。鉴于当前缺少成熟的标注语料, 加之生物医学文献数据体量庞大、学科高度交叉, 第一种映射方法较之后者更加易于实现和扩展。

近年来, 国内外学者在分类体系映射的理论方法和互操作实践等方面开展了许多研究, 其中大部分围绕

收稿日期: 2023-03-08

^{*}本研究得到中国医学科学院医学与健康科技创新工程重大协同创新项目“生物医学文献信息保障与集成服务平台”(编号: 2021-I2M-I-033)资助。

应用较为广泛的《杜威十进分类法》(Dewey Decimal Classification, DDC)、《中国图书馆分类法》(以下简称《中图法》)展开。典型成果包括《瑞典图书分类法》(Sveriges Allmänna Biblioteksförening, SAB)与DDC第19版、第21版的映射,相关结果已经应用于瑞典图书馆联合目录系统;欧盟著名的学科信息门户项目Renardus采用DDC作为交叉浏览的知识框架,将其分别与荷兰、瑞典、德国、法国、英国等国家分布信息源采用的自编分类体系进行映射^[4];英美学者对DDC与《美国国会图书馆图书分类法》(Library of Congress Classification, LCC)^[5]、《国际十进制分类法》(Universal Decimal Classification, UDC)^[6]之间的映射进行了有益探索。国内分类体系映射研究主要聚焦DDC与《中图法》^[7-10]、LCC与《中图法》^[11]、《国际专利分类法》(International Patent Classification, IPC)与《中图法》的映射^[12],以及近年来面向科研评价需求的国家自然科学基金项目分类与中国知网、WoS等国内外大型数据库学科分类体系的映射^[13]。国外学者更加关注学科分类体系映射支持浏览、检索等应用及项目实践,国内学者则致力于改进新兴信息技术驱动下的学科分类体系映射方法。在实证研究方面,国内外学者更倾向于选取内涵凝练、外延清晰的学科,例如化学、农学、林学、法学等;鉴于生物医学领域类目繁杂、学科交叉度高、专业性强,近年来仅有个别学者采用编辑距离法建立了医学分类表之间的映射关系^[14]。随着深度学习、神经网络等技术的深度应用,学科分类体系映射方法不断优化,映射效率和准确率逐步提升,但是相关研究仍未有效解决生物医学领域多源跨库文献数据的学科归属判别问题。

1 国内外主要分类体系及映射方法

1.1 国内外主要分类体系

随着网络信息资源增长及数字图书馆兴起, WoS分类体系、Scopus分类体系、研究领域(Fields of Research, FOR)等文献数据库分类法应运而生,因其均依据现代学科设置类目,故可与教育部《学科目录》建立映射关系。表1简述了本研究涉及的7种国内外分类体系的编制机构、内容结构、编制目的及应用现状等基本情况, WoS、Scopus等数据库面向国际文献分类组织而自建的分

类体系的细粒度专题分类体系或将成为未来的发展趋势。相对而言,国内学科分类体系建设虽起步晚但发展快,日渐赶上国际步伐。《中图法》初版发布于1975年,至今已修订发布了五版^[15-18],第六版本已通过审查,新版可期。《学科分类与代码》(GB/T 13745—2009)、《学科目录》等信息分类标准与学位授予依据,以及支撑文献深度组织与图书编目的《中图法》,成为目前国内学科分类体系的典型代表。

1.2 国内外主要分类体系映射方法

按照分类体系映射过程中的人力参与程度,将现有的主流映射方法分为三大类:人工映射、自动映射及人机结合(表2)。其中,类目相似度计算、同现法(包括数据统计法)、交叉检索法、机器学习法等自动映射方法执行效率高^[12,19-21],并且可通过反复优化不断提升映射效果,成为实践探索的首选。然而,由于编制目的和面向对象不同,文献分类体系与学科分类体系存在类目内涵粒度不一致等情况,导致自动映射结果不准确,因此需要借助领域专家的专业知识进一步从语义层面调整优化。

