

我国近十年知识组织技术研究进展分析*

赖璨 陈雅

(南京大学信息管理学院, 南京 210023)

摘要: 近年来各类现代知识组织技术逐渐出现、迅速发展并投入应用,但目前少见总结。本文旨在梳理近十年知识组织技术的研究进展,探索研究前沿。以“知识”并含“组织”为检索词对CSSCI数据库进行篇名检索,筛选出85篇知识组织技术相关文献进行阅读分析,将我国知识组织技术研究进展归纳为传统知识组织改进技术与新兴知识组织技术两大部分:前者包括知识组织自动化技术与用户参与式知识组织技术;后者包括语义网、本体、关联数据、主题图与知识图谱、知识组织系统互操作与可视化技术。未来知识组织技术研究前沿方向包括细粒度知识组织技术、大规模语义知识组织技术和视听资源知识组织技术。

关键词: 知识组织; 知识组织技术; 语义组织; 互操作

中图分类号: G203 **DOI:** 10.3772/j.issn.1673-2286.2020.12.002

引文格式: 赖璨, 陈雅. 我国近十年知识组织技术研究进展分析[J]. 数字图书馆论坛, 2020 (12): 9-16.

随着信息社会的发展,人们逐渐湮没在文献与信息海洋里,核心需求转变为价值更高的、有序的知识。无论从速度、数量还是质量方面考虑,依赖于人脑分析判断的传统知识组织与服务模式在互联网时代显然难以为继,知识组织新模式受到研究者的关注,各类现代知识组织技术逐渐出现、迅速发展并投入应用,但目前少见总结。在已有研究中,王兰成等^[1]于2008年发表《国外知识组织技术研究的现状、实践与热点》一文,常娥等^[2]于2016年对多种知识组织方法进行了比较,韩燕等^[3]对2006—2016年我国传统知识组织系统关联化研究进行了内容分析,以上都尚未涵盖近几年的新研究。

本文拟对近十年国内知识组织技术研究进展与前沿进行探析,由于CSSCI来源期刊收录学科相关文章较为集中,并且文章质量相对较高,其中的研究较有代表性,故将其选择为文献源,数据库篇名检索“知识”并含“组织”,设定学科类别为“图书馆、情报与文献学”,限定时间为2011—2020年,显示共有198条结果,结合浏览题录与细读文本筛选出85篇技术相关文献,从中将近十年国内知识组织技术研究进展归纳为传统

知识组织改进技术与新兴知识组织技术两大部分。

1 传统知识组织改进技术分析

1.1 知识组织自动化技术

以分类法、叙词表等为代表的传统知识组织手段,用等级结构或聚类揭示概念间关系,人工组织更新缓慢,机器不能理解与处理,跟不上知识的增长速度,在这样的背景下知识组织自动化技术应运而生,主要包括术语自动抽取、词表自动构建与丰富、自动标引、自动分类、自动映射等方面。

术语是知识组织系统中的核心元素,术语自动抽取又称术语自动识别,指从特定领域文本抽取核心概念词汇,可分为基于规则、基于统计、规则与统计相结合3种方式。李智杰等^[4]指出基于规则的方法主要利用术语词典和规则模板,将基于统计的方法归纳为词频统计、TF-IDF、DCDR、互信息、统计机器学习、信息熵等,规则与统计结合的方法则包括C-Value、NC-Value及混合使用法等,使用结合法的术语自动抽取准确度

* 本研究得到江苏省高校哲学社会科学重大项目“公共文化服务效能提升的路径与模式研究”(编号:2017ZDXM001)资助。

更高。

词表自动构建与丰富,通过挖掘词和词间关联,实现词汇聚类。李育嫦^[5]归纳了当时应用的多种词表自动构建方法,包括基于同现分析的方法,概念空间方法,贝叶斯网络方法,基于互操作的方法,利用网络链接结构特征、通过语法分析的方法,利用用户检索策略等。在这些方法中,基于词汇同现频次的统计占据主流,普遍的不足在于识别的仅为相关关系,欠缺等级与等同关系的识别。将标引词与标题中的关键词进行同现分析,一定程度上可以发现并增添新词,促进词表更新。

自动标引可将作者语言自动转换为标引语言,方便检索系统语言匹配,与人工标引相比,在增量、组合、排序等方面应加以优化。自动标引可分为自动抽词标引和自动赋词标引,也可分为概率统计、语言分析和机器学习3类方法,统计法较为简便易行,语言分析法需要利用算法产生标引词,机器学习法需要提前训练。分词尤其是汉语分词是标引的一大难点,汉语词切分技术包括词典法、单汉字标引法、语法分析标引法和神经网络分词法等。李千驹等^[6]对核反应堆工程领域语料进行测试,依托关键词词表和叙词表,通过字符串匹配法获取候选标引关键词,绕开了分词过度问题。

