大数据环境下的农业知识发现服务探索*

赵瑞雪，鲜国建，寇远涛，顾亮亮，朱亮
（中国农业科学院农业信息研究所，北京 100081）

摘要：适应大数据环境下用户新需求，探索新型知识发现服务形态，是图书馆等信息机构提升知识服务能力面临的挑战与机遇。本文在简要分析国内外资源发现与知识发现系统等研究进展基础上，设计大数据环境下农业知识发现服务体系架构，并从农业综合科技数字知识仓储构建、基于元数据搜索的统一发现、基于语义多维知识关联发现、基于深度融合知识挖掘分析和面向特定需求的专题知识发现等方面阐述农业知识发现服务的研究探索。

关键词：大数据；资源发现；知识发现；资源汇聚；知识服务
中图分类号：G254
DOI：10.3772/j.issn.1673-2286.2016.9.005

1 引言

大数据时代，各领域对数据的挖掘与分析日益深入，海量数字信息资源已成为科研用户依赖与使用的主流资源，以数据密集型计算为特征的科学研究“第四范式”方兴未艾[1]。面对日益复杂的科技创新问题，科研用户的信息需求也从单一文献信息向多元综合信息、从简单文献资源发现向细粒度知识单元及知识发现演变。大数据环境下，如何适应和满足用户新需求，以科技文献资源为主体，更加合理、高效地汇聚融合多源异构科技信息大数据资源，并与多类型、多层次知识发现技术相结合，研究构建大数据驱动的新一代知识发现服务系统，是图书馆等专业信息机构面临的挑战，也是优化知识服务体系、提升服务质量的重大机遇[2]。

本文在简要分析国内外资源发现与知识发现系统等方面研究进展基础上，设计大数据环境下农业专业领域知识发现服务的体系架构，并从农业综合科技数字知识仓储构建、基于元数据搜索的统一发现、基于语义多维知识关联发现、基于深度融合知识挖掘分析和面向特定需求的专题知识发现等方面阐述农业领域知识发现系统的研究实践进展。

2 国内外现状

为深入整合图书馆资源及其服务体系，向用户提供从资源发现到资源获取“一站式”服务，Serials Solutions公司于2009年发布全球第一个网络级资源发现系统Summon[3]。随后OCLC、EBSCOhost等数据库生产商分别推出WorldCat Local[4]和EDS[5]，而图书馆集成系统开发商ExLibris公司也发布Primo Central[6]。与提供新型信息资源整合服务的资源发现系统相比，学术搜索引擎则以学术资源为索引对象，涵盖商业元数据、互联网免费学术资源和深层网页学术资源，并将这类资源抓取、索引以统一的界面向用户提供搜索服务[7]，如Google Scholar、微软学术搜索、百度学术搜索等[8]。

一般而言，资源发现系统擅长资源结果的准确定位及综合服务，而学术搜索引擎侧重资源线索的揭示。大数据时代，用户不再满足于简单信息检索和资源发现，而迫切需要从海量信息中发掘更具价值的规律和知识。知识发现是从大量数据集合中抽取或提炼潜在、有用知识的过程。近年来，国内外关于知识发现系统的研究不断升温，2015年，欧洲图书馆协会发布的《数字时代知识发现海牙宣言》指出，内容挖掘、数据抽取工具不仅可以处理大数据，也是数字时代知识发现的关键[9]。

* 本研究得到“中国农业科学院科技创新工程”项目（编号：CAAS-ASTIP-2016-AII）资助。

生物医学领域已基于GO本体和MeSH主题词表开发了语义知识发现工具GoPubMed[10]；百度学术在2015年中国高校图书馆发展论坛上发布“高校图书馆计划”，致力于提供知识发现、连接用户与图书馆的信息服务[11]；基于文献资源的知识发现系统有“中国学术搜索”“超星发现系统”“智立方发现系统”“学知搜索”等[12]。

