大数据环境领域知识组织方法研究*

蒋　勋[1,2]　朱晓峰[2]　肖连杰[3]
([1]无锡环境科学与工程研究中心　江苏　214153；
[2]江苏省数据工程与知识服务重点实验室　南京　210023；　[3]南京大学信息管理学院　江苏　210023)

摘　要: **[目的/意义]**为适应大数据环境下网络资源形式多样化及各领域知识服务需求个性化,本文为揭示知识单元内涵语义、挖掘知识外延关联以及提供丰富灵活的知识服务提供领域知识组织方法。**[方法/过程]**文章针对现阶段大数据呈现出的特质,以面向领域应用的视角设计了知识关联组织、聚类组织及语义组织的方法体系。**[结果/结论]**文章提出的知识组织方法适应了大数据环境知识工程相关技术与领域应用的深度融合与优化创新。
关键词: 知识组织　领域　大数据　方法

Research on Domain Knowledge Organization Method in the Big Data Environment

Jiang Xun[1,2]　Zhu Xiaofeng[2]　Xiao Lianjie[2]
([1] Wuxi Environmental Science and Engineering Research Center, Jiangsu, 214153;
[2] Jiangsu Key Laboratory of Data Engineering and Knowledge Service, Nanjing, 210023;
[3] School of Information Management, Nanjing University , Jiangsu, 210023)

Abstract: [**Purpose /significance**] This research is put forward to adapt to the diversified forms of network resources and the personalized needs of knowledge services in various fields in the big data environment. A method is provided in this paper, domain knowledge organization for revealing the intension and semantics of knowledge units, mining the denotation association of knowledge and providing rich and flexible knowledge service.[**Method / process**] The research work proposed in this paper is based on the characteristics of big data at the present stage. It is also designed a series of methods of knowledge organization from the perspective of domain application, including association organization, clustering organization and semantic organization. [**Result/conclusion**] The method proposed in this paper adapts to the deep integration and optimization innovation of the related technology and field application of big data environmental knowledge engineering.
Keywords: knowledge organization　domain　big data　methods

1　引言

大数据环境下资源对象类型呈现多样化,其组织形式由文件为核心向数据为中心转化,数据和数据之间通过富含语义链接的形式构成了蕴含价值的数据网络。知识组织是关于数据结构的设计、知识内容的记录以及知识集合整序的过程,使之便于揭示知识单元,方便知识发现,能为用户提供有效的知识服务[1]。知识

*本文系国家社会科学基金重点项目“大数据环境下领域知识加工与组织模式研究”(项目编号:20ATQ006)的研究成果之一。

的产生和应用具有特定的领域背景及环境，即不同领域知识具有不同的应用情境，针对不同情境会产生特定的领域知识组织需求。

大数据既为领域应用提供了便利，也为便利性的实现提出了难题，需要采取不同的知识组织方法。元数据、本体、关联数据的提出与应用为面向领域大数据构建多层次、多角度、细粒度的知识组织方法提供了技术路线。知识组织方法不断丰富且功能趋于统一，为了探寻数据中隐藏的规律、深入分析数据内涵知识，可以采用关联技术对知识进行组织；为了进行聚类分析或分类知识检索，聚类组织方法是一个有效途径；实现推理检索或相关知识服务，需要数据间具有语义关系，这可以借助知识的语义组织方法来完成。

2 相关研究工作

2.1 知识关联组织法研究

以网络为平台形成知识间的关联，一直是情报学学科的研究阵地。早期Garfield在*Science*杂志上揭示了引文网络对科学论文中学科知识传承规律[2]，在这一方面国内学者基于科学引证网络揭示出科学知识演化的动力及规律是探究科学知识创造及发展过程的关键[3]，典型的应用是我国中文社会科学引文索引（CSSCI）研制成功并从引文索引数据方面凸显了学科研究特征[4]，而无论是引文网络还是信息网络，知识的传播与交流都与知识关联网络密不可分[5]。在领域知识组织中，相关研究从知识关联视角研究金融领域知识表示与识别[6]，也有针对非物质文化遗产领域设计基于关联数据的知识本体模型[7]，利用关联知识为博物馆馆藏资源实现语义交互等[8]，在工程技术领域基于知识的关联性集结形成主题相关的知识推荐方法[9-10]。