自动分类目前分为基于机器学习和基于知识组织体系两类方法,基于机器学习是主流技术,贝叶斯分类、K-近邻分类、支持向量机、神经网络算法等都较为成熟^[5],但机器学习的自动分类局限在于必须依靠样本数据、关键词机械匹配,且较难解决多层分类问题,因此需要改造分类法(如压缩等级结构、合并稀疏类等),以便于机器学习。基于知识组织体系的自动分类加入了术语关系,即考虑了语义与层次关系,需要在分类前建立知识库而无须样本数据,沈思等^[7]提出以标引词检索关系表得到类主题词,再得到类号与权值的分类算法,但基于知识组织体系的自动分类质量依赖于知识组织体系的完善程度。

自动映射主要分为基于规则和基于统计两类方法。曲建峰等^[8]提出在收集并统计标引记录的分类数据基础上制定映射规则,建立分类法类目的统计映射表,运用人工干预对无映射关系的分类进行补充,以中图法和杜威十进分类法为例建立了两者的映射关系。李千驹等^[9]提出以用户检索日志为语料,通过组配、拆分、抽取、同构、加和、勘误6种处理方式将用户检索用语向叙词表映射,在映射规则中语义关系类型还需要细化扩展。与基于规则的方法相比,基于统计的方法无须

对词表进行语义对比分析,但反映的只是相关关系,准确性与样本数据的大小与质量密切相关。

另外,李晓瑛等^[10]设计了英文同义关系自动发现算法,涉及词形和语义归并、多义术语传递控制及来源词表颗粒度控制等方面,对中文术语同义关系发现也有借鉴意义。毛海波^[11]设计了OA期刊知识自组织系统,其中针对不同网站的异构问题提出新的网络蜘蛛自动收割OA期刊网站深层资源模式,自动标引分为直接标引和间接标引,前者基于中图法与学科关系表,后者通过对题名进行分词获得特征词,经过筛选与TFGINI权重算法计算后进行学科标引,大幅提高了OA期刊库的建设效率。

知识组织自动化技术可以实现知识的动态更新和管理,有助于人员从烦琐的工作中解脱,集中精力解决知识库的设计和优化等更高层次的问题,降低成本并提高效率。在知识组织自动化技术中,基于规则的自动化知识组织在规则设定科学的情况下所得结果较为准确,专业性较强,但对规则设定者的经验提出了较高要求,结果容易受个人主观性影响,也难以适应复杂多变的术语构词规则。而基于统计的自动化知识组织相当于对语料的归纳,实用性更强,结合机器学习可以快速计算语料数据的规律,得出较为通用的结果,但由于术语具体关系识别欠缺、语料库往往不能系统涵盖各类语言现象等原因,知识组织结果的准确率有待提升。因此将规则与统计相结合是一种较好的策略,还可以再结合用户反馈结果加以调整完善。目前知识组织自动化技术仍存在字面机械匹配的问题,在语义和语用知识的组织层面尚待提升。

1.2 用户参与式知识组织技术

大众分类法和网络百科是Web2.0环境下出现的新型知识组织机制,面向用户参与而不是专业群体,具有去中心化、简易性、动态化等特点。

大众分类法的原理是让资源的创建者和使用者主动给资源贴上自定义标签,对内容进行描述予以分类,每个人可以使用不同的标签,同一资源可以使用多个标签,以完善补充网站原有的标准分类,共享于整个站点。大众分类法的缺陷在于缺乏必要的语义,概念间缺乏关联,因此将其应用于知识组织主要结合了概念格,概念格又称形式概念分析,在知识发现领域概念格可以提取分类规则、蕴含规则和关联规则等。如毕达天等^[12]

运用概念格对传统AHP法初始评价结果进行关联规则挖掘, 动态实现合理配置垃圾标签过滤指标权值。滕广青等^[13]运用概念格对关联标签进行聚类, 呈现标签“语义关联词”的关系, 并通过关联规则挖掘揭示了标签“语义下位词”关系。

网络百科是协作开放式的在线百科全书, 里面的条目称为词条, 每个词条独立描述一个主题概念并从属于至少一个分类, 网络百科通过丰富的层次分类体系与超链接将相关词条联系起来, 形成一个巨大的网状体系, 供用户查阅与充实。万力勇等^[14]介绍了维基百科知识组织关键技术, 包括文档网络和分类树、文档链接和分类链接、随机页面、重定向和消歧页等, 但维基百科在知识自组织方面存在不足, 包括分类与链接的准确性等。