上述资源发现系统和学术搜索引擎作为全新的学术信息发现工具，正在以“简单、快速、易用、有效”的创新资源组织方式、全新商业模式颠覆传统图书馆服务理念，带给用户全新的体验[13]。知识发现系统基于快速增长的海量数字资源，通过现代技术手段将资源整合、知识发现、信息推送等服务融为一体，打破以往书刊目录、文献索引和全文获取的局限，为用户提供具有知识挖掘与数据分析功能的知识发现系统，从而实现从资源发现到知识发现的转变[14]，显著提高数字资源利用率和知识服务能力[2]。

3 农业知识发现服务体系架构

通过国内外现状分析不难发现，尽管上述资源或知识发现服务系统已取得重要进展，但面向专业领域服务时，在科技信息资源覆盖类型、资源深度组织与关联、个性化专业化服务，以及线上、线下协同服务等方面还存在不足。近年来，在推进“三农”和现代农业发展过程中，农业科技创新支撑作用日趋明显。大数据环境下的农业科技创新工作，对农业科技信息资源保障与知识服务提出新期望和新要求。如何主动适应大数据环境下农业科研创新和管理决策的用户需求，系统汇聚、有效整合与挖掘利用多源异构科技信息大数据资源，构建大数据驱动的农业专业知识发现服务系统，已成为农业信息机构服务于农业科技创新的当务之急。

本文探讨的农业知识发现服务，是依托并整合国家农业图书馆海量文献资源、各类服务系统和专业人才队伍，在借鉴和集成第三方资源/知识发现系统基础上，研究构建的服务于农业科研创新的新一代农业知识发现系统。该系统旨在全面汇聚与知识化组织融合的多源异构农业领域海量数据资源，实现资源统一搜索与关联发现，加强基于大数据的挖掘分析和知识计算，使得在面向不同用户群体时，能提供专业化、个性化、动态化和集成化的知识发现增值服务。该系统体系架构分为四个层次：多源异构资源层、资源汇聚组织层、知识挖掘分析层、知识发现服务层（见图1）。

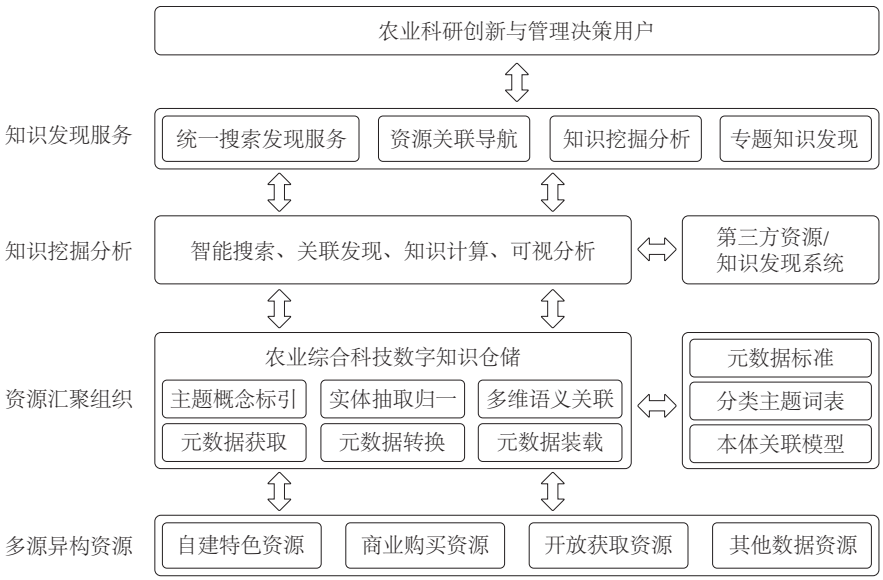


图 1 农业知识发现服务系统体系架构

其中，多源异构资源层是构成整个系统的数据基础，其可整合利用的资源种类及规模都在不断扩大，农业科技信息大数据格局正在快速形成；资源汇聚组织层遵循统一元数据标准实现元数据汇聚，并基于分类主题词表和本体关联模型实现知识组织与多维语义关联，形成农业综合科技数字知识仓储；知识挖掘分析层是系统的核心部分，通过集成应用智能搜索、关联发现、知识计算和可视分析等关键技术，整合Primo

