

基于文献数据规律的 机构知识库关键技术研究

——以北京邮电大学机构知识库为例

周 婕¹ 陈嘉勇^{1,2} 李 玲¹ 侯瑞芳¹

(¹北京邮电大学图书馆 北京 100876;

²北京师范大学政府管理学院 北京 100875)

摘要 文章针对目前国内高校图书馆构建机构知识库普遍存在关联学者、机构、学科或主题不精准等问题,以北京邮电大学图书馆通过研究文献数据预处理的关键技术来探索文献数据中的规律为例,设计了文献实体关系模型并进行高校实体的扩展,在图书馆信息化管理的顶层设计下自主研发了机构知识库。通过认领作者机构的新模式,提出了文献与实体之间可持续的间接精准关联机制,为学术生态圈的形成为和深层次学科服务的开展提供了数据支持。

关键词 机构知识库 数据模型 关键技术 文献计量学 北京邮电大学

Research on the Key Technology of Literature Data Regularity- based Institutional Repository:

A Case Study of Institutional Repository in Beijing University of Posts and Telecommunications

Zhou Jie¹ Chen Jiayong^{1,2} Li Ling¹ Hou Ruifang¹

(¹Library, Beijing University of Posts and Telecommunications, Beijing, 100876;

²School of Government, Beijing Normal University, Beijing, 100875)

Abstract Aiming at the problem of inaccurate data correlation with scholars, institutions, subjects or topics which is widespread in institutional repositories of domestic university libraries, Library of BUPT develops key technologies of preprocessing to explore the regularities in bibliographic data, and designs paper- entity relationship model which is expanded by colleges' entities, and develops institutional repository based on top- level design of information management. By a new mode of identifying author affiliations, a sustainable correlation mechanism for indirect and accurate correlation between paper and entity is proposed, and accurate data correlation provides the support for academic ecosystem and profound subject service.

Keywords institutional repository, data model, key technology, Bibliometrics, Beijing University of Posts and Telecommunications

1 引言

随着近年来开放获取理念的普及和 DSpace 等开源软件的兴起,国内很多高校图书馆开始尝试机构知识库(Institutional Repository, IR)的构建。机构知识库一方面可帮助高校长期保存学术成果、促进学术传

播、提高品牌声誉;另一方面也可辅助科研人员提高成果可见性和学术影响力,并逐渐适应社交环境下的学术交流。对于图书馆加强资源保障、提供深层次学科服务而言也有着重要意义。

根据 OpenDOAR 的统计,截至目前,中国大陆和港澳台地区已有上百所高校和科研机构正式注册机

本文系国家社会科学基金项目“基于多方法融合的中外图书馆学情报学知识图谱实证研究”(编号:11BTQ019)的研究成果。

构知识库^[1],很多高校也开始了探索,并形成联盟。然而,这些年来国内的机构知识库基本上仍是沿袭特色库的思路,并以文库、学者库、机构库命名,在数据来源的获取与组织、数据模型的设计以及服务模式的可持续发展等方面没有统一的标准,具体表现为主要关注数据收割和全文上传,而数据质量参差不齐,关联机构、学者、学科或主题不精准,缺少社交和挖掘等功能;高度依赖开源软件而二次开发不足,缺乏团队管理运营,没有走出一条能让机构知识库活跃于高校学术生态圈中的持续发展道路,国内高校机构知识库的发展到了瓶颈期^[1]。北京邮电大学图书馆在深入分析机构知识库现状的基础上,研究机构知识库的关键技术,拟走出一条可持续发展的道路。

2 研究现状

回到机构知识库的源头,学术出版与学术资源联盟资深顾问 Raym Crow 于 2002 年提出了“典藏、展现说”,认为机构知识库是某个高校把其师生创造的数字知识进行永久保存,以用来提升声望^[1]。美国网络信息联盟常务董事 Clifford Lynch 则认为机构知识库是“服务体系”,是高校为成员提供的管理、传播自己创造的数字资料的系列服务^[4]。因此,机构知识库不止是一个能让学术成果长期保存的数据库,而更是一套包含典藏、管理、运营和服务在内的框架和模式^[5-9]。存储只是基础,由专业化的团队管理、运营和服务才能让机构知识库达到生命周期的高峰,对库中蕴藏的高质量数据进行深层次的分析和挖掘来支持高校决策才能让机构知识库的作用发挥到极致。

