

知识组织与检索语言学术研讨会综述

曹树金 刘慧云 张乐乐 常赵鑫 李慧玲 王雅琪 常惊伟

【摘要】为促进全国知识组织与检索语言领域的学术研讨和交流合作,2018年12月2日至5日,由中国图书馆学会学术研究委员会主办,中山大学资讯管理学院、中国图书馆学会学术研究委员会信息组织专业委员会承办的“知识组织与检索语言学术研讨会”在广州召开。会议邀请了图书情报领域的知名专家就主旨内容作了学术报告,并对知识组织与检索语言的理论、方法和工具发展进行了研讨。本文对此次会议的学术报告和研讨内容进行综述,以供更多的同仁了解和参考。

【关键词】知识组织;检索语言;会议综述

【作者简介】曹树金,中山大学资讯管理学院教授,博士生导师,E-mail:caosj@mail.sysu.edu.cn;刘慧云(通讯作者),中山大学资讯管理学院博士研究生,E-mail:liuhy39@mail2.sysu.edu.cn;张乐乐,中山大学资讯管理学院博士研究生;常赵鑫,中山大学资讯管理学院博士研究生;李慧玲,中山大学资讯管理学院硕士研究生;王雅琪,中山大学资讯管理学院硕士研究生;常惊伟,中山大学资讯管理学院硕士研究生。

【原文出处】《图书馆建设》(哈尔滨),2019.1.155~160

【基金项目】本文系国家社会科学基金重大项目“基于特定领域的网络资源知识组织与导航机制研究”的研究成果之一,项目编号:12&ZD222。

1 会议背景

为促进全国知识组织与检索语言领域的学术研讨和交流合作,2018年12月2日至5日,由中国图书馆学会学术研究委员会主办,中山大学资讯管理学院、中国图书馆学会学术研究委员会信息组织专业委员会承办的“知识组织与检索语言学术研讨会”在广州召开。本次会议的主题是“大数据时代知识组织方法和工具的创新与发展”,旨在探讨知识组织与检索语言领域理论和实践的前沿问题。来自全国相关高校院系的师生、国家图书馆和中国科学技术信息研究所等业界的专家近百人出席了此次会议。会议邀请了图书情报领域的知名专家就主旨内容作了学术报告,与会人员就会议主题进行了热烈的讨论。

会议开幕式由中山大学资讯管理学院博士生导师、中国图书馆学会图书馆学著作出版专业委员会副主任韦景竹教授主持,中国图书馆学会学术研究

委员会副主任、武汉大学信息管理学院副院长黄如花教授,国家图书馆副馆长汪东波研究馆员,北京大学信息管理系主任、中国图书馆学会副理事长李广建教授,中山大学资讯管理学院院长龙乐思教授,中山大学资讯管理学院情报学学科负责人、中国图书馆学会学术研究委员会信息组织专业委员会副主任曹树金教授分别致辞。

