

《中国图书馆分类法》和《美国国会图书馆图书分类法》人工映射分析与差异性探究

童刘奕 张鹏翼

(北京大学信息管理系, 北京 100871)

摘要: 对知识组织系统进行映射, 是解决不同知识组织系统之间互操作的方法之一。本研究以数学、心理学、教育领域为例, 对《中国图书馆分类法》和《美国国会图书馆图书分类法》相关领域类目进行映射, 探讨映射过程中体现的两部分类法的差异性、映射难点和相应的解决办法。在映射过程中, 采用“就上不就下, 就近不就远”的原则, 建立等同、包含于、包含和相关4种映射关系。4种关系各包含类目196条(占24.5%)、423条(占52.9%)、50条(占6.3%)、130条(占16.3%)。影响映射关系类型的主要原因有语言文化差异、社会背景差异、编制原则不同、国内外学科发展水平不同。

关键词: 《中国图书馆分类法》; 《美国国会图书馆图书分类法》; 人工映射; 映射关系类型

中图分类号: G254.1

DOI: 10.3772/j.issn.1673-2286.2018.03.010

知识组织系统是对知识结构进行表达和有组织阐述的语义工具统称^[1], 包括术语表、叙词表、字词典、地名辞典、本体等^[2]。不同知识组织系统由于其组织方式不同且用户难以在一个系统获取所有信息, 而面临在多个系统进行多次检索的困扰, 因此, 如何在各知识组织系统间实现互操作, 成为提高用户检索效率的重要内容。Zeng等^[3]指出, 知识组织系统互操作的实现是当前网络环境下不可避免的问题。实现不同知识组织系统的互操作有多种办法, 其中映射是将两个或多个现有知识组织体系中的概念通过映射建立连接, 从而实现知识组织体系的互操作^[4], 也是最直接的方法之一。

《中国图书馆分类法》是1949年后编制出版的按照一定思想观点, 以科学分类为基础, 结合图书资料的内容和特点, 且分门别类的分类表^[5]。《美国国会图书馆图书分类法》是美国国会图书馆在19世纪末到20世纪初编制的大型综合性分类法, 该分类法将知识分为21个基本大类, 每个大类以单个字母作为标识^[6]。这两种分类法是世界范围内使用广泛的图书分类方法。要同时满足国内外学者的信息资源需求, 就要实现对《中国图书馆分类法》和《美国国会图书馆图书分类法》两

个知识组织系统间的互操作。

本研究以《中国图书馆分类法》的3个二级类目(学科领域) B84心理学、G4教育、O1数学为例, 对这两部分类法的相关类目进行映射, 研究两部分类法部分类目之间的映射关系, 分析映射过程中体现的两部分类法差异性与映射的难点, 并探讨相应的解决办法。目前, 仅有对《中国图书馆分类法》和《杜威十进分类法》(Dewey Decimal Classification, DDC)、《中国图书馆分类法》和《国际专利分类法》的映射研究, 尚未有研究者对《中国图书馆分类法》和《美国国会图书馆图书分类法》进行映射研究, 映射过程中将会遇到何种问题和困难同样值得探索。

本研究在实践方面, 能实现《中国图书馆分类法》与《美国国会图书馆图书分类法》在心理学、教育、数学领域类目的人工映射, 且映射准确率较高, 从而为进一步实现两部分类法的互操作提供数据基础; 在理论方面, 探讨《中国图书馆分类法》与《美国国会图书馆图书分类法》在类目映射上可能遇到的部分问题及其解决方法, 为更多研究者在研究分类法映射时提供理论参考与依据。

1 文献综述

1.1 知识组织互操作

关于知识组织互操作,司莉^[7]总结了实现知识组织系统互操作的9种方法,即派生法、翻译法、系列化分类表或词表、卫星词表法、直接映射、共现映射、中介词典、通过连接数据库实现和宏词汇。如OCLC在《美国国会图书馆图书分类法》和DDC之间使用直接映射法进行一对一的相互映射^[8],目前除D、J、K类尚未完成外,其他部分都已经完成映射,并以表格形式呈现;直接映射法在国内已经使用,黄筱玲^[9]对《中国图书馆分类法》与《中国科学院图书馆图书分类法》的全部类目进行对比和研究,发现两者间有约2/3的类目可形成一一对应的关系。