数据来源的获取与组织是机构知识库构建的基础,在建设初期需要获取并导入历年来学术成果的元数据,并且在建设过程中可持续地导入新成果。一份关于中国机构知识库建设调查分析报告指出,大部分高校图书馆更倾向于从其他专业数据库平台批量导入数据^[1]。多所高校图书馆进行了数据转换的尝试,例如北京科技大学图书馆将 SCI、EI 等题录数据转换为 DSpace 要求的 XML 格式^[3],北京工业大学图书馆选择了高校图书馆已有的科研管理系统、学位论文库的数据,并借助 NoteExpress 提交到机构知识库中^[9]。批量导入数据无疑是高校图书馆青睐的做法,但是数据来源的选择对数据质量有着重要影响,数据预处理时需要转换成中间文件或是借助其他工具完成,增加了机构知识库的管理和运营难度。因此,一步到位的通用题录转换工具和元数据模型是机构知识库需要具备的,有了转换工具和数据模型才能将半结构化的纯文本题录数据向结构化的关系数据库格式进行完整转换。

高质量的元数据需要灵活的模型和完善的平台来支撑数据并提供服务。近年来机构知识库平台的研

发或二次开发受到了各领域的关注。在商业领域,一些公司如 CNKI、万方同盛、联想利用其资源或技术优势提供了机构知识库的产品和解决方案,然而资源优势可能导致局限于其本身的资源而忽略了高校的真实需求;在互联网社交领域,于 2008 年上线的 ResearchGate 也引入机构知识库的思路,拥有丰富的论文数据,通过邮件营销来吸引学者认领论文,学者可以联系同行,分享科研动态以及交流想法,其重点在于社交;在科研领域,国内形成了中科院 IR Grid、北大 CALIS、香科大 HKIR 和台大 TAIR 等牵头的机构知识库联盟,技术路线主要是针对 DSpace 的本地化和二次开发,但是部分高校过于依赖 DSpace,自主灵活定制的能力不足,无法开展社交互动或其他扩展功能。

在对国内外高校机构知识库的调研中,本研究发现不论数据来源和平台如何选择,机构知识库普遍存在与资源关联的机构、学者、学科或主题等实体不够精准的现象,这些实体在库中有多种表达,在不同的库中则有更多不同标准的写法,无法精确地定位到某一实体关联的资源。究其原因是有多种表达的机构、学者、学科、主题数据导入机构知识库后会被识别成不同的实体存在,DSpace 等平台中设计的元数据模型不符合需求,因此有必要设计更灵活的数据模型来融合不同的数据来源,在这些来源数据的外部还需要有高校自身的学者、机构、学科、主题等属性的规范实体库,将导入的实体和高校实体相关联。

相关的研究推动着机构知识库的发展,但同时也反映出国内高校图书馆构建的机构知识库还有较大提升空间。

3 关键技术研究

为了解决上述问题,北京邮电大学图书馆以学科服务为契机,借鉴查收查引等工作的经验,基于图书馆信息化管理的顶层设计自主研发机构知识库的关键技术,引入文献计量学,从模型到模式,为机构知识库的典藏、管理、运营和服务实现了底层架构,也为数据的分析与挖掘、社交功能的扩展、学科服务的支持奠定了基础。

3.1 数据来源与预处理

机构知识库中收录的是高校科研人员创造的数字知识,包括期刊文献、会议文献、专著、学位论文、专利、文献预印本、数据集等,这些数字知识需要由高质量的数据和关联实体组成,需要由专业的数据机构来控制数据质量,才能保证机构知识库平台中的精确导航和检索。

Web of Science、EI、CSCD、CSSCI 等数据库被广泛应用于高校的科研成果评价,以这些数据库中收录的期刊、会议论文作为数据来源,既保证了数据质量和核心性、权威性,又因其文摘库的性质不必考虑版权