2.2 知识聚类组织法研究

在知识服务领域中知识聚类组织是常见方法，如根据关键词承载语义进行分类实现研究主题识别[11]，由词汇粒度不同刻画不同知识实体[12]，利用自动分类器实现科技文献中主题抽取等[13]。在海量的网络资源的知识组织中，以细粒度聚合的方法是网络资源有效的知识组织方法[14]，在这种方法中最常用的实现算法是K-means算法，如在制造加工领域对关键工序的知识工程进行聚类分析，解决关键工序的质量特征数据不足[15]。在大数据环境下对用户历史行为数据、兴趣偏好数据的聚类分析能更精确地推送相关产品服务等条目[16]，有助形成精准、按需的知识服务[17]。面向大数据特征的知识组织，以自然语言描述的服务文档为对象，领域知识的服务聚类为互联网的服务精准推荐提供重要的理论与实际应用支持[18]。

2.3 知识语义组织法研究

知识的语义组织法在各领域中有着许多广泛的应用，如文化遗产领域语义组织的研究解决了传统文化中知识表达体系的自动构建，越来越多的文化遗产项目选择元数据及本体技术进行知识建模[19]。文化遗产领域中法国创建的建筑和移动文物叙词表[20]及欧洲遗产多语言叙词表[21]，以及我国学者针对敦煌壁画叙词表编制研究语义组织过程[22]等，以特定的结构排列显示词汇间关系。在应急管理领域研究中利用本体技术对火灾应急管理进行知识抽取、组织与表示，构建相关本体模型与应急知识库[23]。在不同领域中知识语义组织方法如词表、元数据等能准确反映出语义含义的特征项[24]，词表通过数据添加明确语义，推动了结构化知识组织网络的形成[25]，基于元数据构建了的概念描述与词汇描述[26]。

2.4 述评

在知识组织的基础理论、关键技术、工具方法和知识表示等方面已经取得了很多成果，有效推动了图书情报工作从传统文献信息服务到知识信息服务的转变。也有很多成果探讨了领域知识的服务领域、可能的手段方法，关联组织、聚类组织、语义组织等领域知识组织也有所涌现。这些成果为进行大数据环境下的知识组织的深度探讨提供了深厚基础，为凝聚大数据环境下领域知识的加工、处理和组织的基础理论、技术手段、方法体系提供了有效思路。为了帮助和促进各学术领域的深入研究，尚需要有一套适合于领域知识的科学组织方法体系支撑大数据环境下领域知识加工与组织模式研究。

3 面向领域应用的知识关联组织法

知识关联组织法使原本无序的知识变为易于控制和有序，将原本孤立的数据呈现出有机的联系。例如，阅读中遇到不能理解的专有名词或相关知识时，就希望能够直接阅读到相关知识，知识关联组织法为实现这样的阅读提供了有效地手段，我们可以将文本中的专有名词与知识库中相关知识关联起来，并通过一定的组织结构实现这种连接，为提供知识点知识服务奠定基础。

无论是传统分类法、叙词表，还是领域本体技术，都不是知识的简单排列组合或堆积，而是建立知识概念之间的逻辑关联。知识关联关系，抽象地说是知识概念之间的逻辑关系，在大数据环境下各领域呈现出数据类型复杂、种类繁多、来源广泛以及数据割据分散等特征，需要数据经过关联使其价值得到升华，原本松散的数据在关联的作用下，可能上升为非常有价值的信息或知识，成为知识服务的有效知识源。知识关联就是建立知识之间存在的联系（链接），着重强调揭示知识间的关系性质或类别，知识关联组织是将知识及相互联系关系有效地存储于数据库中。因此，知识关联组织法就是对数据、信息、知识的关系的建立、揭示和组织所采取的一系列技术方法和组织过程。

从领域知识组织出发，进一步将知识关联组织法细分为分类关联、时空关联、统计关联及事件关联。

3.1 知识的分类关联

分类关联是领域大数据的基本特征，用于表达领域中各概念间的层次结构[27]。不同的层级结构与网络分布为不同用户从爆发式的海量知识资源中筛选出需要的知识条目提供了基础。领域知识组织中分类关联组织法进一步细分两类：其一是由事件、人物、领域的行为活动、组织等形成的聚集关系，如工程领域设计者、施工者、管理者等因生产制造活动归属为聚集关系；其二是由归属不同类别知识但由事件驱动而关联，各类别知识归属的聚集关系导致关联传播扩展，如图1所示。第一类分类关联中以医疗系统领域为例，医生、医院、病人因医疗救治归属一类；第二类分类关联中以应急救援领域为例，疫情防控直接关联着医疗救治，因应急事件将医疗系统与应急管理等领域关联。