2 生物医学领域多源文献数据学科映射规则与方法研究

2.1 多源文献数据学科映射规则研究

为了充分保障WoS、Scopus、FOR文献分类体系与教育部《学科目录》之间的映射结果质量,同时避免后期反复修正,提前制定映射规则。本研究在遵循国际标准ISO 25964-2:2013的基础上,提出面向生物医学领域不同分类体系映射的等同、上位、下位和相关4种映射通用规则以及针对一对一类目映射关系要求的专用规则。

(1) 等同映射。当文献分类体系与学科分类体系的类目内涵相同时,通过等同映射,在二者之间建立同义关系。

(2) 上位映射。WoS文献分类体系的类目揭示粒度较细,但教育部《学科目录》的类目粒度粗,导致大部分类目之间无法建立等同映射。对WoS非交叉类目,以就宽不就窄的上位映射规则与同一个学科类目建立多对一关联关系; WoS交叉学科类目经上位

表1 国内外主要文献学科分类体系概况

分类体系	内容结构	编制目的	应用现状
WoS分类体系	科睿唯安公司针对WoS数据库核心合集采用的学科分类体系, 内容涉及自然科学、工程技术、生物医学、社会科学、艺术与人文五大领域, 包含22个一级大类、254个二级学科类别	对WoS数据库中SCIE、SSCI等独立子库收录的2万余种期刊, 依据现代学科进行分类组织, 便于用户从现代学科角度检索获取论文	细粒度, 适用于期刊级别的学科分类 (见WoS数据库)
Scopus分类体系	爱思唯尔公司用于Scopus数据库的期刊分类体系, 包含生命科学、社会科学、自然科学、医学四大一级领域、27个二级大类、334个三级学科子类	对Scopus数据库收录的4万余种期刊进行学科分类	粗粒度, 适用于期刊级别的学科分类 (见Scopus数据库)
FOR	数字科研解决方案有限公司基于《澳大利亚与新西兰标准研究分类法》(Australian and New Zealand Standard Research Classification) 改编形成; 为两级体系, 包含一级大类24个、二级类目212个	对Dimensions数据库收录的1.3亿篇论文、1.5亿件专利、620万项基金等数据进行分类组织, 支持用户按FOR体系分类导航进行过滤检索	细粒度, 适用于篇级文献的学科分类 (见Dimensions数据库)
IPC	改编自《国际专利分类斯特拉斯堡协定》(International Patent Classification Strasbourg Agreement), 严格按照部、大类、小类、大组和小组的五级体系组织, 最新的第八版基本版约有2万条类目	对海量专利文献进行组织和管理, 为适应世界各国专利新颖性、创造性和检索有效性等特点, 改革形成第八版	细粒度, 国际通用的专利分类和检索工具 (见Dimensions数据库)
ICRP CT	由国际癌症研究合伙组织 (International Cancer Research Partnership, ICRP) 建立并维护的癌症分类体系 (Cancer Types, CT), 共62个类目	对癌症研究相关的文献、基金、奖励等数据, 按照具体的癌症类别进行分类组织	细粒度, 适用于肿瘤领域篇级文献的分类 (见Dimensions数据库)
《学科目录》	国务院学位委员会及教育部发布的《学科目录》是我国开展学位授权审核与学科管理、高校开展学位授予与人才培养工作的依据, 分14个学科门类、181个一级学科	为适应我国经济、社会、科技和高等教育的发展, 规范学科专业的设置与管理而编制的学科分类体系	粗粒度, 适用于研究生专业设置及学士、硕士、博士学位授予工作, 亦是现阶段高校和科研院所普遍采用的学科分类与评估依据
《中图法》	由中国国家图书馆编制与维护, 是我国图书馆和情报单位普遍使用的综合性分类法。包含马列主义、毛泽东思想、哲学、社会科学、自然科学、综合性图书五大部类、22个大类。第五版为最新版, 共53 713个类目	充分吸取国外分类法编制理论和技术, 针对我国图书资料的内容与特点, 依据学科分门别类形成的一套分类体系	细粒度, 被我国高校及专业图书馆和情报机构广泛使用