问题。为此本研究选择科学评价数据库中纯文本格式的文献题录数据作为机构知识库数据来源的突破口,进行数据预处理的关键技术研究。

作者重名是预处理需要解决的关键问题,本研究以相同机构一般不存在同名作者为前提假设,从题录数据中识别出作者与机构之间的关系,将作者与机构的组合视为作者实体。除了题录数据中的主题词和标引词之外,标题和摘要等自然语言中的专业术语和语义信息也具有相当的价值,需要使用自然语言处理的相关技术进行分词、去除停用词、词归一化,抽取标题和摘要中的词。此外,对题录数据中有意义的 DOI 以及地理数据也不应忽略。

从不同来源批量导出的题录数据一般是用不同结构的纯文本文件记录的,字段标识有所区别,字段值也可能用了不同的表达或分割方式。本研究针对不同数据库的格式标准,分别设计出适用于 Web of Science(包含 CSCD)、EI、CSSCI 这 3 类格式标准的正则表

达式,以及自动识别和匹配的方法,该方法用于抽取题录各字段的数据,不需要生成中间格式或借助其他工具,降低了数据来源定期获取与导入的难度。

3.2 文献实体数据模型设计

虽然来自数据库的纯文本格式的题录数据使用了不同的格式标准,但是文献与学者、机构、学科和主题等实体都隐藏在文献题录数据中,并且都能从中提取出文献与这些关联实体间的直接关系,进而获得实体间的间接关系,以及深层次的网络关系。本研究对 Web of Science 的题录格式做了深入分析,并提出了文献实体关系的通用模型^[8],实现了从半结构化纯文本文献数据向结构化关系数据库格式的完整无损转换,该模型为机构知识库融合不同来源格式的文献数据提供了保障,也为机构知识库数据的深层次分析挖掘奠定了基础。