随后,会议进行了特邀报告、专题报告和专题研讨。下面就本次会议的特邀报告、专题报告和专题研讨进行综述,以供更多的同仁了解和参考。

2 特邀报告

会议特邀报告由武汉大学教育部人文社会科学重点基地信息资源研究中心主任、教育部长江学者李纲教授主持。

武汉大学人文社科资深教授、国家教学名师马费成教授以“大数据环境下用户需求信息组织”为题,阐释了随着大数据处理技术和机器学习技术的

发展和应用,大数据中的知识正在迅速地被发现和积累,原始的行业大数据正逐步演变为知识大数据。分析了大数据环境对用户需求信息组织提出了新的要求,包括提高用户需求信息组织的智能化水平、建立用户需求信息描述的标准和规则、实现用户需求信息的关联、加强用户需求信息组织与分析中的个人隐私保护等。他认为现有的用户需求信息组织方法不够理想,可以将关联数据技术应用于用户需求信息的组织中,构建用户需求语义网络。用户需求语义网络是一种利用关联数据技术创建和发布关于用户需求、行为及用户需求之间关系的规范化描述信息,以用户需求和行为特征为结点,以它们之间的关系为边而构成的语义化网络。随后,他构建了用户需求语义网络的理论框架,包括数据层、需求信息层和应用层,并认为用户需求语义网络可以运用于科技成果转化和技术交易以及图书馆用户需求信息组织等场景。最后,他指出用户需求语义网络很好地回应了大数据环境对用户需求信息组织提出的要求和挑战,因为用户需求语义网络将为智能化的用户需求信息处理提供基础和保障,采用RDF作为需求信息描述的统一规范,可以建立用户需求信息之间的关联,通过设置多种协议保护用户个人隐私。

南京大学信息管理学院学科带头人、教育部长江学者苏新宁教授以“面向知识服务的知识组织思路与方法”为题,论述了通过知识组织实现知识服务的新思路和方法。首先,他认为知识服务是指从各种资源中按照人们的需要有针对性地提炼知识,并用来解决用户具体问题的高端信息服务过程。这种服务的特点就在于,它是一种面向知识内容和解决方案的服务。图书情报领域提供的知识服务,应当充分发挥图书情报传统的知识组织色彩和知识服务特性。接着,他分析了知识服务与知识组织的关系。一方面,知识服务需要对数据组织有更高的要求;另一方面,科学的知识组织是实现知识服务的保证,有效的知识组织能够保证知识的准确提取,能够提炼出新的知识。随后,他具体提出了利用数字技术、可视化技术等进行词典的知识组织与服务的思路和方法,为百科知识建立分类知识体系和建立主

题关系的知识体系从而提供百科知识服务的思路,阅读文本中的知识点链接的思路和方法,知识组织中语义关系建立的四歌要点,包括语义表达简单化、语义关系表达的编码化、已有知识体系充分运用和重视动态知识的关联与服务。最后,他指出未来应当加强面向知识服务的知识组织研究与实践,发挥知识服务在社会发展、科学研究、政府管理中的重要作用。

3 专题报告

继特邀报告之后,是3个时段的专题报告,分别由国家图书馆中文采编部主任、中国图书馆学会学术研究委员会信息组织专业委员会副主任王洋副研究馆员,武汉大学信息管理学院珞珈特聘教授、中国图书馆学会学术研究委员会信息组织专业委员会副主任司莉教授,国防大学政治学院军事信息与网络舆论系教研室副主任包冬梅副教授主持。

3.1 知识组织与检索语言前沿理论

中山大学资讯管理学院曹树金教授以“大数据环境的知识组织”为题做了报告。他认为在大数据环境中知识组织至少发生了3点变化:知识组织对象的扩展,既要组织传统的文献、信息和知识,又要组织大数据中的数据;知识组织原则的变化,反映用户端的大数据会越来越成为知识组织的依据知识组织技术的发展,大数据技术和智能技术将越来越多的应用于知识组织。大数据对知识组织来说是挑战,更是机遇。大数据环境中知识组织有两个核心任务:一是结合用户情景,对更细粒度的知识单元进行更为细致的揭示与关联,二是从具体领域的需要、实践和研究中概括出更为一般的、可以跨领域和任务应用的知识组织理论和方法。在此基础上,他提出了大数据环境中知识组织研究的3个主要方向是面向情景的知识组织,数字人文中的知识组织,大数据治理中的知识组织。

中国人民大学信息资源管理学院贾君枝教授以“数据起源描述框架及应用”为题,报告了数据起源的产生背景、概念以及起源描述框架、起源管理框架和起源应用等内容。她在说明数据起源产生的四大背景基础上,界定了数据起源的概念,认为数据起源是描述数据产生、修改、拥有及其他影响的元数据。