3.2 知识的时空关联

时空关联基于知识的时间与空间二维特性将领域知识精确刻画事物结构与状态的有效关联与演化过程。如3.1节中所举例的应急管理领域知识体现了时空的二维特性。时间维度方面，应急管理领域知识在时间上具有记忆性与延续性，通过追溯历史信息能复盘应急响应全程，根据知识在时间序列的演变规律预测领域知识所描述事件的可能趋势。空间维度方面，领域知识描述了应急管理领域空间特征相关信息，由地质地貌、交通位置、建筑用地等构建的物理层面空间维度，以及由政治法律、风土人情、经济环境等构建的人文空间维度。领域大数据所蕴含知识的时空关联性刻画了大数据动态演化的特征，如应急管理领域中针对火灾救援就需要密切结合火灾时间（上班上学高峰、深夜时刻）、建筑用地（学校等人口密集区、森林易燃区）等时空关联性因素采取不同的应急对策。

3.3 知识的统计关联

无论是知识的分类关联亦或时空关联都基于知识间可抽取的逻辑关联，而知识的统计关联是由事件发生后由统计方法解析出的可能存在的关联关系，知识的统计关联无法基于概念表示直接组织形成。如经典的购物车案例中“尿布”与“啤酒”的关联“神话”，由分类关联、时空关联都无法逻辑解释其相关性，只有通过数据仓库将超市各门店的详细原始交易数据进行分析和挖掘，才能形成统计学意义上的关联，为超市商品陈列布局提供了非常有价值的参考。各领域所积累的海量行业数据类型复杂、来源广泛，简单的抽样分析与因果假设无法应对大数据分析需求，必须用统计手段从大数据中发现相关性。基于知识的统计关联，我们能更容易、更清晰地分析出事物发展中隐藏的关联规律。

3.4 知识的事件关联

在不同的领域中都有各自规律性的知识集合。如知识密集型制造行业其需要的知识服务通常具有学科融合、领域交叉特点，其知识来源主要是经验案例知识、规范知识、专利知识；应急管理领域知识更多由常识知识、经验知识与推理知识组成。这些知识集合基于领域的各类事件过程形成关联关系，体现了学科的融合性与领域的交叉性，呈现出事件演化发展的动态关联。例如公众在掌握应急常识知识基础后，进一步希望获取相关的完整的案例描述；应急救援在已有经验知识背景下实施救援行动，同时需要补充施救地区的地形、地貌、风俗等知识；决策层既需要与公众互动，辟出谣言猜想、引导正确舆论，又需要实时地关联环境变化、医疗资源分布、物资匹配信息[28]。

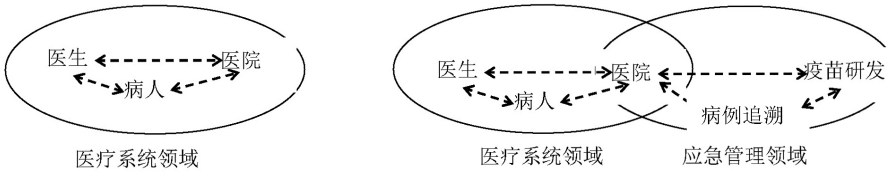

图1　知识分类关联示意图

4 面向领域应用的知识聚类组织法

随着领域知识的积累与应用,一部分在学科领域、实践应用领域等处于重要地位的知识关联从其他的知识关联中脱颖而出,形成知识聚类。领域知识发展过程中知识关联的“富者更富”是知识聚类的根源。知识聚类组织实际上也是知识关联组织的一种,是其内涵的发展与衍生。它与上文所述关联不同在于,知识聚类不局限于依赖相互关系联系起来,根据领域特征与需要赋予相关类别(号),并借助类别(号)聚集在一起,形成分类有序的数据组织。聚类组织法是根据一定的规则将文献、信息或知识按类聚集起来,分别给予相应的类别标记,并将类号赋予相关信息,存于数据库中的过程、技术与方法。

知识聚类组织法将原本分散的信息或知识按主题特征等聚合在一起,使聚合在一起的信息或知识具有某种共同特征或关联。数据仓库的建立就是有效采用了这一知识组织方法。数据仓库的一个主要特征就是“面向主题”的数据集市,是从多个数据库中将某一主题相关信息汇集到数据仓库中,并通过一定的关联组织于数据仓库中。这种聚类的知识组织为聚类分析与知识服务打下了基础。

在聚类组织信息的过程中,信息的类分层是很重要的,也就是类的上下位关系要体现出来,并将这些关系存储于数据库中,只有这样对信息分析、知识服务才更有价值。

4.1 领域词汇功能划分架构

领域知识组织中词汇粒度是聚类组织的基本单元,以词汇为特定的目标信息进行识别与抽取,赋予词汇领域内特定含义,并使用符号完成信息的固化与表示,形成表达概念的知识单元。