如图 1 所示,文献实体关系模型中的文献实体由代表文献特性的字段组成, AU、BE、CA、RID、RP、EM

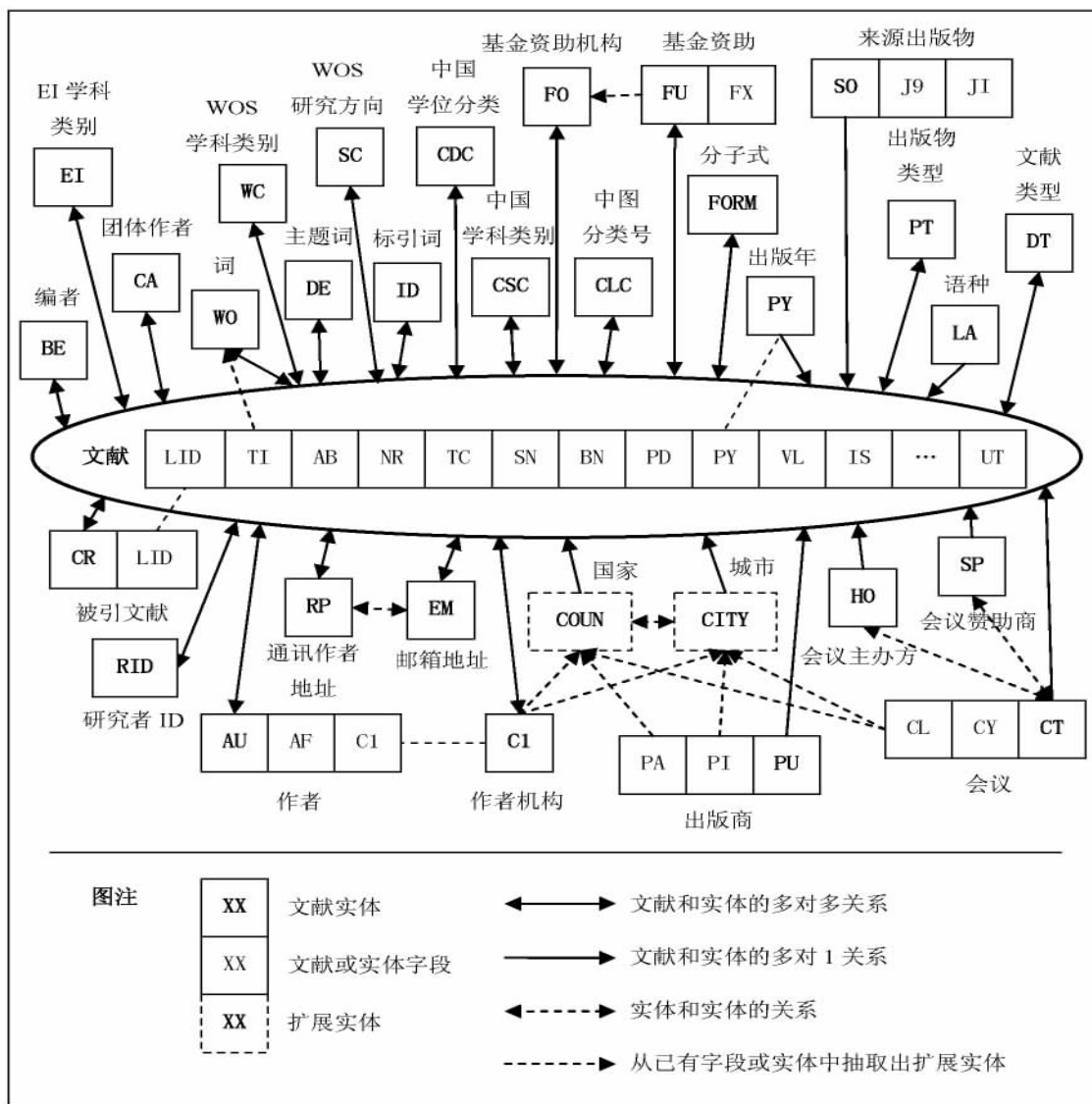


图 1 文献实体关系模型

等实体记录了相关学者信息, C1、RP 实体记录了机构信息, DE、ID、WO、WC、SC、CLCN、FORM、PT、DT、LA 等实体记录了专业术语或标识, FO、SO、CL、PA、SP、HO 等实体记录了相关基金或组织信息, CR 记录了文献间的引用关系, PY 为所有实体打上了时间标签。

3.3 高校规范实体模型扩展

机构知识库要做到论文与机构、学者、学科和主题等实体的精准关联, 需要解决这些实体在不同数据库中不同表达的问题, 这是高校图书馆构建机构知识库面临的主要难点。

在题录数据中, 作者和机构根据不同期刊的投稿要求或数据库的格式标准呈现了不同的表达方式, 如姓与名的前后位置, 姓名与地址的全称与缩写, 名称中间的分隔号, 中英文数据库中相同作者的不同语种表达方式, 以及拼写错误; 不同数据库中使用了不同的学科分类体系, 如英文数据库普遍使用的 Web of Science 学科类别 (Subject Category)、Web of Science 研究方向 (Research Areas)、EI 学科类别以及中文数据库普遍使用的《中国学科分类与代码国家标准 (GB/T 13745-2009)》、《中国学位授予和人才培养学科目录 (2011 年)》和《中国图书馆分类法 (第 5 版)》等体系。此外, 不同数据库的标引词与主题词等关键词字段也有明显的区别。

这些表达各异的数据导入机构知识库后会被识别并保存成不同的实体, 如图 2 所示, 在 SCI 和 EI 数据库中有两种不同写法的机构都属于北京邮电大学的理学院。同样, 可能会有上百条类似但格式不同的作者与机构的组合实际上是同一人, 某个学科在不同学科体系中可能以不同的名称存在。

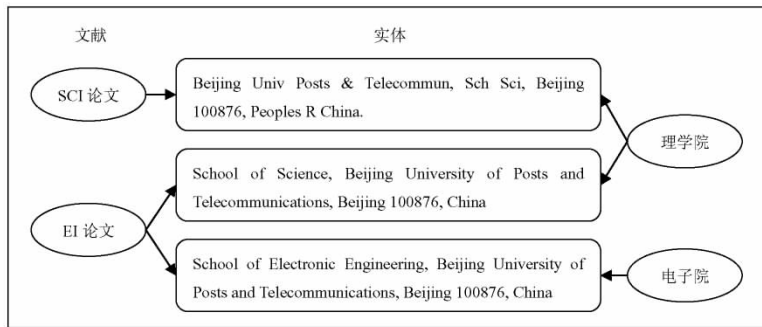


图2 论文与机构的关联, 其中多个机构均为理学院

文献实体关系模型虽然完整地保存了文献和实体的数据和关联, 但还是无法做到与高校的某学者、机构等高校规范实体的精准关联, 因此非常有必要对其进行扩展, 设计出更灵活的数据模型来融合不同的数据来源, 即在来源数据的外部关联高校自身的机构、学者、学科、主题等真实数据的高校实体, 将题录实体和高校实体相关联。