随后,她分析了数据起源、元数据和信任之间差异,说明起源描述是起源利用的基础,是通过运用起源模型,对数据的信息进行RDF描述并使其形式化表示的过程,目的是方便人或机器获取数据的出处信息。接下来,她介绍了开放起源模型、W7模型、PROV模型等不同类型的数据起源模型,认为起源模型是起源数据特征的抽象,用于表示实体类及属性间的关系。她概括了起源管理框架包括的出处记录获取、起源分析准备、起源模型构建、起源描述生成、起源校验发布、起源查询及可视化6个环节。最后,指出起源数据可以应用在数据质量评估、数据追踪和数据分析等领域。

中国科技信息研究所常春研究馆员以“知识组织生态系统研究进展”为题,提出知识组织生态系统概念,从实例、概念、词表、系统4个层次,说明知识组织生态系统的研究内容。他以《汉语主题词表》的修订与重编进展为背景,将生态学的原理和方法,运用到知识组织系统中,构建了知识组织的生态系统,从而实现了生物个体与实例的对应、生物种群与概念的对应、生物群落与词表的对应、生物环境与文献环境的对应,以及生态系统与知识系统的对应。又将生态位法则引入知识组织,得出了知识组织系统的生态位法则;将生态学中的种群增长规律同知识组织系统相结合,得出了知识组织的种群增长规律。最后,他指出了知识组织生态系统的在实例个体、词表群落、概念种群和知识系统等层面的研究方向。

中国科技信息研究所张运良研究员以“知识服务中的知识组织挑战”为题,结合他在科研与工作中的经验,论述了当前知识服务中知识组织所面临的挑战。他认为知识组织是知识服务的基础,没有知识组织就无法升级创新成知识服务。目前的知识组织面临一些挑战,主要是基于语料库的知识组织系统构建中面临挑战,具体包括语料库规模大,处理困难,检索策略构建困难;知识的极端不平衡性;知识的不完整性;知识错误的偏离与传播等问题。知识组织体系的现代化面临挑战,在本体层面存在本体共识及共享效果不佳、本体推理偏弱、对实例理解缺乏等问题;在知识图谱层面,存在知识图谱体量巨大,为存储和利用带来挑战的问题;在资源媒体形式

层面,存在多媒体、跨媒体、富媒体、新媒体和媒体融合挑战。最后,在关系类型对应指引中,由于数据不断增加而带来的挑战是巨大的,阈值低,得到的关系数量就多,但关系之间又是不平衡的。基于众多关系数量所构建的知识图谱是否有意义,是一个值得探讨的话题。

3.2 知识组织与检索语言技术与方法

北京大学信息管理系主任、中国图书馆学会副理事长李广建教授以“用网络数据揭示非正式交流过程”为题,报告了如何借助信息抽取和深度学习技术,利用网络新闻等开放数据,揭示非正式交流的过程。他提出在大数据、互联网环境下,多源数据融合技术的不断发展使得网络新闻等为代表的公开情报源的价值不断凸显。他们在研究中尝试利用表层网络信息这一公开情报源研究线下的非正式的交流,突破以往基于特定业务系统数据而构建的用户画像,尝试选择公开的数据源来构建用户画像,以应对公开情报源的稀疏性,尝试引入机器学习的技术来进行公开情报源的信息组织。他指出当前人工智能最新发展阶段的特征是以机器学习为基础,将人的智能——“学习”赋予机器,机器学习将成为公开情报处理的重要工具和方法。应用深度学习的方法可以计算概念之间的关系以应对公开情报源信息表达中的离散性,进而梳理其内在逻辑关系;应用机器学习可以进行推理,发现概念之间的关系。

南京农业大学人文社科处处长黄水清教授以“古文信息处理:概念、现状与趋势”为题,首先阐述了古文信息处理的概念与对象,古文信息处理不包括古籍的载体特征,特指古代汉语文本呈现出来的音、形、义等具体形式。他认为目前的古文信息处理研究主要集中在数字化、智能处理、人文计算三个方面,其中数字化为另外两项研究奠定了基础,智能处理的主要贡献在于古文的词汇处理、自动断句、语义和句法标注三个方面。而后他介绍了其团队所做的研究工作:(1)运用条件随机场模型构建古汉语自动分词模型,结合字词结构、词长、拼音等特征信息构建古汉语词性标注模型,实验结果显示,其团队构建的自动分词模型的调和平均数F值稳定在95%以上,古汉语词性标注模型实验F值均值为90.4%,人名实体