映射, 与多个《学科目录》类目之间形成多对多映射关系。

(3) 下位映射。Scopus分类体系中带有“miscellaneous”杂项标识的类目揭示粒度较粗, 无法对应教育部《学科目录》, 只能采用下位映射, 将大概概念映射到具有包含关系的小概念中。

(4) 相关映射。当上述3种映射规则都不适用时, 考虑语义关系较弱的映射关系, 用于辅助映射但不建立映射关系。

以上4种映射规则并非同级并列关系, 而是具有明确的优先级, 优先级从高到低依次为等同映射、上位映射、下位映射、相关映射。此外, 本研究增加一条专用规则, 即在一对多关系中区分最优映射和推荐映射, 通过添加不同标识, 为对学术成果进行唯一分类提供便利。例如将Scopus分类体系中的“2732 Orthopedics and Sports Medicine”最优映射到教育部《学科目录》中的“1002 临床医学”(标识P), 同时推荐映射到“1009 特种医学”(标识S)。

表2 国内外主要分类体系映射方法概况

类别	方法名称	基本原理	优势	不足
人工映射	专家映射法	领域专家依据其自身经验与主观判断,在分类体系的类目之间建立对应关系	映射结果汇聚了领域专家的智慧,准确性较高	工作量大、执行周期长;易产生结果不一致的问题,缺乏衡量结果的金标准
自动映射	类目相似度计算	计算类目文本之间的相似度,选取与当前类目文本相似度最高的类目作为映射结果	文本相似度计算过程简单,易于工程化	仅计算类目字面相似度,忽略了深层次的语义信息
	同现法	利用分类标引后的期刊文献或书目数据,通过统计类目在同一批标引数据中的同现频次,计算类目之间相关度	映射结果来源于大量分类标引数据,具有一定的科学性与普适性	对分类标引数据的完整性和准确性要求高,否则所得映射结果无意义
	交叉检索法	用一种分类体系检索另一种分类体系标引的资源,通过分析检索结果,建立两种分类体系之间的映射关系	处理过程直观、易懂,具有较高的可操作性	得到的映射结果多为一对多的关系,粒度较粗
	机器学习法	将人工映射结果作为标注语料,训练机器学习模型,产生分类器,进一步预测类目之间的归类匹配关系	高性能计算设备的普及、深度学习模型的兴起使得该方法受到极大关注	需要预先准备质量高且具有代表性的语料,且映射结果可解释性较差
人机结合	基于众包模式的分类映射法 ^[22]	通过自动映射方法生成候选映射类目,在此基础上引入众包模式,利用群体智慧解决计算机难以完成的任务,提高映射准确率和覆盖率	在自动映射的基础上进行人工干预,有效纠正错误并扩展更多映射关系,提高了映射结果质量	大众参与分类映射存在一定难度,需要专业培训

2.2 多源文献数据学科映射方法研究

本研究在上述映射规则的指导下,充分发挥类目相似度计算、交叉检索法、同现法、专家映射法等已有映射方法的优势,提出一种面向生物医学领域分类体系映射的融合方法(见图1),涉及预处理、自动映射、人工映射及映射优化4个关键环节。