如图 3 所示, 扩展后的文献实体关系模型由文献实体、题录实体与高校实体三层组成。图 2 所示的文

献实体关系模型在图 3 中简化了文献实体层与题录实体层, 它们的数据与关系来源于对题录数据的预处理。高校实体层来源于高校信息网络中心的统一标准或接口数据, 题录实体层与高校实体层之间的关系由学者或科研秘书主动认领来确定, 或由学科馆员人工关联。

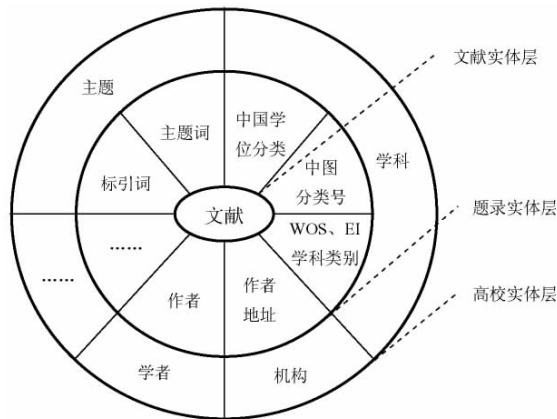


图3 扩展高校实体后的文献实体关系模型

经过扩展了高校实体后的文献实体关系模型具有扩展性和灵活性, 文献实体层与题录实体层之间的关系完全由科学评价数据库提供的题录数据完整无损地转换而成, 而题录实体层与高校实体层之间的关系由学科馆员灵活地设置。如果学者调整了所在机构, 或者高校进行了机构重组, 则只需在高校实体层中进行内部调整, 不会影响到文献实体层与题录实体层。用户在机构知识库中看到的机构与学者等信息来自于高校实体层, 不会受到题录实体层中有着多种表达的实体数据干扰, 机构和学者关联的论文则是根据高校实体层、题录实体层与文献实体层逐级关联的结果。

虽然理论上可以有更彻底的预处理技术将题录数据进行批量转换或替换, 让文献实体与高校实体直接关联, 从而免去题录实体层。然而, 本研究非常强调题录数据的完整无损转换, 不推荐使用批量替换或其他任何方法破坏科学评价数据库所提供的题录数据, 目的在于记录高校科研成果在科学评价数据库中的原貌, 探索海量文献数据中的规律, 让新的题录数据按照自然形成的规律对应到学科馆员所关联的高校实体中。

3.4 系统架构设计

关键技术的解决扫除了构建机构知识库的障碍, 北京邮电大学图书馆基于管理信息化理念的顶层设计自主研发了机构知识库, 与信息管理平台 and 门户网站使用了统一架构, 实现了与图书馆信息化管理平台和门户网站的有机结合。

北京邮电大学机构知识库的基础框架将扩展高

校实体后的文献实体关系模型作为元数据模型来典藏、管理与运营机构知识库的数字资源,并支持深层次的挖掘分析和学科服务工作。同时,机构知识库还使用业务流程引擎、邮件接口、微信接口以及社交平台接口来支撑学科馆员基于学科服务的协同工作,以及机构学者在社交网络模式下的分享协作。

在机构知识库的用户界面,本研究设计的便捷导航可以让用户方便地通过数据来源、语种、学科、机构、年份、学者、主题等实体与文献的直接或间接的关联关系,快速检索出所需文献的列表,用户界面设计如图4所示。