识别效果优势明显。(2)基于不同的标注体系对古汉语进行自动断句,准确率有明显的提升。(3)古文语义和句法的标注,是该领域的一个难点。最后,他认为古文人文计算或可成为未来的热点问题,并提出研究古文人文计算的关键点在于研究对象与问题、数据与语料、算法模型3个方面。古文数字化发展中应重视其规范化,古文智能处理需要融合机器学习技术构建机器学习模型。

南京理工大学经济管理学院章成志教授的报告以“基于全文内容分析的算法使用行为与影响力研究”为题。他在介绍当前学术实体评估方法的基础上,引入基于全文内容的评估,以大数据挖掘算法为研究对象,以算法的提及数、提及位置、使用动机为指标,考察特定领域学术全文内容中算法的使用行为,并分析其影响力。从考虑全文本内容的算法学术影响力分析、特定任务关联的算法影响力分析、基于全文本内容的算法使用行为研究和特定领域中算法使用行为的中外差异研究四个方面展开论述。基于当前研究的成果,章教授提出了未来可以对基于大规模全文语料的算法自动抽取与评价、特定领域算法或研究方法的抽取与评价和面向给定任务的算法或研究方法推荐三个主题开展更为深入的研究。

四川大学公共管理学院信息管理技术系主任范炜副教授的报告题目为“《情报语言学词典》语义化进展”。他回顾了张琪玉老师在情报检索语言领域聚焦提高检索效率所做的贡献,认为《情报语言学词典》是情报学领域的信息检索知识体系基础,未来应当与时俱进修订专业词典,与现有信息检索专业教材形成补充;在面向情报学研究生的信息检索教学场景中增强语义模型理解与技术实训。接着,他对《情报语言学词典》的语义化研发进行了论述,并对词条、正文、大体量词条的拆解表示、引见关系、子概念处理、衍生概念、主题概念及其性质的语义关系建立等等进行了举例说明。最后,他认为下一步的词典语义化工作的重点分为以下4个方面:(1)词典正文部分要以概念为中心设计概念模型,以“引见”与释义结合梳理主要语义关系类型,并制作生成RDF数据;(2)情报检索语言语种简目方面,要以中文词表优先继续搜集整理,进行词表级描述并开展网络服

务注册与发布;(3)在情报检索语言文献简目(文库)上要搜集整理经典文献并基于关联数据开发论文网络数据集;(4)参照开放关联数据发布要求确定语义技术方案。

3.3 知识组织与检索语言应用与实践

国家图书馆副馆长、全国信息文献标准化技术委员会识别与描述分委员会主任、全国图书馆标准化技术委员会副主任汪东波研究员做了题为“信息组织标准化进展”的报告,对知识组织的相关标准进行了系统的梳理。首先,他按照制定主体、实施效力、标准表述三个维度对现有的标准展开了一个较为具体的分类说明,并纵向分阶段阐述了标准制定的程序。接着,他分别对国际、国内信息组织领域的标准化机构及其制定的标准进行了论述,包括国际标准化组织(ISO)、国际图书馆协会联合会(IFLA)、我国的国际标准化管理委员会,中央政府和地方政府相关部门,学协会、联合会等社团,企业或企业联盟,技术委员会等。最后,他对比了国际国内信息组织领域的若干代表性标准,并对具体标准之间联系进行了详细的阐述,包括对“文献工作——文献审读、主题分析与选定标引词方法”“信息与文献——索引内容、编制和表述指南”以及电子政务、军队机关公文、中文新闻信息等专业领域的标准进行了——阐述。