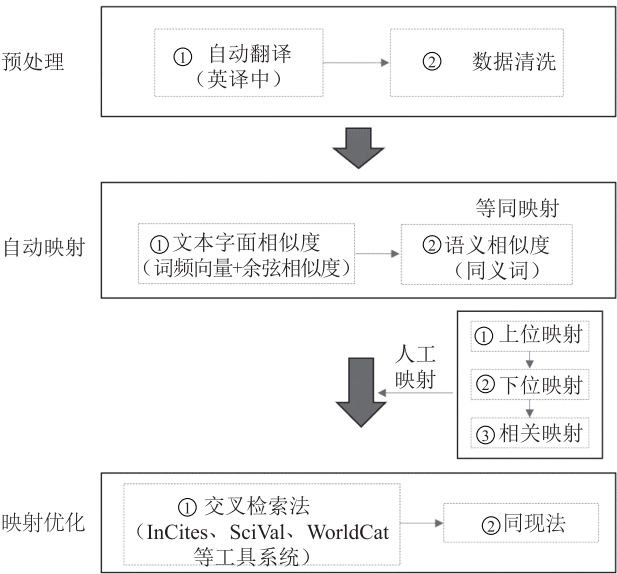


图1 分类体系映射融合方法框架

(1) 预处理。预处理主要包括自动翻译和数据清洗。自动翻译指将英文语种的文献分类体系类目翻译成中文,减少在外文类目识读上的时间和精力消耗,提高分类体系映射工作效果。可利用的成熟翻译工具较多,如百度翻译、CNKI翻译助手等,这在一定程度上保障了自动翻译的可行性与准确性。数据清洗指对个别未整体翻译的中英混合类目进行拆分清洗及人工翻译,为后续的自动映射处理做好数据基础。

(2) 自动映射。按照制定的4种分类体系映射规则,基于同义关系的等同映射规则具有最高的优先级,且容易通过计算机算法模拟实现。从类目术语的文本字面相似度和语义相似度两方面入手,完成文献分类体系与学科分类体系中同类目自动映射。前期实验表明,融合词频向量的余弦相似度法因具有良好的结果区分度^[23],优于Jaccard、Dice等同类算法,故成为本研究的首选。此外,医学领域积累了《中文医学主题词表》(CMeSH)等成熟的医学词表^[24]。现有医学词表提供了规范的术语同义关系,从语义层面提升了自动映射效果。

(3) 人工映射。鉴于计算机程序较难自动判断文献分类体系与学科分类体系的类目内涵,无法准确地完成上位映射、下位映射及相关映射,本研究在前两个阶

段自动处理的基础上,参考类目释义等注释信息,进行必要人工调整与增补。

(4) 映射优化。交叉检索法与同现法是一类复用现有资源和成果的分类体系映射方法。科睿唯安In-Cites、爱思唯尔SciVal等科研学术分析工具支持用户按照教育部《学科目录》交叉检索WoS、Scopus等数据库。通过获取这些数据库收录期刊的分类信息,可侧面了解国际出版商建立的WoS、Scopus等文献分类体系与教育部《学科目录》的映射关系,用于优化验证本研究初步得到的映射结果^[25]。此外,由于WoS、Scopus、Dimensions等数据库收录的国际期刊存在较高的重合度,可借助同一种期刊在WoS、Scopus、FOR等国外文献分类体系中的分类数据,采用同现法优化完善其与我国教育部《学科目录》相关类目的映射关系。

3 生物医学领域多源文献数据学科映射实践

生物医学领域文献量大,以期刊为分类对象的WoS、Scopus分类体系类目设置不均衡,特别是临床医学相关类目粒度过细,无法较好地满足按类统计分析的需求。以FOR为代表的国外学科分类体系受国情和学科发展背景影响,分类原则和类目划分具有鲜明的地域特色。目前我国教育部《学科目录》根据学科专业发展规律,将医学专业分为基础医学、临床医学、口腔医学、公共卫生与预防医学、中医学、中西医结合、药理学、中药学、医学技术、法医学、护理学11个大类,各类之间专业粒度较为均匀,同时符合教育统计分类的要求。因此本研究在期刊层面,将WoS、Scopus、FOR、IPC等国际主流分类体系与教育部《学科目录》进行映射,以适应国内分析研究和教育统计等工作需要。由于教育部《学科目录》分类粒度较粗,在单篇文献层面,选取细粒度并以学科为分类标准的《中图法》作为映射目标。

3.1 WoS、Scopus与教育部《学科目录》映射

为了满足我国生物医学相关高校和科研院所建设一流学科的学术成果分析评价和国际对标等需求,在自动映射的初步结果基础上,基于InCites分析工具及WoS、Scopus类目完成双重优化,最终建立了WoS的104个类目与教育部《学科目录》14个生物医学类