基于知识聚类的需要,将领域词汇功能划分为领域无关词汇功能与领域相关词汇功能,所形成的领域相关词汇功能是领域聚类的空间,如图2所示。其中,领域无关词汇功能表示各领域通用的词汇,任意一个领域的研究问题与研究方法;而领域相关词汇功能囊括了领域适用性的功能词汇。在不同学科领域、实践应用领域可通过定义相关的功能类别,如在医疗系统领域中领域相关词汇功能应包含疾病、诊疗、医护等。

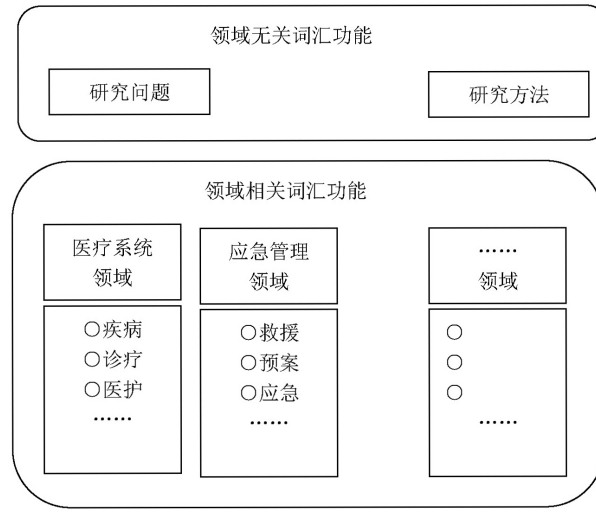

图2 面向领域的词汇功能划分架构

领域相关词汇的生成方法通常有两种:基于预设类别的分类标注方法、基于限定内容的文本生成方法。基于预设类别的分类标注方法是对预先设定好的词汇功能类别进行分类与序列标注,一般采用监督与半监督训练方式实现分类器的模型学习过程,在大数据环境下尤其是自然语言处理(Natural Language Processing,NLP)的成熟推动了分类器进行端到端的深度学习。另一种基于限定内容的文本生成方法,通过约束序列语言模型最终输出文本的内容,实现目标功能词汇的获取。所使用的自动文摘生成策略可分为抽取式与生成式,抽取式自动文摘生成策略是对词句的重要性排序,生成式自动文摘生成策略则是在理解文本语义的基础上实现对原文本的复述。抽取式使用效果通常优于生成式,现阶段基于神经网络的生成式文本摘要方法在众多自然语言处理任务中表现突出。

4.2 领域相关词汇的聚类方法

基于上述领域相关词汇功能,进一步细分各领域中知识聚类的方法。知识粒度刻画越细则表述的准确性越高,知识聚类的内在逻辑性越强。对上述形成的领域相关词汇功能以细粒度聚类的思路进行构建:序化文档信息、构建聚合单元、细化领域概念、构建领域情境,如下页图3所示。

4.2.1 序化文档信息

序化文档信息是聚类的前提,只有理清了文档信息才能辨析概念,才能识别聚合单元,最终才能构建聚合单元。一般的文档信息包括了体裁、单位机构、内

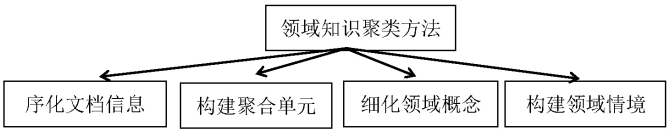

图3　领域相关词汇的聚类方法

容、来源、作者、题名等。

4.2.2　细化领域概念

领域知识组织中知识体系构建的细致程度对文档信息的检索与利用效率会产生较大影响，领域概念细化程度越高，所构建的聚合单元片段越小，其检索利用的相关性与准确性越高，如表1所示的领域概念粒度层级。

表1　领域概念粒度层级

层级	领域概念粒度
K1	主要概念
L2	子概念
Kn	……

领域概念粒度的细化有利于领域知识的准确性，有助于文档信息序化后的准确性与相关性的聚合效率与效用。

领域概念的结构与相互间关系，结合5.1.2节具体论述。

4.2.3　构建领域情境

领域情境定义了知识体系所在应用领域特定的概念环境，从而构建领域情境与聚合单元概念之间的关联关系提供基础，支持序化后文档信息聚合。领域情境的构建可从多个方面进行考虑，如图4所示，不同领域其任务需求不同，在实际构建中可调整。