与传统的知识组织体系相比, 用户参与式知识组织具有较高的自由度, 用户自主提供的标签或分类可能是主题, 也可能是时间、人物等, 从而实现知识的多维化描述, 再借助链接技术形成知识间的关联共享, 这些标注的动态更新还可以反映词汇与社会热点的变迁。此类技术的不足之处在于随意性强, 导致标注的同义词、冷僻词乃至垃圾标签激增, 影响知识聚集与结构形成, 与概念格的结合一定程度上缓解了这一问题, 如何提升概念分类与关联的准确度仍需进一步研究, 如为用户引入术语库词汇、采用词义计算技术等。

2 新兴知识组织技术分析

语义网、本体、关联数据、主题图等是近年来蓬勃发展的知识组织技术, 可以建立数据间的联系, 从单纯的显性知识描述转变为揭示其潜在深层的语义关系, 支持机器理解与推理。与传统知识组织系统化构建知识体系相比, 这类知识组织技术更侧重呈现知识间的多元关系, 提供更多检索入口, 并促进知识组织对象向细粒度方向发展。

2.1 语义网

语义是对数据的解释, 语义网对万维网进行了扩展, 目标是利用元数据描述资源, 使计算机能理解网上的信息和知识, 并可以进行一定的自动处理和推理。语义网结构整体分为7层, 语义关系更为灵活, 主要基于XML、RDF和本体技术实现。曾新红^[15]利用RDF、

RDFS、OWL对SKOS进行了扩展, 以实现传统受控词表全描述。

关联知识聚合是语义网应用的具体形式, 关联知识元数据可通过被动服务器端采集或数据端采集, 知识元数据预处理可采用知识集成、知识清理、知识归约等。张海涛等^[16]基于自组织神经网络构建知识聚合模型, 引入模糊均值聚类算法, 利用分布式计算系统进行知识抽取、知识映射, 实现图书馆关联知识聚合。马晓悦^[17]提出结合语义网和社会标注建立社会化语义网知识组织模型, 实质是组织反映内容概念的知识文本, 分类识别知识主题与类别, 与一般语义网相比突出“用户参与”核心, 强调用户理解的动态化和不同理解间的协调。

语义网将事物的属性以及事物间的语义联系明确而简洁地表示出来, 利用结构化数据加快了数据流通, 但尚须解决多源数据融合方面的问题, 以实现高效信息共享。此外, 由于RDF和OWL等语言逻辑性太强, 对普通开发人员而言具有理解和掌握的门槛, 语义网技术的发展速度面临瓶颈。

2.2 本体

本体起源于哲学领域, 在信息时代被应用于人工智能领域, 本体技术通过重用、集成、映射等方式将异构数据集中到一个本体, 以实现不同资源或概念的共享。

本体构建方法可分为人工、半自动与自动3类。马创新等^[18]将当时人工方法总结为七步法、TOVE法、骨架法和METHONLOGY法等。王颖等^[19]参考了七步法和骨架法等, 结合文本挖掘与专家判断, 在国史本体的基础上对国史知识和事实进行语义揭示, 构建了国史知识网络。传统知识组织系统是投入大量人力物力构建而来的智力成果, 包含丰富的术语和语义关系, 完全可以加以改造用以本体构建, 以免另起炉灶导致资源浪费。何琳^[20]提出通过映射分类体系、整合叙词表, 利用机器学习建立语义细分分类器, 从而构建领域本体。毕强等^[21]将基于概念格的半自动化本体构建方法概括为Obitko方法、Haav方法和Cimiano方法等, 前两种方法直接将概念格与本体中的概念等同, 后者将本体中的概念匹配概念格中的属性, 并提出运用概念格构建领域本体的技术路径, 分别构建基于结构化资源(如主题词表等)和非结构化资源(如文本等)的形式背景及相应概念格, 通过并叠置运算获得新的概念格, 最终转

化为目标领域本体。这种方法易于实现自动化,资源预处理环节可使用汉语词法分析系统,概念格构建环节可使用ConExp、Lattice Miner等工具,本体构建可使用Protégé、KOAN等工具。

白华^[22]提出结合大众分类本体与一般概念本体(知识组织系统),基于用户标签和浮出语义方法,构建以概念为中心的用户标签集成本体,整合标签和知识框架,形成适应大数据环境的新本体。浮出语义是信息源多次交换协商资源和标签的含义基础上进化而来的统一本体,具有动态性和交互性,问题在于大众分类标签可能与知识组织系统中的概念不一致。张娴等^[23]以变量农作机械专利为研究对象,基于语义TRIZ对专利技术进行建模,构建领域本体、专利技术供给本体和需求本体,提出利用本体求解专利技术供需匹配的知识组织方案。