目之间的147对语义映射关系,以及Scopus的154个类目与教育部《学科目录》13个生物医学类目之间的217对语义关联(见图2)。在语义映射中不仅从类目字面上开展直接映射,而且通过深入分析类目内涵,增加更多的映射关系。以WoS类目“Food Science & Technology”与教育部《学科目录》“1007 药学”的映射为例:通过WoS官网了解到该类目注释为“Food Science & Technology covers resources concerning various aspects of food research and production, including food additives and contaminants, food chemistry and biochemistry, meat science, food microbiology and technology, dairy science, food engineering and processing, cereal science, brewing, and food quality and safety”,表明其内涵与药学相关;进一步分析WoS期刊分类数据,发现大量药学期刊的分类为“Food Science & Technology”和“Pharmacology & Pharmacy”,这验证了从语义层面建立“Food Science & Technology”与教育部《学科目录》“1007 药学”这一映射关系的可行性。

3.2 FOR与教育部《学科目录》映射

FOR按照24个大类、212个子类向下设列类目,其中与生物医学相关的类目主要集中于“31 Biological sciences”“32 Biomedical and clinical sciences”“42 Health sciences”及“52 Psychology”。经过自动映射、人工映射,最终确定FOR分类体系33个生物医学类目与教育部《学科目录》12个一级类目之间的47对映射关系(见图3)。优先进行同义匹配(如“2900 General Nursing”与“1011 护理学”),当无法进行等同映射时则关联到《学科目录》相近类目中(例如“2916 Nutrition and Dietetics”与“1011 护理学”)。FOR与教育部《学科目录》的映射有助于我国高校科研管理者与研究人员按照熟悉的《学科目录》归类组织Dimensions数据库中的海量生物医学文献。

3.3 IPC与教育部《学科目录》映射

IPC是管理、分析专利文献的有效工具,已成为国际通用的专利文献分类体系。IPC兼容面向功能与面向应用的编制原则,注重按照技术、工艺的基本作用分类,其中生物医学相关类目主要集中于“A61医学或

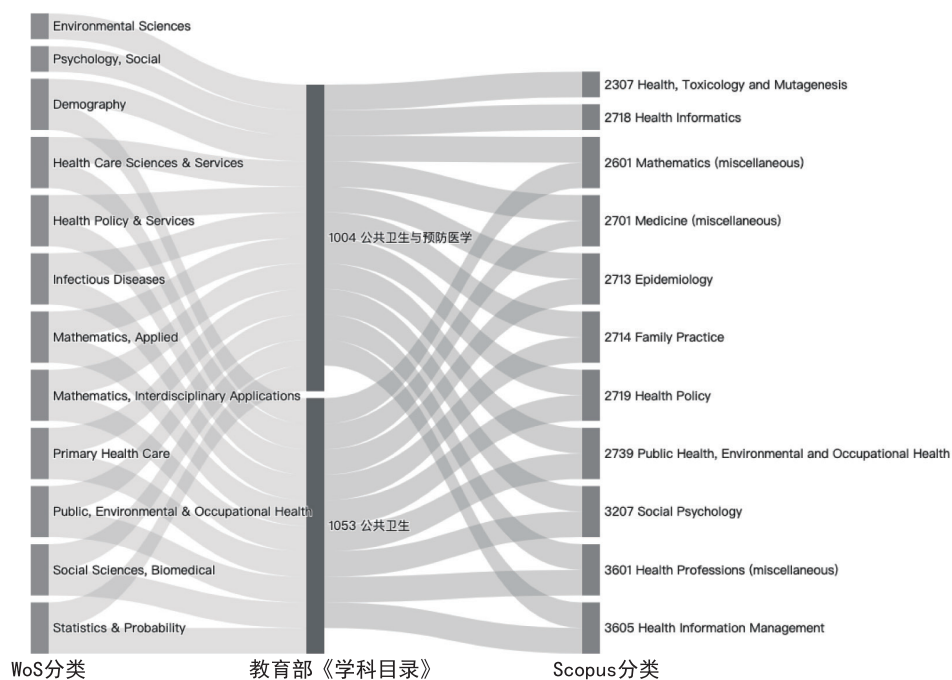


图2 WoS、Scopus生物医学类目与教育部《学科目录》语义映射结果(部分)