1.2 映射方法

从建立映射的方法看,映射可分为直接映射和共现映射;根据知识组织系统映射关系建立过程中人力的参与程度,可分为人工映射和自动映射^[10]。直接映射指不同的词汇与分类号建立对等的关系,目前已实现的不少分类法映射项目都采用直接映射法。如《美国国会图书馆图书分类法》与DDC的相互映射,OCLC在两部分分类法的各类目间都直接建立一一对应的关系^[10]。共现映射指在同一元数据方案或书目记录中同时出现的不同知识组织系统术语间建立映射关系^[11]。1994年,由OCLC发起的从LCSH到DDC的映射^[11]使用DDC电子版进行了约9万次共现映射。

人工映射和自动映射是从映射的主体角度进行划分。人工映射的主体是人,需要投入大量人力和时间对分类法的每个类目进行比对和映射。Zins等^[12]以“10-pillar of knowledge”模型为基准,将《美国国会图书馆图书分类法》、DDC、《通用十进制分类法》三部分分类法的一级大类、二级大类映射到该知识分类法,以论证当前使用的三部分分类法是否有足够的系统性,能否覆盖人类的所有知识。自动映射以计算机为映射主体,基于语义相似度、交叉浏览等方法设计映射算法,对类目进行自动映射。如Vizine-Goetz等^[11]将主题词表以MARC21格式进行编码后,使用自动映射法以ERIC和LCSH相互映射为例,论证该算法的可行性。

1.3 映射关系类型和影响因素

首先,在进行类目映射前要了解映射的影响因素。1990年,崔明爱^[13]从分类标准及其引用次序、分类法类型、列类方法、语言标记,分类法的倾向性、动态性、标引规则和复合主题等多个方面总结分类法兼容过程中可能遇到的问题。李珂等^[14]将影响因素分为宏观因素和微观因素,宏观因素包括编制原则、体系结构、语言文化等,微观因素包括先组程度、类目专指度和表述差异。

其次,要对映射的关系类型进行一定了解。分类法类目之间表达概念的不完全一致性,决定类目之间的映射存在多种语义对应关系^[15]。如Renardus项目在进行DDC与本地网关分类法映射过程中,规定完全等同、包含、包含于、大部分重叠和小部分重叠5种对应关系^[16];戴剑波等^[15]在进行《中国图书馆分类法》和DDC映射实践时,规定等价、包含、包含于和相关4种语义映射关系;此外,也有研究在进行映射的过程中不对关系类型进行定义,通过定义相似度的计算方法,找到两部分分类法之间相似度最高的类目直接进行一一对应^[16-17]。

最后,对类目之间进行映射的判断依据也需要参考。贾君枝等^[18]提出要依据类名、注释、主题词、类目间关系等因素确定映射关系类型,而不能仅凭借上述某一项就直接确定类目的映射关系类型,并对具体情况进行详细分析。对于组合类目(如《中国图书馆分类法》中O15代数、数论、组合理论类目),贾君枝等^[19]也针对类目完全等同、类目交叉、类目包含3种情况进行具体讨论,并对每种情况给出相应处理办法。

2 研究方法

2.1 抽样类目情况

根据美国国会图书馆官方网站提供的《美国国会图书馆图书分类法》全文^[20]和国家图书馆网站提供的《中国图书馆分类法》^[21],本文对数学、心理学、教育学3个学科共计799条《中国图书馆分类法》类目进行人工映射。

O1数学为二级类,该二级类的下位类(不含O1)共237条,其中三级类目17条(约占比7.2%)、四级类目(约占比26.2%)62条、五级类目158条(约占比66.7%)。

数学领域大部分三级类目及其下位类的数量分布

呈金字塔形结构,部分类目下位类较多,专指性较强,在映射过程中找到等同关系的映射概率较大。此外,还存在类目设置不均衡的情况。如O10数学理论等三级类目没有设置下位类,导致语义不明确,确定涵盖范围较难。另外,赵冬梅^[22]指出,为体现国内学者对数学学科体系构建的认识,《中国图书馆分类法》设计了体现中国特色的特殊性类目(如古典数学、初等数学、高等数学等),这些类目在结构上与《美国国会图书馆图书分类法》的类目设置差异较大,在映射过程中需要特殊考虑。