本体可以表达概念间的复杂关系,解决语义异构,实现分布式共享,但不能直接建立知识单元与资源实体的关联。本体不仅关联概念间显性的关系,还具有推理功能,通过隐含关系推导出新知识,但本体构建需要领域专家参与,较为复杂,如何在高效构建本体的同时降低成本是研究的重点。目前研究中的本体构建多是基于特定领域,有助于用户对某一领域的知识达成一致,但通用本体缺乏,且本体语言及构造标准多元化,已构建本体的扩展和共享性、本体的更新演化带来的版本控制等是尚待研究的问题。

2.3 关联数据

关联数据的URI既是知识概念的唯一标识,也可以对资源实体进行导航定位,实现概念的规范控制与多元化表达。作为将数据链接起来的语义资源,关联数据在异构异质知识资源的组织聚合中发挥日益重要的作用。关联数据自身不具备语义功能,需要将词表或本体作为其语义描述框架。关联数据可以标注文本和多媒体等资源,进行知识描述与揭示,核心在于找到与待标注对象相匹配的数据集,如文本标注可以利用基于数据集内部结构的PageRank方法或基于数据集内容描述的Context Similarity方法^[24];也可以通过连接以关联数据形式发布后的知识资源实现知识单元间的互联;还可以对不同来源的知识进行网状结构组织序化。

董坤^[25]基于关联数据提出高校知识资源语义化组织框架,构建了高校知识管理原型系统,在资源描述本

体的基础上使用D2RQ平台将关系数据库语义映射至关联数据,改变了传统高校知识组织系统的封闭状态,实现开放式语义检索与获取。曾子明等^[26]基于关联数据构建数字人文视觉资源知识组织模型,分为资源采集层、语义描述层、知识关联层和智慧服务层,以敦煌学为例在构建本体的基础上建立关系数据表实现关联。

关联数据将概念与实体置于统一的RDF框架下,可实现概念间、实体间、资源与实体间的语义关联。关联数据强调在不同的数据集之间创建链接,还能融合元数据与本体,让用户能更加方便地进行访问、查询和处理,可以有效实现语义推理和知识发现。知识组织应用关联数据技术面临的问题在于,由于数据集存在内容局限性,直接导致了语义标注的局限性,同时知识单元互联与组织受到数据源开放共享程度与知识产权的限制。

2.4 主题图与知识图谱

主题图可以从异构库中抽取关键数据,通过主题、关联及资源出处进行导航,揭示网状-核心-边缘的知识结构,基于XML Topic Maps和HyTM等语法实现功能。胡娟等^[27]应用主题图技术构建学术博客知识组织模型,主要分为主题关联层和blog层,学术博客知识可以根据需要添加不同类型的主题来多方面分类揭示。陈婷等^[28]将主题图技术和社会化标签相结合,将标签转换为主题,使用主题图的“范围”要素解决标签一词多义的问题,从而将扁平化的标签转化为网状结构,并以超星数字图书馆为例,基于Ontopia引擎构建了标签主题图。相较传统主题图,智能主题图可采用全信息相似性算法进行聚类,匹配语法、语义和语用,并且扩展了知识元导航功能与推理功能,应用基于规则引擎的推理机制,由内部或自定义规则获取隐性知识^[29]。

知识图谱有两种含义,一类是谷歌公司提出的基于语义网的大规模知识库,另一类是科学知识图谱。后者将图形学、计量学、引文分析和共现分析等结合,用于展示学科整体知识架构、核心结构与前沿领域等,揭示知识领域的动态发展情况。娄国哲等^[30]基于前一类知识图谱,将时间属性添加至基本组成单元,形成实体、关系、时间区间、实体四元组和实体、属性、时间、属性值四元组,对网络舆情进行知识组织,先创建网络舆情本体后构建网络舆情知识图谱,可直接通过内容采集、文本处理与事件发现实现网络舆情事件的抽取。

与语义网、本体的知识概念不同,主题图将人、时

间、动作等所有标引对象都视为主题,利用智能主题图可以在网络知识组织中实现知识表示、推理、导航、推送与可视化,但不能在网络发布和获取资源。由于用户可以自由定义主题图的分类词汇,存在用词不规范导致连接错误或遗漏信息等风险。知识图谱规模大、语义丰富,能将分散的知识序化,用可视化方式展示知识的分布及关系,由显性知识引导获取隐性知识,还能向用户提供信息智能查询与深度问答服务等;缺点在于为了纳入更多知识,在设计时允许模式定义不完善或缺失,从而带来了数据质量控制问题,进而影响知识图谱的完整性与准确性,此外知识图谱算法依赖现有数据集,如果要扩展则需修改数据结构和业务逻辑,维护成本较高。