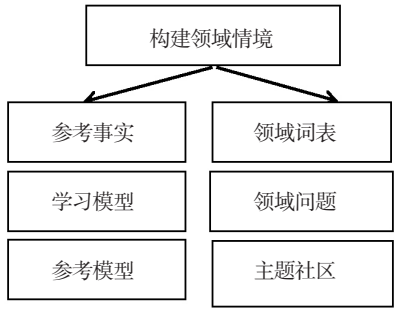

图4　领域情境构建模型

4.2.4　构建聚合单元

在4.1节形成的领域的词汇功能基础上构建聚合单元概念及概念之间关系，一般采取自上而下的方式逐次确定聚合单元的概念及属性，使得每个构建的聚合单元都具有独立的概念。所构建的聚合单元在领域情境下可用性越高，该聚合单元的有用性越强，即反映聚合单元构建的精确性；所构建的聚合单元越多可适用于同一领域情境，该聚合单元的覆盖面就越广，反映了这些聚合单元构建的领域全面性。

如图4中所构建的聚合单元之间的关联关系包括了相似或相近领域概念（即内容相似度）、相似情境下知识被重用（即情境相似度）、解决相关领域问题（即面向主题社区的相关性）。

5　面向领域应用的知识语义组织法

知识语义组织法是根据大数据环境下领域信息资源的特点，以某种方式实现领域资源的序化、规律化和系统化。具体而言，领域知识语义组织就是将信息及信息间的语义关系存储起来，构成具有语义关系的数据库，在检索和分析时，通过建立的语义关系，进行语义推理实现知识服务。知识的语义组织将数据库中原有记录、字段（属性）的关系上升到数据、知识间的语义关系，保证了知识的关联、再生以及隐性知识的呈现。知识组织中的语义关系丰富多彩，例如，主题知识词典中的语词关系有同义、属分、参见等语义关系，人与人之间有各种亲属关系、研究中有合作关系、工作中有上下级之间的关系以及各类人际关系，事物、知识间的语义关系更加复杂。但这些关系在数据或知识之间建立以后，将大大提升信息系统的知识服务能力。

5.1　元数据

元数据是目前领域知识组织中实现文本（非结构化）到数据（结构化）的通用技术方案，其优势是通过语义互操作实现多源数据的异构数据共享、交互与整合，主要包括了元数据扩展与元数据对齐两种技术路线。元数据的基础与关键是元数据标准。由元数据标准与词表层构成了元数据的领域知识语义组织的立体模型，如图5所示。

词表属于上层，是数据网络中的模式层，定义信息描述使用的基本模式、所选描述元素与相应取值类型及范畴。根据不同领域需求、不同数据源特征、标引人员偏好等，所选用的词表差异性较大，通过元数据标准层与词表间的映射解决词表之间的互操作困难。特别针对一些跨学科、跨领域较强的知识组织过程，如面向突发事件的应急决策领域，需要多学科、多部门快速协同，此时由多学科、多行业的多个词表与元数据标准形成映射，相互补充才能完成。

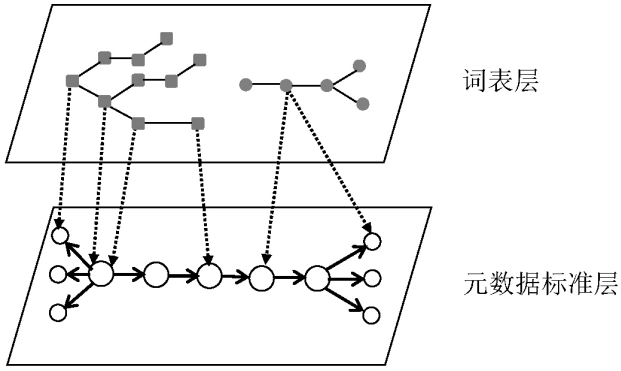

图5 元数据的领域知识语义组织示意图

5.1.1 元数据标准

元数据标准旨在构建一个对数据语义内容能达成共同理解的模式，能实现知识组织中数据对齐与合并，最终提高信息处理标准化与数据交换效率。在元数据标准下生成各种类型的元数据集，定义元数据间的关系及语义、语法、可选值等。在大数据环境下，由数据描述、表示及存储技术等推动，信息处理遵循元数据标准从类元数据集向内容类数据集的转换，具体如表2所示。

表2 元数据标准结构

元数据标准	描述对象	描述工作	元数据集	描述资源
类元数据	文献单元	非结构化数据转换为结构化数据	MARC、DC、EAD等	外部属性（如作者、题名、出版机构等）
内容类数据	知识单元	语义内容特征提取	RDA、DACS、CCO等	概念间关系表示（RDF词汇表）