B84心理学为二级类,该二级类的下位类(不含B84)共71条,其中三级类目9条(约占比12.7%)、四级类目45条(约占比63.4%)、五级类目17条(约占比23.9%)。

心理学领域和数学领域有较大区别,9条三级类目中只有3条细分到五级类目。类目设置不均衡与学科历史发展有关,19世纪末,心理学才成为一门独立的学科,因此,该学科领域的进一步划分不如数学领域细致。此外,心理学的五级类目较少,专指性较低,难以揭示细小专深的主题。

G4教育为二级类,该二级类的下位类(不含G4)共488条,其中三级类目30条(约占比6.2%)、四级类目232条(约占比47.5%)、五级类目226条(约占比46.3%)。

教育领域是本次映射研究中类目最多、类目设置最复杂的二级类目。三级类目中,G5按照世界各国教育事业进行划分,G6按照各级教育进行划分,G7按照各类教育进行划分。划分原则不同,导致各类目之间有大量重叠部分。如四级类目中的“教育政策”“教学理论”“教学管理”等概念多次重复出现在三级类目的下位类中,使这部分类目复杂冗长,寻找映射关系较难。

2.2 映射方法

本文采用直接映射和人工映射的方法对两部分分类法的相关类目进行映射。本研究的映射为单向映射,即将《中国图书馆分类法》单向映射到《美国国会图书馆图书分类法》上。

本研究将映射关系类型定义为等同、包含于、包含、相关4种。采取的映射关系判断依据是以概念涵盖范围为主,以类名语义关系为辅,结合相应注释进行判断。在判断过程中,优先寻找与《中国图书馆分类法》

类目在映射上存在等同关系的《美国国会图书馆图书分类法》类目,“就近不就远,就上不就下”(即如果有含义更加接近的上位类,则优先选取上位类),允许一对多的关系(即在找不到一对一映射的情况下,如果《中国图书馆分类法》中某一类包含《美国国会图书馆图书分类法》的多个类目,则取多个类目,属于包含关系),但将一对多关系限定在3个以内,等同关系的判定只需语义和涵盖范围基本相当即可。此外,若复分类目所指不明确,找不到对应类,则直接归入相关的上位类。

3 类目映射结果

3.1 数学领域

《中国图书馆分类法》数学领域的238条(含O1数学)类目中有108条(占比45.4%)类目的映射关系为等同,86条(占比36.1%)类目的映射关系类型为包含于,21条(占比8.8%)类目的映射关系为包含,23条(占比9.7%)类目的映射关系为相关。

许多专有数学名词与概念,在《中国图书馆分类法》和《美国国会图书馆图书分类法》中的语义和概念范围一致,能快速建立等同映射关系。

根据“就上不就下”的原则,包含于关系类型远多于包含。映射关系类型为包含于的类目都是五级类目,因为《中国图书馆分类法》数学领域的类目一般仅细分到五级。此外,由于《中国图书馆分类法》所表达的部分概念在《美国国会图书馆图书分类法》中不再进行细分,其下位类只能选择包含于的映射关系类型与《美国国会图书馆图书分类法》中的上位类概念建立映射。如《美国国会图书馆图书分类法》中的QA248-248.7类,包括集合论和算术基础等概念,但该类不存在下位类,因此,《中国图书馆分类法》中的O144集合论的下位类O144.1基本概念、O144.3公理集合论等类目,均只能以包含于的关系与其建立映射。

建立包含关系的类目,大部分是组合类目。如《中国图书馆分类法》的三级类目O1-6数学参考工具书拥有两个四级下位类O1-64数学表和O1-8计算工具,分别与《美国国会图书馆图书分类法》的QA47-59 Tables和QA47-59 Tables建立等同映射关系;结合“就近不就远”原则,该三级类目与《美国国会图书馆图书分类法》的QA47-59 Tables和QA47-59 Tables建立包含的映射关系。除组合类目外,还有概念本身的包含关