图3 FOR生物医学类目与教育部《学科目录》语义映射结果(部分)

兽医学; 卫生学”“C07有机化学”“C08有机高分子化合物; 组合物”“C12生物化学; 微生物学; 酶学; 突变或遗传工程”。鉴于IPC类名多为描述性文本, 基于等同映射的自动处理并不适用, 故专业人员在理解IPC类目含义和分析专利分类数据的基础上, 建立了IPC与教育部《学科目录》6个类目之间的78对语义关联。例如, IPC“A61K9 以特殊物理形状为特征的医药配制品[2006.01]”被最优映射到教育部《学科目录》“1002 临床医学”(标识P), 同时推荐映射到“1007 药学”(标识S)。IPC与教育部《学科目录》的映射, 有助于我国学者遴选与分析国际专利数据, 能够为我国制订专利战略规划、专利布局及竞争技术发展政策提供决策依据与信息支撑。

3.4 ICRP CT与《中图法》“R73 肿瘤学”映射

肿瘤学是医学科学中十分重要的一个专门分支, 精准医学计划项目推动了对肿瘤发生发展机制及临床治疗的深入研究。ICRP CT体系将目前已知的肿瘤划分为61种类型(如“Lung Cancer”“Brain Tumor”等), 将其余肿瘤统一标记为“uncoded”, 一共62个类目。鉴于具体的肿瘤中英文名称共识度比较高, 提出的等同映射处理效果较好, 完成了对57个ICRP CT类目与《中图法》“R73 肿瘤学”50个子类目的准确映射, 其余类目经人工参照《中图法》类目释义完成语义映射。以“Melanoma”为例, 百度翻译工具译为“黑色素瘤”, 人工参照“R739.5 皮肤肿瘤”类目释义“专论上皮性瘤、黑痣及恶性黑色素瘤的著作入此”, 建立语义映射关系。ICRP CT体系目前已用于Dimensions数据库1 200万篇肿瘤英文文章的篇级分类, 建立其与《中图法》“R73 肿瘤学”161个子类目的语义映射, 有助于我国科研人员按照熟悉的学科分类体系细粒度地组织与归类肿瘤领域国际学术成果, 并为深度学习模型训练语料准备与结果验证等提供数据支撑。

4 生物医学领域多源文献数据学科映射研究意义

4.1 助力科研管理与学术分析

WoS、Scopus两大国际引文数据库收录的学术成果的数量、影响因子、被引频次等指标是高校开展科研

管理、学术分析、优势学科评估的重要评价维度。我国科研管理人员需要花费较多时间和精力从这些数据库中筛选学术成果, 并判断成果在教育部《学科目录》中的所属学科。建立WoS、Scopus文献分类体系与教育部《学科目录》生物医学领域相关类目的关联映射关系, 可助力我国新医科背景下的科研管理与学术分析。

4.2 辅助我国学科专业目录调整和人才培养

WoS、Scopus、Dimensions等国际数据库基本覆盖了全球最新的高质量学术成果, 反映了欧美发达国家的重点、热点。通过跟踪这些资源的学科属性并与我国教育部《学科目录》进行对比, 为我国调整学科设置与科研布局提供启示。通过及时增设相应学科, 尽早制定相关学科人才培养计划, 提高我国在前沿热点领域的综合实力, 提升中国在科技创新中的国际话语权。

4.3 赋能跨库文献数据的交叉浏览与检索

通过广泛集成WoS、Scopus、Dimensions等在国际上具有较高影响力的文献资源, 建立不同文献分类体系与我国教育部《学科目录》之间的语义映射, 将具有世界影响力的外文学术成果均按照我国科研人员熟悉的学科分类体系进行归类与组织, 实现多源跨库文献资源的交叉浏览和一站式检索, 提升科研学术成果的开放共享与管理利用效果, 有助于加快我国高端学术交流平台构建和科技自立自强。