5.1.2 词表构建

词表是语义网环境下描述与表示领域的概念与关系，通过添加数据明确词表语义，推动结构化数据网络的形成。领域词表定义了领域内术语、概念及相互关系，包括了分类表、名词规范档、叙词表等术语集合。在不同领域根据资源的特征，选择适宜词表，准确理解词表中概念及相互关系，注重描述标引的规范性与一致性。

大数据的特质使得网络资源呈现多样性，驱动着新型词表不断出现，词表作为各领域知识概念的载体，经过生成、发布、维护、修正等过程趋于稳定，如图6所示的应急救援领域的词表承载的概念体系。

词表与概念的关系中，当领域概念跨学科、跨部门，如应急救援领域，词表不足以完整刻画领域概念，需要多个词表互相补充完成领域概念的描述。

词表间关系中尤其在词表构建阶段，重用关系与映射关系最为常见。重用关系通过共享术语、定义交流信息模式实现，目标在于降低词表构建成本并提高彼此间互操作性；映射关系是建立词表间两两对应关系，是跨领域知识组织的基础，映射关系的类型可概括为表3中的三种。

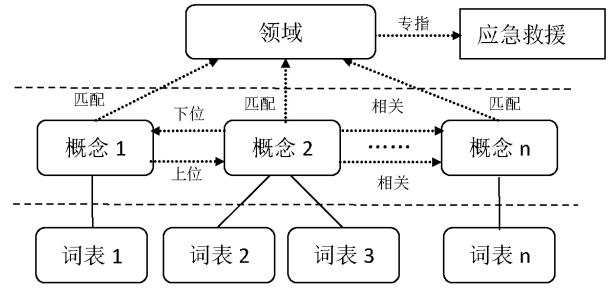

图6 领域词表概念与关系

表3 词表映射关系

映射关系	含义
等同关系	词表间两个术语相同或相近
等级关系	词表间两个术语存在包含关系
相关关系	词表间两个术语有关联性

5.2 关联数据

关联数据技术是领域知识组织中实现互通互联的一种技术，是一种轻量级语义网实现方式，能实现海量异构数据资源的形式化、语义化、关联化及共享化。具体方法是：在领域知识组织中采用关联数据技术在网上发布领域词表等知识组织的方式，进一步与外部开发数据集建立关联关系，使外部数据资源成为关联开放数据的一部分，为领域知识组织提供多元环境。实现过程中关联数据技术更多与开放数据技术集合，以RDF（Resource Description Framework，资源描述框架）为基础，利用OWL与SKOS工具，将不同领域中内容数据转化成标准化、结构化的关联数据。

关联数据优势在大数据的应用环境下得到充分体现，一方面是大数据环境形成的数据网络基于RDF数据模型能实现多源数据各实体的链接；另一方面各个数据网络之间的关联为所聚集的海量数据开发与利用提供了保障。基于上述两点越来越多的领域机构加入关联数据，关联数据成为跨领域、跨部门的数据集成与应用的技术路径。关联数据目前使用最多的是公共文化领域，如图书馆、博物馆等，实现馆藏资源的集成与关联。

5.3 本体技术

本体技术是知识语义组织应用最多的形式之一。本体是对客观存在事物的一个系统的解释或说明，它

关心的是客观现实的抽象本质。也就是说，把对象之间的关系抽象出来，这些抽象出来的关系再通过语义关系表达出来。本体中的语义关系有概念之间的语义关系，即类与类之间的语义关系，如历史事件中的“事件”和“地名”的关系，“事件”和“时间”的关系等；也有客观事物对象之间的语义关系，如“父子关系”、同事关系、上下级关系等。具体到领域本体技术应用，领域本体已构建了更细粒度的应用场景，如应急救援领域本体可细分为应急预案本体、应急组织本体、应急资源本体、应急案例本体。

5.3.1 领域本体构建流程

面向不同领域本体构建其流程相同（如图7），由五个主要步骤构成：一是确定所构建本体的领域与范畴。如明确为应急救援领域中疫情防控工作的本体构建，在这一步骤中需将疫情防控工作的本体覆盖到医疗救治、疫苗研发、疫情输入等要素，需要调用已有公共卫生突发事件领域的应急预案本体等，需要指明各要素与所构建领域本体的关系。