2.5 知识组织系统互操作与可视化技术

知识组织系统包括分类法、叙词表、语义网、本体等,无论是同构知识系统,句法、术语或概念异构知识系统,还是跨语言、跨领域的知识组织系统,实现多系统彼此间的兼容转换互操作,让词汇与检索式直接适用或转换后适用于多个系统,可以让分布式资源实现统一组织与利用,是知识组织的一大目标。分类与主题语言互操作是实现跨语言检索的有效手段,语义互操作是更高层次的互操作,目标是达到异构系统语义层面的机器理解与交互,是最难实现的。

李育嫦^[5]将知识组织系统互操作实现技术概括为演化\建模、翻译\改编、映射、链接以及中心转换;胡滨等^[31]将知识组织系统互操作方法概括为推导/建模(派生法)、翻译/转译、映射、转换/中介词典、卫星子表、元叙词表、集成词表和链接;郭思成等^[32]运用深度学习工具Word2Vec将词条进行向量形式转化,自动匹配筛选,建立医学词表间的语义映射,实现知识组织系统互操作,消除歧义、统一概念逻辑,并指出后续可以综合使用TF-idf、SVM分类器等技术,或结合相关词对齐库;李慧佳等^[33]提出对智库资源加以元数据规范,最大限度整合元数据标准,并且构建元数据本体及关联关系进行语义化描述,从而实现语义互操作,元数据本体可以通过集成或映射的方式构建。

目前映射、链接和翻译等是最常用的知识组织系统互操作技术,映射建立了知识组织系统间的概念对照,也是一些互操作技术的基础,其复杂程度受系统结构、概念定义等的异质性影响,链接通过中介词表建

立间接关系,不改变原词表的体系,翻译有助于促进多语种知识组织系统间的互操作,但可能会导致语义的损耗或歧义的出现。知识组织系统互操作已拓展至语义互操作层面,如结合元数据标准,并有研究开始结合深度学习加以辅助实现互操作的自动化,较为依赖训练语料的质量,仍需结合专家意见与人工标注干预来保证互操作效果。

网络环境下,传统知识组织系统结构无法全面、直观、动态地显示词间关系,与单纯的文本展示相比,可视化用图像显示多维数据,可以加深用户对数据含义的理解,增强用户对抽象事物的认知。张运良等^[34]总结了当时Web前端可视化技术包括Java Script、Flash、Java Applet、Silverlight等,将D3.js类库应用于知识组织系统动态交互的Web可视化,为了实现节点合理分布使用力定向布局功能。D3是数据驱动文档,可以在文本对象模型上绑定任意数据,应用数据驱动转换到文档,利用HTML、级联样式表和可缩放矢量图形实现可视化展示。知识组织系统可视化的常用方法包括缩进树、节点树、可缩放展示、焦点加上下文和3D展示等^[35],但图形可视化展现的是概念之间的关系,缺少对概念语义的展示,王福等^[36]总结了可以通过可视化概念格、可视化聚类层次、可视化本体结构和可视化关联图谱等形式展现知识聚合。

3 知识组织技术研究前沿

3.1 细粒度知识组织技术

粒度是对不同层次知识的度量,一本书或一篇论文包含许多知识,属于粗粒度,而一个术语词汇等则属于细粒度知识。现在知识组织已经由粗粒度向细粒度转变,但如何更好地实现对细粒度知识的抽取、挖掘与组织,仍然是值得研究的问题。研究者对细粒度知识的对象着眼于知识元、知识单元和单元信息等。

知识元是不能分割的最小的独立知识单位^[37]。刘杰等^[38]通过抽取、标引、显隐性链接和集成知识元,构建了四层次六梯度的知识元映射,以组织科技文本资源内容,实现知识的精准定位与拓展延伸。其局限在于由于中文语义语法的复杂性,知识元自动抽取技术结果准确性还不高,需要加入人工操作。

李祯静等^[39]将知识单元定义为深入资源内部、能够完整表达知识、具有相对独立性的基本单元,是知识

内容本身。他们基于文本结构的抽取方法将科技文献划分为以五元组表示的多个知识单元,采用基于文本集合关联规则的算法自动生成语义链接网络,最后使用Cytoscape软件实现网络构建与可视化。

张娟等^[40]认为知识元的语义太零碎,知识单元又过于宽泛,提出知识组织对象是单元信息,即文献中内容意义完整的一个知识单元,如观点、数据、方法等,综合运用元数据、本体和关联数据进行组织,如何识别、抽取及关联异构知识还有待进一步研究。

对细粒度知识的组织有助于实现更精准有效的检索与挖掘,细粒度知识内部根据单元大小也有层次划分,组织的技术路径主要为细粒度知识抽取与语义关联,知识抽取的准确度影响知识组织的质量,现有研究主要面向科技文献,未来有待改进自然语言处理和使用结合规则与统计的算法。