系,但无合适的上位类情况。如CLCO159模糊数学就与《美国国会图书馆图书分类法》中最接近的下位类QA248.5 Fuzzy sets建立包含的映射关系。

建立相关关系的类目大部分是表达概念范围有交错,且无法在《美国国会图书馆图书分类法》相应类的上下位类中找到能建立包含或包含于映射关系的类目。这种情况大部分是由于《美国国会图书馆图书分类法》的相关类目是组合类目,且不再进行细分。如《中国图书馆分类法》的O123.5轨迹与几何作图和《美国国会图书馆图书分类法》的QA464 Practical geometry. Geometrical drawing,两者的几何作图部分是等同映射关系,但在《美国国会图书馆图书分类法》中该类目不再有以下位类,因此只能建立相关映射关系。

3.2 心理学领域

在进行映射的《中国图书馆分类法》心理学领域的72条(含B84心理学)类目中有22条(占比30.6%)类目的映射关系为等同,39条(占比54.2%)类目的映射关系为包含于,5条(占比6.9%)类目的映射关系为包含,6条(占比8.3%)类目的映射关系为相关。

等同映射关系的建立大多集中在三级类目、四级类目。心理学领域包含于映射关系比例多于数学领域,建立包含关系的类目大部分是组合类目,原因与数学领域类似。

建立相关关系的类目除组合类目前,还存在其他情况。如《中国图书馆分类法》中的三级类目B84-0心理学理论与《美国国会图书馆图书分类法》的BF38-64 Theory. Scope. Relations,从语义看,两者应建立包含于关系,但观察两个类目的下位类会发现,其概念范围存在部分交叉,不属于包含于的关系,因此,两个类目间仅能建立相关的映射关系。

3.3 教育领域

在进行映射的《中国图书馆分类法》教育领域的489条(含G4教育)类目中有66条(约占比13.5%)类目的映射关系为等同,298条(约占比60.9%)类目的映射关系为包含于,24条(约占比4.9%)类目的映射关系为包含,101条(约占比20.7%)类目的映射关系为相关。

《中国图书馆分类法》教育领域的等同关系类型比例在三个领域中最低,说明在教育领域能建立等同

映射关系的类目很少。因为教育领域的四级类目相似度过高,使每个三级类目的下位类结构和概念都有很多重叠部分,这种重叠导致能建立等同映射的四级类目极少。但在四级类目概念未重叠的部分,其专指性较高,如《中国图书馆分类法》的G432广播、电视教学能与《美国国会图书馆图书分类法》中LC6571-6581 Radio and television extension courses. Instruction by radio and television建立等同关系的映射。

教育领域建立的包含于映射关系类型比例高于其他两个领域(60.9%),这与教育领域复杂冗余的四级类目有关。在多个三级类目下出现“思想品德教育”“教学理论、教学法”“教学研究与改革”的四级下位类。当出现类似的概念,均采用“就上不就下”的原则,统一归入《中国图书馆分类法》中上位类所对应的《美国国会图书馆图书分类法》类目中。如果该四级类目前上位类与《美国国会图书馆图书分类法》相应类目建立的是等同关系,则该四级类与《美国国会图书馆图书分类法》相应类目建立包含于关系;如果该四级类目的上位类与《美国国会图书馆图书分类法》相应类目建立的是相关关系,则该四级类与《美国国会图书馆图书分类法》相应类目也建立相关关系。建立包含关系的类目大部分为组合类目,建立相关关系的类目主要是《中国图书馆分类法》中上位类已经与《美国国会图书馆图书分类法》相应类目建立相关关系。

在教育领域中,值得注意的还有复分类目的映射。

《中国图书馆分类法》中G51-57各地区教育事业依据地区复分表复分,相对应的,《美国国会图书馆图书分类法》中同样存在地区复分表,只是复分的具体细节和《中国图书馆分类法》不一样。《美国国会图书馆图书分类法》相应类目若详细列出各地区,则将地区进行对应;如果没有列出各个地区,则直接与《美国国会图书馆图书分类法》中的上位类进行映射,大多数建立包含于关系的映射。

4 结论

在数学、心理学、教育领域,《中国图书馆分类法》与《美国国会图书馆图书分类法》类目映射关系的建立与分析研究可以提高外文资源批量编目的工作效率,有利于实现对全球资源的揭示、组织和共享,为图书情报机构的工作人员、相应学科领域的信息组织者和科研人员提供参考。同时,也可为《中国图书馆分类法》的