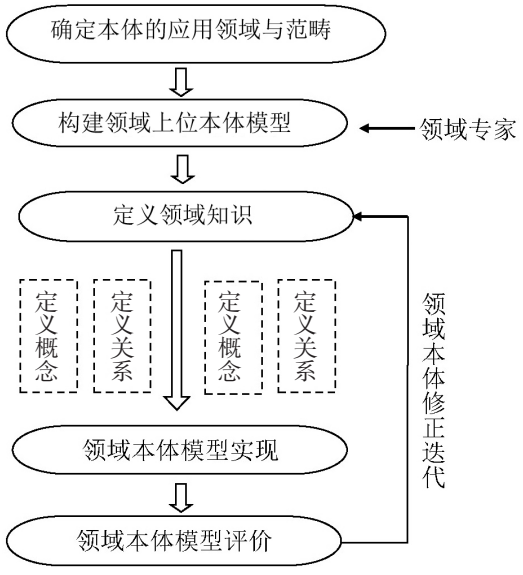

图7　领域本体构建流程

二是基于领域专家介入构建工作，尤其是领域的上位本体，形成可靠、可扩展的领域本体框架。在领域专家的帮助下能促使领域内各本体模型的共享与互操作，并对历史相关本体进行修改，实现本体的重用。

三是定义领域知识以填充所构建的上位本体，形成领域本体的模型。定义的领域知识中包括了概念、关系、属性及实例。

四是本体实现，通常采用OWL作为本体的描述语言。

五是对所构建的本体进行评价、修正、迭代优化，形成领域本体的螺旋式上升。

5.3.2 构建面向领域的上位本体

上位本体也称顶层本体，是抽象层次最高且不依赖特定问题的本体，即上位本体与领域无关，是基础性的概念模型和知识结构。通过重用上层本体方法，一方面建立面向不同领域的概念和关系框架；另一方面，提高本体模型语义操作能力，实现不同领域内本体模型的共享与集成。

著名的上位本体如SUMO（Suggested Upper Merged Ontology，推荐上层合并本体），是由多个上位本体整合形成的超级上位本体，融合了政府、经济、地理等20多个领域。基于SUMO构建面向领域的上位本体，通过提炼相关概念及关系形成*-SUMO。面向领域的*-SUMO一方面能涵盖特定领域本体的主要概念，另一方面能实现与已有本体的重用与互操作。

5.3.3 获取领域本体知识

除了构建领域本体架构之外，另一项重要工作是获取领域本体知识，尤其是核心概念与组成概念的关系。核心概念来源的广泛性决定了领域本体知识的全面性，专业性则决定了领域本体的精准性。以公共卫生应急管理的知识体系为例，涉及疾病预防控制体系知识、公共卫生服务体系知识、重大疫情防控救治体系知识，以及应急物资保障体系知识、国家储备体系知识、应急物资采购供应体系知识。六个方面的知识融合一起，是一项非常复杂的系统工程。首先分析公共卫生应急管理各方面知识体系，在领域专家指导下形成公共卫生应急管理领域的基本概念、概念间关系、概念属性及专业术语等知识；其次根据公共卫生应急管理领域中已有的细粒度概念实例与关系对实体类别进行归类，设计该领域中实体与关系的抽取模型；最后对六个方面获得的概念、关系进行泛化、聚类确定主要概念与关系，填充到公共卫生应急管理领域本体模型中。

6 结语

大数据环境与知识组织理论与方法息息相关，知识组织揭示了数据内容与语义内涵。大数据环境下领域知识组织，不仅需要传承过去知识组织方法，更需要创造和完善知识组织新方法、新策略，具体而言，就是在已有文献组织方法、信息组织方法、通用知识组织方

法的基础上，系统研究并重新构建领域知识组织方法。当然，知识组织技术与方法很多，本文主要针对受大数据影响较大且对知识服务产生明显效果的几种方法进行研究，包括：第一，基于海量数据内涵知识的领域知识关联组织法；第二，基于大数据技术的领域知识聚类组织法；第三，基于语义网、元数据、关联数据等领域知识语义组织法。这三类知识组织方法的水平影响着大数据所提供的数据质量及后期能挖掘的应用价值，大数据的演化也推动着知识组织技术方法的提升与创新。文章剖析了知识组织方法的层次并阐述这些技术方法在一些经典应用领域或场景中的特点，明确了三种方法在当前大数据环境下的特点。面对各领域精细化需求，信息服务不仅转型为知识服务，更将转型为智慧服务，领域知识组织方法在不断延伸拓展。

参考文献

[1] 苏新宁等.面向知识服务的知识组织理论与方法[M].北京：科学出版社，2014.

[2] Garfield Eugene.Citation indexes for science: a new dimension in documentation through association of ideas [J]. Science, 1955,122(3159) :108-111.