3.2 大规模语义知识组织技术

虽然新兴知识组织技术有效弥补了传统知识组织无法精准发现特定主题知识这一缺陷,发展了语义抽取、描述与组织方法,但仍然受到数据集规模和质量限制。大数据时代,各种大规模数据集涌现,内含知识密度低、关联多,需要通过数据挖掘等手段发现。高效处理动态发展、海量增长的数据集,对其进行语义分析,将其中的知识资源实现有序化、系统化组织有赖于大规模语义知识组织技术的发展。

李旭晖等^[41]提出基于语义数据管理进行知识组织,需设计合适的语义数据模型,数据结构应支持大量归纳计算,并提供原子计算,将人工智能技术应用于自主、自动对海量知识数据的循环迭代挖掘,同时知识库在本体驱动下自主归纳知识内容和结构演化模式。在人工智能时代,还可以将基于神经网络算法的深度学习技术应用于知识组织。孙坦等^[42]提出将知识组织与认知计算相融合的开放知识服务系统改进思路,其中需要突破大规模语义知识库构建技术等。

大规模语义知识组织可以适应知识的动态演化,进行知识重组与服务,实现知识发现与增值,已有研究主要聚焦在基于新兴技术的知识计算,提高知识组织效率,而如何与传统知识组织相结合,统一表现处理非结构化与结构化语义信息,提升语义组织的准确度尚待探索。此外,由于数据集体量庞大,还要加强对数据存储结构和索引机制的设计研究。

3.3 视听资源知识组织技术

如今视听资源数据规模日益大幅增长,已有文献大多数是对文本资源的知识组织技术进行研究,较为成熟,而传统对图像、音频、视频等形式的资源多是简单著录,转换成基于关键词或文本标注模式,对其具体知识内容的组织技术的专门研究很少,特别是其中的语义如何识别并深度描述、知识间如何关联、实时编辑等问题亟待研究。

视听资源的知识组织方式大致分为基于人工、基于图像匹配和基于上下文3类。庄文杰等^[43]认为非遗视频知识元一般只能采用人工提取,借鉴DC、CDWA和VRA,从基本信息、视频属性、非遗属性、知识元属性和可扩展属性五方面进行元数据描述,再进行链接与可视化表示,这种方法可以保证较高的准确性,但人力与时间成本较高。张兴旺等^[44]以平面图结构算法描述非遗视觉资源,并采用基于邻居向量的近似子图匹配算法,在视觉知识库中搜索非遗视觉资源局部特征相似的资源,通过与其他物体间的位置与出现概率比较计算得出彼此关联关系,这种方式的挑战在于如何实现绘制画面与实景的相似性判断。曾子明等^[26]以敦煌学视觉资源为例,通过采集相关文本,抽取上下文的语义信息来进行资源描述与标引,将文本与图像关联起来;李建伟^[45]利用口述历史元数据同步器功能插件,实现关联检索多媒体语义内容与上下文,提供口述记录与同步转录文本,完成信息资源多维度揭示,这种方式难以涵盖图像或画面的方方面面,如人物的服饰、神态等细节,仍需要人工补充。

视听资源的知识组织在语义层面要深入至资源的时空背景、资源内的主体客体、动作神态等,对视觉资源而言,综合图像与文本描述的知识组织更为全面,音频资源则主要结合语音识别转录与传统的描述方式。随着人工智能技术的应用,基于机器学习的图像识别对比与资源标注等将在视听资源知识组织中发挥更大的作用,未来还要加强对视听资源知识推理机制以及与其他类型知识的关联研究。

4 结语

21世纪以来,人们对知识的需求日益强烈,并向精细化发展,不再满足于一本书、一篇文章、一段视频等检索结果,而是寻求其中更深度相关的片段,推动知识

组织技术的研究与实践应用。本文对近十年CSSCI收录的知识组织技术代表文献进行了阅读分析,发现已有研究在自动化、语义化和关联化知识组织技术等方面都取得了较大进展,并出现部分对技术的组合应用研究,前沿研究领域细粒度知识组织技术、大规模语义知识组织技术和视听资源知识组织技术等仍需深入探索,未来还要重点攻克知识组织系统的扩展和共享、多源数据的歧义和噪声、用户参与知识组织的规范控制等问题,为向用户提供更全面更精准的知识服务奠定基础。