来源	SCI	SSCI	A&HCI	CPCI-S	CPCI-SSH	EI	CSCD				
语种	中文	英文									
学科	通信电子	计算机	数学	物理	光学	经管	法律	政治	语言	文学	图情
机构	信通	电子	计算机	自动化	软件	数媒	经管	人文	马研	公管	理学院
国际	网教	民教	网研	光研	感研	图书馆	体育部	世纪			
年份	2014	学者	全部	主题	optical communication	搜索					
1. Bifurcation analysis and solutions of a three-dimensional kudryashov-sinelshchikov equation in the bubbly liquid 田播 2014-04-01 SCI											
2. Power dependent pulse delay with asymmetric dual-core hybrid photonic crystal fiber coupler 任晓敏 2014-02-01 SCI											

图4 机构知识库用户界面设计

4 应用分析

北京邮电大学机构知识库在保存了文献数据原貌的同时,也尝试着探索隐藏在数据中的规律来帮助实现文献的精准关联,并为学术社交和学科服务提供数据支持。

4.1 题录数据呈现二八定律

本研究将北京邮电大学被 Web of Science、EI、CSCD、CSSCI 收录的文献数据导入到机构知识库中,文献数据量如表1所示。

表1 北京邮电大学文献数据量统计(截至2014年4月15日)

文献总数为38723,但其中存在1篇论文同时被多个数据库收录的情况,实际的论文总数为31244。本研究更关注的是与文献关联的题录实体的情况,将题录实体关联的文献实体数量定义为题录实体频次,以机构实体为例,如图5所示。从图的曲线中我们可以发现,机构实体的分布符合二八定律:约80%的机构实体频次来自于约20%的高频机构实体中,另外约20%的机构实体频次来自于约80%的低频机构实体中。

根据文献数据规律中出现的二八定律现象,在面对机构知识库中识别出来的海量题录实体

数量并不感到无从下手,因为只需重点关注20%的高频题录实体与高校实体的关联,就可以基本上展现文献题录数据的全貌。以作者机构为例,高频作者机构一般是正确的书写形式,未来新收录文献的题录实体也很可能落入到学科馆员已经关联过的高频作者机构中,低频作者机构则一般来自于科研秘书不推荐的书写形式或其他院校的实体数据。但是由规范体系中生成的题录实体(如学科分类),不论高频还是低频,都是不可忽略的,低频的学科分类可能是一个高校需要扶持或新兴的学科分支。

4.2 认领模式实现精准关联

为了辅助教学、科研、人事等部门的相关工作,实现论文与学者的精准关联,目前高校图书馆的馆员普遍采用学者自行认领论文的模式,让学者定期自行识别哪些文献归入自己名下,随后馆员进行审核。然而,这种被动等待学者来认领的模式需要相关部门的制度推动,也需要机构知识库在新论文入库时有能自动推送给学者进行认领确认的功能,否则很难落实。

论文是随着科学发展不断产生的,每天都有学者的论文被数据库收录。然而,学者和所属机构是相对不变的,本研究主张“以不变应万变”的思路,基于文献实体关系模型,让题录实体作为文献实体与高校实体的桥梁,让学者自行认领作者与作者机构的组合,让科研秘书对机构认领作者机构,或者让学科馆员主动关联作者实体与学者、作者机构实体与机构,这种方式能持续有效地解决文献实体与高校实体的关联,实现文献与学者、机构等实体之间可持续的间接精准关联机制。

4.3 社交互动形成学术生态圈

本研究在机构知识库中引入营销和社交等互联网因素,基于作者机构的认领模式,对新收录的论文自动识别关联的学者与机构,同时还挖掘出学者个性

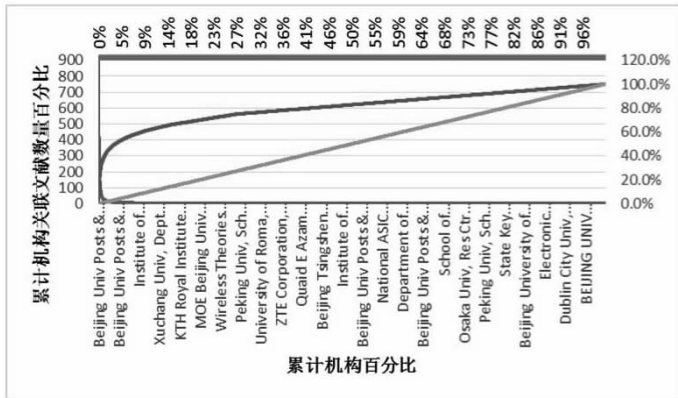


图5 机构实体关联文献数量的二八定律分布

化的学术需求(如文献中高频使用的关键词、经常投稿的期刊、经常参加的会议等),通过邮件和微信营销等方式将学者最新成果的收录情况以及校外相关研究动态推送给学者,吸引学者到机构知识库中吸收最新科研动态,了解校内外最新科研成果,并将自己的研究成果的全文保存在平台中。