[3] 刘 向，马费成.科学知识网络的演化与动力——基于科学引证网络的分析[J].管理科学学报，2012,15(1)：87-94.

[4] 苏新宁.中文社会科学引文索引（CSSCI）的设计与应用价值[J].中国图书馆学报，2012,38(5) :95-102.

[5] 滕广青.基于频度演化的领域知识关联关系涌现[J].中国图书馆学报，2018,44(3) :79-95.

[6] 唐旭丽，马费成，傅维刚等.知识关联视角下的金融知识表示及风险识别[J].情报学报，2019,38(3)：286-298.

[7] 侯西龙，谈国新，庄文杰，等.基于关联数据的非物质文化遗产知识管理研究[J]. 中国图书馆学报，2019,45(2)：88-108.

[8] Giannis Skevakis, Konstantinos Makris, Varvara Kalokyri, et al. Metadata management, interoperability and linked data publishing support for natural history museums[J].International Journal on Digital Libraries，2014 (14)：127-140.

[9] 王临科，蒋祖华，李心雨.面向工程领域的主题多样性知识推荐方法[J].计算机集成制造系统，2021,27(1) :214-227.

[10] Li X , Chen C H, Zheng P, et al. A knowledge graph-aided concept - knowledge approach for evolutionary smart product - service system development[J]. Journal of Mechanical Design, 2020, 142(10):e101403.

[11] 陆 伟，李鹏程，张国标，等.学术文本词汇功能识别——基于BERT向量化表示的关键词自动分类研究[J].情报学报，2020,39(12) :1320-1329.

[12] Ying D, Min S, Han J, et al. Entitymetrics: measuring the impact of entities[J]. Plos One, 2013, 8(8):e71416.

[13] Heffernan Kevin，Teufel Simone. Identifying problems and solutions in scientific text[J]. Scientometrics，2018, 116(2):1367-1382.

[14] 马翠嫦，曹树金.网络学术文档细粒度聚合本体构建研究[J].图书情报工作，2019,63(24)：107-118.

[15] 陈克强，刘伟军，姜兴宇，等.面向多品种小批量制造过程的关键工序识别与聚类分析方法[J].计算机集成制造系统，2021,27(2) :104-126.

[16] 黄 颖，蒋祖华，刘璞凌，等.考虑时序性和动态信任的工程经验知识推荐技术[J].上海交通大学学报，2016,50(9)：1422-1429.

[17] 章成志，王玉琢，王如萍.情报学方法语料库构建[J].科技情报研究，2020,2(1)：30-45.

[18] 赵 一.基于领域知识的服务聚类与个性化推荐方法[D].武汉：武汉大学，2018.

[19] 李章超，何 琳.文化遗产语义组织研究进展[J]. 图书情报工作，2020,64(7) :4-12.

[20] Departement du Systemed' Information de l'Architecture et du Patrimoine. Thesaurus for the Designation of Architecture and Structural Works[EB/OL].[2021-04-14].http://data.culture.fr/thesaurus/page/ark:/67717/T96 .

[21] Council of Europe. Thesaurus[EB/OL].[2021-04-14].https://www.coe.int/en/web/herein-system/thesaurus .

[22] 王晓光，侯西龙，程航航，等.敦煌壁画叙词表构建与关联数据发布[J].中国图书馆学报，2020,46(4)：69-84.

[23] 王 芳，杨 京，徐路路.面向火灾应急管理的本体构建研究[J].情报学报，2020,39(9)：914-925.

[24] 贾君枝. 面向数据网络的信息组织演变发展[J].中国图书馆学报，2019,45(5)：51-60.

[25] 贾君枝.词表生态系统:构成要素及关联关系[J].中国图书馆学报，2020,46(4)：60-68.

[26] 吴雯娜，鲍秀林. 国家叙词库的体系结构与数据模型[J].中国图书馆学报，2016,42(2) :81-96.

[27] 唐旭丽，马费成，傅维刚，等.知识关联视角下的金融知识表示及风险识别[J].情报学报，2019，38(3)：286-298.

[28] 蒋 勋，苏新宁，陈祖琴.多维视角下应急情报管理体系的知识库构建研究[J].情报学报，2017,36(10):1008-1022.

[作者简介]蒋 勋，男，1980年生，无锡环境科学与工程研究中心教授、江苏省数据工程与知识服务重点实验室教授。

朱晓峰，男，1975年生，江苏省数据工程与知识服务重点实验室教授。

肖连杰，男，1990年生，南京大学信息管理学院博士研究生。

收稿日期：2021-04-30