等其他第三方资源/知识发现系统，以最大程度实现仓储中各类资源的可见性和可获得性，并从中挖掘有价值的知识；知识发现服务层面向农业科研创新和管理决策的用户需求，提供人机友好交互的统一搜索发现、资源关联导航、知识挖掘分析和专题服务等知识发现服务。

4 农业知识发现服务探索重点

农业知识发现服务旨在集资源、技术、知识和服务于一体，实现农业信息资源整合并为用户提供优质高效的知识服务。本文重点从以下方面论述研究实践。

4.1 农业综合科技数字知识仓储构建

4.1.1 综合科技数字资源体系建设

近年来，国家农业图书馆资源建设在加快转型。除做好学术期刊、图书、会议录、学位论文等传统馆藏文献资源建设外，还引进标准、专利等特种文献，开展国家农业科学数据共享中心作物科学、农业区划、动物科学等专业领域科学数据资源的整合，加强政策纲要、科技报告和宏观统计数据（人口、耕地、生产、贸易）等情报资源的收集整理，扩大农业领域专家学者、科技机构、基金项目和学术期刊等规范库建设规模，启动开放获取期刊论文、学位论文、图书、会议录、机构仓储和开放关联数据集等开放资源的采集，实施农业知识百科、专业术语、叙词表和学科分类体系等知识组织体系的修订完善等工作。至此，集商业购买、自主构建和开放获取于一体的多源、异构农业综合科技数字资源体系正在形成。

4.1.2 多源异构资源汇聚与关联融合

为适应大数据环境下用户多样化、个性化、知识化服务，以及内容计算和深度分析的需求，本文综合应用元数据、词表、本体和关联数据等技术方法，对上述多源异构资源开展集成汇聚、知识组织与语义关联。首先，基于统一的元数据描述标准规范，通过元数据获取、转换、装载以及数据互操作访问接口等方式，实现对农业领域科技文献、科学数据、事实型数据、规范库、宏观情报资源等多源异构资源的统一元数据存储，完成资源初级汇聚整合；其次，综合应用分类主题词表和本体关联模型，进行各类资源的规范描述、主题概念与学科分类的标引，以及对科研机构、专家学者等实体对象的抽取和归一[15]，并显性建立资源间多维度语义关联，在此基础上构建农业综合科技数字知识仓储；最后，将各类资源发布为富含语义关联关系的数据和知识网络，并与SFX等资源动态链接技术进行集成，从而将资源整合提升到知识组织与深度关联融合的层次，最终为农业领域知识发现与知识服务，提供一个内容密切关联、多维立体、多层次、网络化的综合科技数字资源保障体系。