4.4 为深度学习模型提供分类标注语料与结果对照集

随着计算机软硬件能力的提升, 深度学习模型得到广泛应用, 使得大规模文献篇级粒度的自动分类成为可能。本研究得到的学科分类体系映射结果能够辅助国际论文成果在沿用其原有期刊级别分类数据的基础上, 按照我国教育部《学科目录》进行初步学科归类。通过人工手段进一步细化调整, 形成篇级文献分类数据, 从而在分类标注语料准备方面为训练基于深度学习模型的文献自动分类算法提供便利。此外, 通过参考国际肿瘤分类体系ICRP CT与《中图法》“R73 肿瘤学”的映射结果, 能够快速实现对Dimensions数据库

中上千万篇肿瘤英文文献的分类,亦可将映射结果作为对照集,与由深度学习模型直接得出的肿瘤文献篇级分类结果进行对比分析,这亦是本研究今后继续推进的一个重要方向。

5 结语

建立WoS、Scopus、Dimensions等国际文献数据库分类体系与我国现行的常用学科分类体系(如教育部《学科目录》、《中图法》)之间的映射关系,是多来源文献数据学科归属问题的一种行之有效的解决方案,在国内外学术成果的集中检索、分类导航、统计评价等科研管理与学术分析,乃至全球科研成果的共享利用等方面起到促进作用。本研究在调研国内外主要文献分类体系编制目标与应用现状的基础上,分析了已有映射方法的优势与不足,提出面向生物医学领域多源文献数据学科映射的规则与融合方法,进一步开展了WoS、Scopus、FOR等文献分类体系与我国教育部《学科目录》、《中图法》的映射实践,解决了多源跨库生物医学文献数据的学科归属判别问题。今后,将基于现有多源文献期刊级别分类数据及分类体系映射结果,结合深度学习模型,推进大规模生物医学篇级文献分类,实现对国际学术成果的细粒度组织与揭示。

参考文献

- [1] 张蒂,郝晋清,王珺.映射之难:人文社科学科评价中跨库数据的学科归属问题[J/OL].图书馆论坛:1-12[2022-10-11].<http://kns.cnki.net/kcms/detail/44.1306.G2.20220715.1316.004.html>.
- [2] 刘江峰,林立涛,刘畅,等.深度学习驱动的海量人文社会科学学术文献学科分类研究[J].情报理论与实践,2023,46(2):71-81.
- [3] 张智雄,赵旸,刘欢.构建面向实际应用的科技文献自动分类引擎[J].中国图书馆学报,2022,48(4):104-115.
- [4] KOCH T, NEUROTH H, DAY M. Renardus: Cross-Browsing European Subject Gateways via a Common Classification System (DDC) [M]//MCILWAINE I C. Subject Retrieval in a Networked Environment. Berlin: De Gruyter Saur, 2003: 25-34.
- [5] ZINS C, SANTOS P L V A C. Mapping the knowledge covered by library classification systems[J]. Journal of the American Society for Information Science and Technology, 2011, 62(5): 877-901.
- [6] From UDC to DDC: reclassification at the University of Bath[EB/OL].[2023-03-01].<https://researchportal.bath.ac.uk/en/publications/from-udc-to-ddc-reclassification-at-the-university-of-bath>.
- [7] 戴剑波,侯汉清.文献分类法自动映射系统的构建:以《中国图书馆分类法》与《杜威十进分类法》为例[J].情报学报,2006,25(5):594-599.
- [8] 赵冬梅.基于数学类目的DDC22与CLC5映射分析[J].图书馆论坛,2013,33(3):126-130.
- [9] 李珂,宋文.农业科学领域DDC和CLC类目映射分析[J].图书馆理论与实践,2015(10):58-62.
- [10] 原小玲.DDC22与CLC5化学类目映射分析[J].图书馆理论与实践,2013(3):71-74.
- [11] 蒋彦廷.依据《中图法》的英文图书分类探索:结合图情与一般领域的文本增强方法[J/OL].北京大学学报(自然科学版):1-11[2023-01-03].<https://doi.org/10.13209/j.0479-8023.2022.070>.
- [12] 何贤敏,李茂西,何彦青.基于孪生BERT网络的科技文献类目映射[J].计算机研究与发展,2021,58(8):1751-1760.
- [13] DU W, LAU R Y K, MA J, et al. A multi-faceted method for science classification schemes (SCSs) mapping in networking scientific resources[J]. Scientometrics, 2015, 105(3): 2035-2056.
- [14] 单连慧,赵迎光,钱庆.基于词汇相似度的医学分类体系映射研究与实现[J].医学信息学杂志,2016,37(11):46-50.
- [15] 司莉,饶曦.DDC新特点及其对《中图法》修订的启示[J].数字图书馆论坛,2015(3):58-62.
- [16] 曹树金,颜丽君,汪东波.DDC、LCC、UDC网络版评析[J].中国图书馆学报,2002,28(6):61-65.
- [17] 肖竹青.《中图法》(第五版)修订后类目设置的研究[J].图书馆研究,2017,47(3):1-6.
- [18] 吴润珍.探讨使用《中图法》第五版后的图书改编问题[J].图书馆工作与研究,2013(12):73-76.
- [19] 贾君枝,陈瑞.基于关联书目数据的DDC与LCC类号映射实现[J].图书情报知识,2019(2):81-91.
- [20] 李悦,苏成,潘云涛.分类法映射研究综述[J].情报理论与实践,2018,41(9):154-160.
- [21] 李珂,宋文.分类法映射研究[J].图书馆杂志,2014,33(12):49-56,68.
- [22] 陈瑞,贾君枝.基于众包模式的分类法映射研究[J].情报理论与实践,2020,43(7):137-143.
- [23] 武永亮,赵书良,李长镜,等.基于TF-IDF和余弦相似度的文本