参考文献

- [1] 王兰成, 敖毅, 曾琼. 国外知识组织技术研究现状、实践与热点[J]. 中国图书馆学报, 2008 (2): 93-97.
- [2] 常娥, 夏婧. 多种知识组织方法比较[J]. 图书馆论坛, 2016, 36 (8): 1-6.
- [3] 韩燕, 何琳, 彭爱东. 近10年我国传统知识组织系统关联化研究述评[J]. 图书情报工作, 2017, 61 (24): 135-143.
- [4] 李智杰, 曾文, 乔晓东. 知识组织系统构建技术研究[J]. 情报理论与实践, 2017, 40 (1): 115-120.
- [5] 李育嫦. 传统知识组织系统的重构及其在网络环境下的应用[J]. 情报杂志, 2011, 30 (7): 114-118.
- [6] 李千驹, 李思达, 刘建毅. 一种基于知识组织的关键词自动标引方法[J]. 情报科学, 2016, 34 (11): 107-110, 139.
- [7] 沈思, 苏新宁. 知识服务环境下分类表的知识组织探究[J]. 图书情报工作, 2014, 58 (7): 113-118.
- [8] 曲建峰, 李芳, 张铁华, 等. 知识组织系统自动映射规则研究与实现——以《杜威十进分类法》和《中国图书馆分类法》为例[J]. 现代图书情报技术, 2012 (10): 83-88.
- [9] 李千驹, 宋培彦. 基于会话合作原则的知识组织方法及其实证研究[J]. 情报理论与实践, 2015, 38 (8): 91-94, 99.
- [10] 李晓璜, 李丹亚, 钱庆, 等. 面向知识组织系统整合的英文同义关系自动发现算法研究[J]. 现代图书情报技术, 2014 (5): 26-32.
- [11] 毛海波. OA期刊知识自组织系统设计[J]. 情报杂志, 2015, 34 (2): 154-158.
- [12] 毕达天, 邱长波, 滕广青. 基于概念格的Folksonomy知识组织研究——Tag Spam过滤指标权重配置[J]. 情报学报, 2011, 30 (10): 1078-1085.
- [13] 滕广青, 毕强, 高娅. 基于概念格的Folksonomy知识组织研究——关联标签的结构特征分析[J]. 现代图书情报技术, 2012 (6): 22-28.
- [14] 万力勇, 黄传慧. 用户生成性学习资源的知识组织机制研究[J]. 图书馆学研究, 2014 (19): 47-53.
- [15] 曾新红. 中文知识组织系统形式化语义描述标准体系研究(一)——扩展SKOS实现传统受控词表全描述[J]. 中国图书馆学报, 2012, 38 (3): 57-68.
- [16] 张海涛, 张连峰, 王丹, 等. 基于自组织神经网络的图书馆关联知识聚合研究[J]. 情报理论与实践, 2015, 38 (9): 73-78.
- [17] 马晓悦. 考虑观点多样性的社会化语义网知识组织模式探究[J]. 情报科学, 2016, 34 (7): 25-30.
- [18] 马创新, 陈小荷. 基于学科本体的训诂学知识组织体系初步构建[J]. 图书情报工作, 2013, 57 (12): 118-122.
- [19] 王颖, 张智雄, 孙辉, 等. 国史知识的语义揭示与组织方法研究[J]. 中国图书馆学报, 2015, 41 (4): 55-64.
- [20] 何琳. 基于知识组织资源仓库的领域本体构建研究[J]. 图书馆杂志, 2011, 30 (12): 59-62, 112.
- [21] 毕强, 鲍玉来. 数字图书馆知识组织体系构建的发展路径——概念格与本体的互补融合[J]. 华中师范大学学报(人文社会科学版), 2011, 50 (5): 130-136.
- [22] 白华. 大众分类本体与知识组织系统融合研究[J]. 图书馆学研究, 2016 (10): 53-59.
- [23] 张娴, 胡正银, 茹丽洁, 等. 专利技术供需信息关联知识组织模式研究[J]. 图书情报工作, 2016, 60 (8): 118-125.
- [24] RUSU D, FORTUNA B, MLADENIC D. Automatically annotating text with linked open data [J/OL]. LDOW, 2011, 813 [2020-12-01]. https://www.researchgate.net/publication/265929816_Automatically_Annotating_Text_with_Linked_Open_Data.
- [25] 董坤. 基于关联数据的高校知识资源语义化组织研究[J]. 情报理论与实践, 2016, 39 (3): 91-95.
- [26] 曾子明, 周知, 蒋琳. 基于关联数据的数字人文视觉资源知识组织研究[J]. 情报资料工作, 2018 (6): 6-12.
- [27] 胡娟, 程秀峰, 叶光辉. 基于主题图的学术博客知识组织模型研究[J]. 图书情报工作, 2012, 56 (24): 127-132.
- [28] 陈婷, 胡改丽, 陈福集, 等. 社会标注环境下的数字图书馆知识组织模型研究——基于标签主题图视角[J]. 情报理论与实践, 2015, 38 (3): 63-70.
- [29] 余利娜, 盛小平. 利用智能主题图开展网络知识组织研究[J]. 图书情报工作, 2011, 55 (20): 115-120.
- [30] 娄国哲, 王兰成. 基于知识图谱的网络舆情知识组织方法研究[J]. 情报理论与实践, 2019, 42 (1): 58-64.
- [31] 胡滨, 吴雯娜. 国内外知识组织系统互操作模式及方法研究[J]. 情报科学, 2012, 30 (9): 1291-1297.