修订完善提供参考。

本文对《中国图书馆分类法》中数学、心理学、教育3个领域共799个类目建立与《美国国会图书馆图书分类法》相应类目的映射关系,其中196条(占比24.5%)类目建立等同映射关系,423条(占比52.9%)类目建立包含于映射关系,50条(占比6.3%)类目建立包含关系,130条(占比16.3%)类目建立相关映射关系。

分析类目映射的结果发现,《中国图书馆分类法》与《美国国会图书馆图书分类法》的各级类目在大部分情况下无法建立等同关系映射,只能建立包含或包含于的映射,在某些学科领域,只能建立起相关映射。出现这种现象的原因很复杂,总结起来有以下4点。

(1) 语言文化背景不同。Lee^[23]在研究中表示分类法能有效地展示提出该分类法的主流文化,因而文化背景不同导致分类法差异较大。同时,文化背景不同,导致社会学科、人文学科有大量概念难以对应。

(2) 社会背景差异较大。中国和美国在社会背景、人口构成、经济发展等方面差异较大,从教育领域看,两部分分类法体现出极大的社会背景差异性。如《中国图书馆分类法》中G473.3学籍,在《美国国会图书馆图书分类法》没有任何相关概念。

(3) 编制原则不同。《中国图书馆分类法》在编制过程中符合科学性原则,以科学分类为基础,采取从总到分,从一般到具体的逻辑系统,同时要考虑图书资料分类的特点;而《美国国会图书馆图书分类法》是根据美国国会图书馆藏书进行分类需要进行编制,造成两部分分类法在结构和细分程度上的差异性,从而使映射建立困难。

(4) 国内外学科发展水平不同。国内外的学者对该学科某些概念的涵盖范围认知不同,导致编制过程中类目设置的差异。如《中国图书馆分类法》数学领域的三级类目O22运筹学,在《美国国会图书馆图书分类法》中无法找到相对应的类目。《美国国会图书馆图书分类法》的运筹学类目被归入T科技总论中,因为中美学者对运筹学的学科归属理解不同,中国学者更倾向于将运筹学归入数学领域,而美国学者倾向于将其归入工程、管理领域。

本文的局限在于进行映射的类目有限。在《中国图书馆分类法》5大基本部类、22个一级大类中,仅各在3个一级大类中选取了1个二级类目及其下位类进行映射,采样数量过少。无法映射指《中国图书馆分类法》中存在,但在《美国国会图书馆图书分类法》中无法找

到等同、包含于、包含或相关的类目。对于无法映射概念,本文采取“就上不就下”的原则,直接与其上位类所对应的《美国国会图书馆图书分类法》类目建立相应的映射关系。这种简单的处理方法会使这种情况与普通的包含于映射关系混淆,影响对结果的解读。

另外,本研究在方法上仍存在一定不足。虽然人工映射建立在详细分析类语义和概念范围的基础上,但一方面由于学科知识的专业性和主观理解的局限性,有可能使参与映射的人员对某些类目理解不当,导致建立错误的映射关系;另一方面由于人工映射的工作量较大,而映射关系的建立是一项综合性的智力劳动,在映射过程中难免出现疏漏。研究的下一步是进行实证检验,即引入自动映射作为参考,通过比对自动映射与人工映射的结果,以检验研究成果的准确性。

参考文献

- [1] HILL L L, BUCHEL O A, JANEE G, et al. Integration of knowledge organization systems into digital library architectures [J]. 2002, 43 (7) : 331-335.
- [2] Systems of knowledge organization for digital libraries: beyond traditional authority files [EB/OL]. [2017-05-12]. <https://www.clir.org/pubs/reports/pub91/#1>.
- [3] ZENG M L, CHAN L M. Trends and issues in establishing interoperability among knowledge organization systems [J]. Journal of the Association for Information Science and Technology, 2004, 55 (5) : 377-395.
- [4] 宋文. 知识组织体系语义互操作研究 [J]. 图书馆论坛, 2012, 32 (6) : 117-121.
- [5] 马张华. 信息组织: 第3版 [M]. 北京: 清华大学出版社, 2008: 137-139.
- [6] Library of Congress Classification [EB/OL]. [2017-05-12]. <http://www.loc.gov/catdir/cpsolcc.html>.
- [7] 司莉. 知识组织系统的互操作及其实现 [J]. 现代图书情报技术, 2007, 2 (3) : 29-34.
- [8] Map LC (LCC) to Dewey (DDC) Classification [EB/OL]. [2017-05-12]. http://www.questionpoint.org/crs/html/help/zs/ask/ask_map_lccddc.html.
- [9] 黄筱玲. 高校图书馆文献分类工作改革的思考——兼谈《中图法》和《科图法》的兼容 [J]. 图书馆学研究, 2001 (4) : 29-31.
- [10] 刘晓鹏, 真漆, 于洋. 基于统计的知识组织系统自动映射方法研究 [J]. 数字图书馆论坛, 2009 (12) : 75-78.