国家图书馆《中图法》副主编卜书庆研究员以“《中图法》的发展历程与发展方向探讨”为题,总结了《中图法》在不同发展时期的特点及成果。她将《中图法》发展分为“创建、统一、一体化(1975—1999)”“机读化、电子化、网络化(1999—2009)”“语义化、关联化、可视化(2010—2013)”“自动化(智能化)、最终用户服务化(2014至今)”4个阶段,每一阶段有不同的研发重点和成果特征。目前《中图法》对数字资源、多媒体信息、动态信息组织的能力较差,其自动标引、自动分类、自动维护方面还不能满足用户需求;虽然局部采用分面分析、多重列类、交替类目等技术建立立体的类目结构,但总体上是一个线性体系,提供检索途径单一;其易用性、交互动态性、人性化等方面也都有待进一步改善。针对这些问题,她提出了《中图法》最终用户版即分类搜索引擎版的系统构建设想,希望通过分面分析法,对《中图法》等级列举式体

系结构进行改造,使其形成多面并列且可后组配的等级体系,来满足最终用户(读者、研究者)从多角度、多属性分类检索资源等需求。

武汉大学信息管理学院副院长、中国图书馆学会学术研究委员会副主任黄如花教授以“面向多源数据的信息资源整合”为题做报告。她首先对多源数据融合的必要性进行了解释,国家政策的支持与“云物数智”技术的发展应用为图书馆将丰富的开源数据纳入已有的封闭数据,实现信息资源整合提供了条件。开源数据的来源广泛,包括国际组织,各级政府数据门户,公共部门、非营利组织,企业,文化机构,大学(科研院所),家庭、个人等。而后,她以案例展示的方式重点阐述了数字信息资源的整合方式,在分类和编目方面,她展示了世界银行、世界卫生组织的数据目录;在导航或学科指南方面,她介绍了伯克利大学等机构利用政府数据、科学数据等开放数据资源制作的导航以及基于导航之上的开发与应用;在数据门户方面,她展示了英国数据服务发现网、Data.gov、哈佛大学建立的融合了封闭资源和开放资源的数据平台。她指出开源的系统应该作为未来资源整合的方向,开放资源和封闭资源都是资源整合的对象,借助关联数据、语义网、人工智能等技术,在图书情报人员、企业、政府机构、用户等开放社群的共同参与之下,才能有望实现信息资源整合的数字化、数据化乃至智慧化发展。

中国人民大学信息资源管理学院图书情报教研室主任周晓英教授以“政府健康信息治理与信息构建”为题,论述了信息组织、信息构建在政府健康信息治理中的应用。她认为,提升国民健康素养是世界各国推进健康国家建设中面临的共同问题,全民健康素养的提升和维护,需赖于健康信息的支持,这离不开信息平台和信息工具的支持,只有内外因相结合才能够提升健康素养。人民的健康是“健康中国”建设的中心,信息在人民健康的建设和维护中发挥着举足轻重的作用。由于社会传播的健康信息质量堪忧,公民健康素养低,信息碎片化严重,信息无法作为资源直接利用,因此健康信息需要治理。健康信息的核心是信息质量,但目前通过互联网搜索到的健康信息、医疗产品知识、疾病知识、治疗建议、

医疗咨询、在线问诊等丰富的健康信息常常缺乏精心的组织,而且还存在判断真伪的风险。最后,她归纳出四种政府健康信息治理策略,包括通过行政手段进行预防、监督和处理;通过多种形式的教育以提升国民的健康素养;评估筛选健康信息,导向高质信息;建成健康信息服务体系,提供公众利用。健康信息构建是实现健康信息治理的主要手段,应该建立起可信、易用、用户体验佳、服务和建设可持续的健康信息传播与服务平台。