- 分类方法[J]. 中文信息学报, 2017, 31 (5) : 138-145.
- [24] 李晓瑛, 李军莲, 邓盼盼, 等. 医学知识组织系统构建研究与应用实践[J]. 数字图书馆论坛, 2020 (7) : 30-35.
- [25] 唐小利, 李晓瑛, 刘懿, 等. 基于分析工具与期刊同现双重优化的生物医学领域学科分类体系映射研究[J]. 医学与社会, 2022, 35 (12) : 108-113, 144.

作者简介

李晓瑛, 女, 1982年生, 博士, 副研究员, 研究方向: 医学知识组织。
 刘懿, 女, 1995年生, 硕士, 助理馆员, 研究方向: 信息计量与科学评价。
 李爱花, 女, 1985年生, 博士, 副研究员, 研究方向: 医学科技情报研究。
 杨雪梅, 女, 1995年生, 硕士, 馆员, 研究方向: 文本挖掘与知识发现。
 唐小利, 女, 1966年生, 硕士, 研究馆员, 通信作者, 研究方向: 科学计量学, E-mail: tang.xl@imicams.ac.cn。

Research on the Subject Mapping Methods Optimization of Multi-Sources Biomedical Literatures

LI XiaoYing LIU Yi LI AiHua YANG XueMei TANG XiaoLi

(Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100020, P. R. China)

Abstract: Given the inconsistency and the difficulty in the subject attribution of dominant biomedical discipline classifications, outstanding discipline classifications and well-known mapping methods are first investigated, the semantic mapping principle and fusion method integrating automatic semantic matching, document classification data validation, and manual review are then proposed especially for biomedical discipline classifications. Diverse mappings are established between Web of Science, Scopus, and Fields of Research categories and biomedical items of *Catalog of Postgraduate Education Disciplines and Specialities (2022)* designed by Ministry of Education separately. Practical explorations including the mapping of domestic and international neoplasm classifications are also carried out. The research achievements can promote scientific research management and academic analysis work such as centralized retrieval, classified organization, and statistical evaluation of international academic achievements, and provide decision-making basis and information support for the adjustment of disciplines and professional catalogs, talent cultivation, and scientific research layout in China.

Keywords: Semantic Mapping; Subject Attribution; Discipline Classification; Mapping Principle

(责任编辑: 雷雪)