- [11] VIZINE-GOETZ D, HICKEY C, HOUGHTON A, et al. Vocabulary mapping for terminology services [J/OL]. <https://journals.tdl.org/jodi/index.php/jodi/article/view/114/113>.
- [12] ZINS C, SANTOS P L, CIDA L V A C. Mapping the knowledge covered by library classification systems [J]. *Journal of the Association for Information Science and Technology*, 2011, 62 (5) : 877-901.
- [13] 崔明爱. 图书分类法兼容理论问题探讨 [J]. *图书馆建设*, 1990 (1) : 32-35.
- [14] 李珂, 宋文. 分类法映射研究 [J]. *图书馆杂志*, 2014, 33 (12) : 49-56.
- [15] 戴剑波, 侯汉清. 图书分类法映射系统设计原理——以《中国图书馆分类法》和《杜威十进分类法》为例 [J]. *情报学报*, 2005, 24 (3) : 299-303.
- [16] HEERY R, CARPENTER L, DAY M. Renardus project developments and the wider digital library context [J]. *D-Lib Magazine*, 2001, 7 (4) : 1082-9873.
- [17] VIZINE-GOETZ D. Popular LCSH with Dewey numbers: subject headings for everyone [J]. *Journal of Library Administration*, 2001, 34 (3/4) : 293-300.
- [18] 贾君枝, 郝倩倩. DDC到《中图法》类目映射方法研究 [J]. *中国图书馆学报*, 2013, 39 (1) : 43-50.
- [19] 贾君枝, 郝倩倩. DDC与《中图法》组合类目映射探讨 [J]. *中国图书馆学报*, 2012, 38 (4) : 63-70.
- [20] *Library of Congress Classification PDF Files* [EB/OL]. [2017-05-12]. <http://www.loc.gov/aba/publications/FreeLCC/freelcc.html>.
- [21] 《中国图书馆分类法》编委会. 中国图书馆分类法 (第五版) Web版 [EB/OL]. [2017-05-12]. <http://clc5.nlc.cn/login.aspx>.
- [22] 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.
- [23] LEE W C. Culture and classification: an introduction to thinking about ethical issues of adopting global classification standards to local environments [J]. *Knowledge Organization*, 2015, 42 (5) : 302-307.

作者简介

童刘奕, 男, 1995年生, 硕士研究生, 研究方向: 信息组织、用户研究, E-mail: tly124@pku.edu.cn。

张鹏翼, 女, 1981年生, 博士, 副教授, 研究方向: 信息与知识组织、信息行为与认知, E-mail: pengyi@pku.edu.cn。

Research on the Manual Mapping and Discrepancies of *Chinese Library Classification* and *Library of Congress Classification*

TONG LiuYi ZHANG PengYi

(Department of Library and Information Science, Peking University, Beijing 100871, China)

Abstract: Mapping knowledge organization system is one of the methods to solve the interoperation between different knowledge organization systems. In this paper, we mapping *Chinese Library Classification* and *Library of Congress Classification* in the fields of mathematics, psychology and education, and explore the differences between the two classifications, discuss the difficulty of mapping, and explore the solution. During mapping *Chinese Library Classification* and *Library of Congress Classification*, the principle is "narrow-match first, more relevant term first". And there are four match types: exact-match, narrow-match, broad-match and relevant-match. Each of the four relationships contains 196 items (24.5%), 423 items (52.9%), 50 items (6.3%) and 130 items (16.3%). The main factors that influenced the match type are: different cultural and language background, different social backgrounds, different principles when compiling the classifications, the gap of subject development in America and China.

Keywords: *Chinese Library Classification*; *Library of Congress Classification*; Artificial Mapping; Match Types

(收稿日期: 2018-02-01)