南京理工大学经济管理学院副教授薛春香的报告以“在线学术社交平台信息组织调查与分析”为题。她的研究思路源自Nature对科学家使用社交媒体进行学术行为的调查,由此开始对学术社交平台的发展脉络进行梳理,将学术社交平台分为研究内容分享、学术资源共享、学术成果交流等三个类型。她选取Academia.edu、ResearchGate、Mendeley和科学网,对其信息服务对象、信息服务内容、信息资源构成、信息流、信息组织模式进行实证分析,总结出当前学术社交平台信息组织呈现的特点。最后,她对学术社交平台信息组织提出了四点建议:信息组织方式多元化,解决信息与信息、人与信息、人与人之间的关系和链接问题;更加注重用户参与的信息组织,提供用户参与机制和利用手段;个人信息组织和平台信息组织结合;信息自组织与他组织结合。

4 知识组织工具发展研讨

本次会议对知识组织工具的发展进行了研讨,研讨的主要内容包括《中国图书馆分类法》(简称《中图法》)和《中国分类主题词表》(简称《中分表》)的相关修订和维护问题。讨论了《中图法》第五版的修订尤其是A大类的修订问题,演示并讨论了《中图法》第五版电子版和基于互联网环境的《中图法》与《中分表》修订维护系统的开发成果,研究了《中分表》新增主题词的规则和相关问题。

关于《中图法》第五版修订问题,与会专家肯定了《中图法》第五版启动修订的必要性。但是,有专家提出应该先确定好修订的方针原则,再来讨论细节性问题。对“习近平新时代中国特色社会主义思想”在《中图法》中如何设类的讨论非常热烈,与会专家一致同意在《中图法》中应该设立相应类目,至于

是在A大类设立类目还是在D大类设立类目,以及设立类目的具体方法,应该尽快启动研究。

对《中图法》第五版电子版和基于互联网环境的《中图法》与《中分表》修订维护系统,在观看了成果演示之后,与会专家在给予肯定的同时,也对界面设计、功能优化提出了建设性的意见。其中,有专家对《中图法》与《中分表》修订维护系统指出了应该增加类目拖动调整的功能,系统应该满足多人在线编辑对处理效率和反应速度的较高要求。另外,有专家认为该系统的操作页面过于文字密集,在实际工作中会给编目人员带来较大的困扰,建议对界面和交互设计进行优化。最后,相关人员对《中分表》新增主题词的规则进行了说明,大家就实际工作中遇到的问题 and 解决办法进行了充分的讨论,达成了共识。

5 结语

作为图书馆学、情报学的核心领域,知识组织经历了传统环境、网络环境,当前已经开始进入大数据环境。在“知识组织与检索语言学术研讨会”中,与会者对大数据时代知识组织所面临的新挑战和新课题做出了敏锐的反应,虽然相关的探索还是初步的。大数据环境中的大数据理念、大数据方法以及各类智能技术给知识组织带来了机遇和挑战,面对庞大、繁杂、急剧膨胀的大数据,知识组织的对象、原则、方法和技术都发生了或大或小的变化。如何在大数据环境中抓住机遇、应对挑战,发扬传统、适时创新,适应知识组织泛化和细化的要求,推进知识组织的发展,发挥知识组织的重要作用,需要各位同仁的共同努力。

Knowledge Organization and Retrieval Language Seminar Summary

Cao Shujin Liu Huiyun Zhang Lele Chang Zhaoxin Li Huiling Wang Yaqi Chang Jingwei

Abstract: Aiming to promote the field of knowledge organization and retrieval language research, strengthen researchers' communication and cooperation, and enhance China's the core competitiveness and influence in the field of knowledge organization, the knowledge organization and retrieval language seminar which sponsored by the China library association academic research committee, and organized by the school of information management of Sun Yat-sen University and Library society of China's academic research committee was held in Guangzhou, during December 2-5, 2018. With the theme of "Innovation and development of knowledge organization methods and tools in the era of big data", the seminar invited some famous experts in the field of knowledge organization to share their current research findings and thoughts. Besides, the theoretical experts and the practical workers also discussed the development of knowledge organization tools. This paper tried to report the conference details so that the domestic scholars and practitioners from the field of library science and information science can quickly know the latest research trends, the frontier work, and future development directions in the field of knowledge organization.

Key words: Knowledge organization; Retrieval language; Seminar review