- [32] 郭思成, 李纲, 周华阳. 基于Word2Vec的医学知识组织系统互操作研究——以词表间语义映射为例[J]. 情报理论与实践, 2019, 42(9): 160-165, 176.
- [33] 李慧佳, 王楠. 基于语义关联的智库资源知识组织研究[J]. 图书与情报, 2020(1): 120-126.
- [34] 张运良, 张兆锋, 张晓丹, 等. 使用D3.js的知识组织系统Web动态交互可视化功能实现[J]. 现代图书情报技术, 2013(Z1): 127-131.
- [35] SWAMINATHAN V, SIVAKUMAR R. Comprehensive ontology cognitive assisted visualization tools-a survey[J]. Journal of Theoretical and Applied Information Technology, 2012, 41(1): 75-81.
- [36] 王福, 毕强. 不规则知识晶格在知识的多维度柔性组织研究[J]. 情报理论与实践, 2016, 39(1): 16-23.
- [37] 高国伟, 王亚杰, 李永先. 我国知识元研究综述[J]. 情报科学, 2016, 34(2): 161-165.
- [38] 刘杰, 秦春秀, 赵捧未, 等. 基于知识元的科技文本资源内容组织方法[J]. 情报理论与实践, 2018, 41(4): 128-133.
- [39] 李祯静, 秦春秀, 赵捧未, 等. 科技文献的资源语义空间: 一种细粒度知识组织方法[J]. 情报杂志, 2019, 38(2): 158-165, 180.
- [40] 张娟, 陈人语. 语义网背景下基于单元信息的知识组织框架研究[J]. 国家图书馆学刊, 2018, 27(6): 54-59.
- [41] 李旭晖, 秦书倩, 吴燕秋, 等. 从计算角度看大规模数据中的知识组织[J]. 图书情报知识, 2018(6): 94-102.
- [42] 孙坦, 刘峥, 崔运鹏, 等. 融合知识组织与认知计算的新一代开放知识服务架构探析[J]. 中国图书馆学报, 2019, 45(3): 38-48.
- [43] 庄文杰, 谈国新, 侯西龙, 等. 非物质文化遗产视频知识元组织模型研究[J]. 情报科学, 2018, 36(12): 25-32.
- [44] 张兴旺, 卢桥, 田清. 大数据环境下非遗视觉资源的获取、组织与描述[J]. 图书与情报, 2016(5): 48-55.
- [45] 李建伟. 内容深度发现下的非结构化数字记忆遗产知识组织——以口述梅州侨批史料保护为例[J]. 图书馆论坛, 2019, 39(4): 166-173.

作者简介

赖璨, 女, 1996年生, 硕士研究生, 研究方向: 公共文化服务、图书馆管理与服务。

陈雅, 女, 1965年生, 博士, 教授, 通信作者, 研究方向: 公共文化服务、图书馆管理与服务、图书馆学基础理论, E-mail: yachen@nju.edu.cn。

Analysis on the Research Progress of Knowledge Organization Technology in China in Recent Ten Years

LAI Can CHEN Ya

(School of Information Management, Nanjing University, Nanjing 210023, China)

Abstract: In recent years, various modern knowledge organization technologies have gradually appeared, rapidly developed and put into application, but there are few summaries at present. This article aims to summarize the research progress of knowledge organization technology in recent ten years and explore the frontiers of research. The CSSCI database was searched by title with "knowledge" and "organization" as the search term, then 85 articles related to knowledge organization technology was selected for reading analysis. The research progress of knowledge organization technology in China was summarized as two major parts: traditional knowledge organization improvement technology and emerging knowledge organization technology. Traditional knowledge organization improvement technologies include knowledge organization automation technology and user participatory knowledge organization technologies, and emerging knowledge organization technologies include semantic web, ontology, linked data, topic maps and knowledge graphs, and knowledge organization system interoperability and visualization technology. The frontier research directions of future knowledge organization technology include fine-grained knowledge organization technology, large-scale semantic knowledge organization technology, and audiovisual resource knowledge organization technology.

Keywords: Knowledge Organization; Knowledge Organization Technology; Semantic Organization; Interoperability

(收稿日期: 2020-11-24)