此外,在机构知识库的社交网络中,高校师生通过实验室圈子、感兴趣的领域关注学者,以获得被关注学者的最新动态,并且能和学者在社交环境下进行学术交流。

4.4 数据支持学科服务

为了挖掘出机构知识库的潜力,北京邮电大学图书馆充分利用机构知识库数据的价值,识别高产和活跃学者,有针对性地提供深层次的学科服务。学科馆员开展的学科服务工作能促进机构知识库的构建,同时学科馆员利用机构知识库的数据能更好地开展深层次的学科服务,让学科服务和机构知识库构建互相促进。

5 结语

北京邮电大学图书馆根据查收查引等工作的经验,长期研究机构知识库的关键技术来探索文献数据中的规律,并设计模型付诸实践,走出了一条可持续发展的道路。北京邮电大学机构知识库在长期保存学术成果的同时,今后还要重点发挥学科服务团队中学科馆员的作用,基于文献数据深入开展学科服务、为人事科研部门提供数据支持、开展学科分析辅助高校决策,形成一种良性循环的机制,让机构知识库活跃于北京邮电大学的学术生态圈。

参考文献

- [1] The Directory of Open Access Repositories – OpenDOAR [OL]. [2014-09-01]. <http://www.openoar.org/>
- [2] 张巧娜. 我国大陆机构库实践的“冷现象”研究[J]. 大学图书馆学报, 2010(6): 48–52.
- [3] Crow R. The case for institutional repositories: A SPARC position paper [R]. ARL Bimonthly Report 223, 2002.
- [4] Lynch C A. Institutional repositories: Essential infrastructure for scholarship in the digital age [J]. Libraries and the Academy, 2003, 3(3): 327–336.
- [5] 张智雄, 高嵩. 机构仓储及其在数字图书馆服务中的应用模式研究[J]. 图书情报工作, 2006(8): 59–62.
- [6] 赵晓晔. 大学机构知识库研究[J]. 图书情报工作, 2009(5): 5–7.
- [7] 聂华. 中国机构知识库建设调查分析报告[R]. 中国机构知识库推进工作组, 2013.
- [8] 李国俊. 基于元数据的高校机构知识库建设研究——以北京科技大学机构知识库为例[J]. 大学图书馆学报, 2012, 30(4): 55–60.
- [9] 邓红. 高校机构知识库建设实践与探索——以北京工业大学图书馆为例[J]. 现代情报, 2013, 33(7): 80–83, 129.
- [10] 肖明, 陈嘉勇, 李国俊. 文献计量系统的文献实体关系通用模型研究[J]. 图书情报工作, 2012, 56(23): 129–134.

[作者简介] 周婕,女,1974年生,北京邮电大学图书馆信息咨询部主任。

陈嘉勇,男,1987年生,北京邮电大学图书馆馆员,北京师范大学政府管理学院兼职讲师。

李玲,女,1980年生,北京邮电大学图书馆馆员。

侯瑞芳,女,1979年生,北京邮电大学图书馆馆员。

收稿日期:2014-09-03

图书、情报、信息、资料工作者自己的刊物

欢迎订阅《情报资料工作》全文数据库

中国人民大学书报资料中心现隆重推出《情报资料工作》回溯数据库。数据库以一张光盘形式提供。1980年—1994年数据报价为340元。1995年后每季度更新数据,全年更新费为130元。

该数据库可以全文检索,检索结果可以复制、拷贝、打印,或者根据用户的需求进行再编辑。

联系单位:中国人民大学书报资料中心

地址:北京9666信箱市场部

联系电话:010-82503412/38/40 62512171

邮政编码:100086

户名:中国人民大学书报资料中心

账号:344156031742

网址:www.zlzx.org

开户银行:中国银行北京人大支行