4.2 多层次知识发现服务

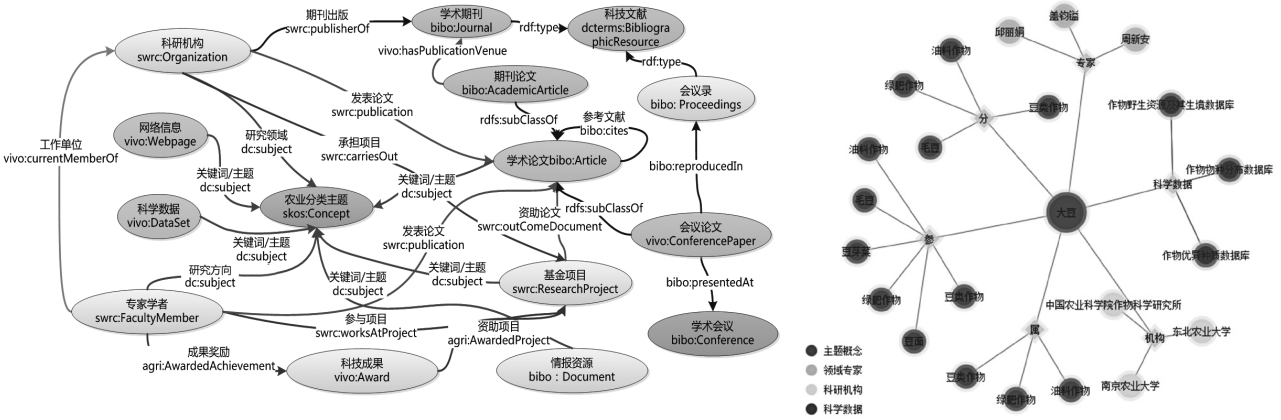
4.2.1 基于元数据搜索的统一发现

提供基于元数据搜索的统一发现是知识发现服务系统的“标配”功能，也是各类资源深度聚合的进一步体现，可提高资源统一发现的水平和能力。为解决异构资源的组织、标引和检索问题，人们开始将搜索引擎和专业知识库相结合。本文基于开源的企业级全文搜索引擎Solr和改进后的中、英文分词器插件IKAnalyzer，建立各类资源元数据的多核索引（Multi-Core）体系[16-17]。其中，在建立索引过程中，综合应用农业科学叙词表、农业百科、联合国粮农组织的多语种词表*AGROVOC*，以及从文献抽取的高频关键词（中、英文）等语料，提高资源切词、分词的准确性和专业性[18]。

目前提供的统一搜索发现服务，初步实现跨库、跨资源、跨语言的一站式检索、多维分面、学科导航、语义扩展等功能。基于词表及词间语义关系，从概念匹配的角度建立语义交互，初步实现对自然语言检索式进行语义浅层理解、分析、匹配，提供相近检索词提示和中、英文智能检索等功能，提供按资源类型、学科分类和TopN等多种分类和排序方式对检索结果进行过滤、聚合与导引，方便用户快速过滤定位所需资源；系统还集成了ExLibris公司动态链接产品SFX，为文献资源建立情景敏感的多种全文获取路径；此外，系统也提供实体命中功能，可根据不同输入主动输出用户关注的核心内容，如对命中的科研人员、科研机构、基金项目等实体对象优先展示，也提供类似百度“框计算”的嵌入式APP深度分析服务结果，如搜索“水稻”，将直接命中产量分析APP，在地图上展示水稻产量统计、预测和对比分析结果。

4.2.2 基于语义的多维知识关联发现

在第二届世界互联网大会开幕式上，习近平指出：“网络的本质在于互联，信息的价值在于互通”[19]。大数据的巨大价值在于依据数据间关联性而建立的复杂关系网络中蕴含的知识[20]。基于科研本体语义关联模型驱动（见图2a），农业知识发现系统实现了“知识立方”和“专家学术圈”等功能，以直观、可视化、多维度立体展示系统中人与人、人与知识、知识与知识、知识与机构、机构与人、机构与机构等资源间的关联关系[21]。例如，检索“大豆”时，“知识立方”模块以检索词“大豆”为中心，检索并可视展示相关概念、专家、机构、科学数据等资源及其相互间关联关系（见图2b）。“专家学术圈”以专家为中心展示Profile、学术论文、基金项目、科技成果等信息，以及合著关系等的可视化关联展示，系统还提供共词作者、共词机构、相似文献等自动推荐功能。



a 科研本体语义关联模型　　b 多维知识关联导航

图 2 基于语义的多维知识关联发现

农业知识发现服务系统通过引入本体模型和简单推理规则，以科技文献为基础，与其他科技资源进行碎片化深度聚合，组合成一个高度集成的信息资源体系，初步实现内容的整合和语义上的无缝链接。通过计算分析处理，可根据知识与检索主题间的语义相关程度为用户呈现结构清晰的知识体系，帮助用户快速形成对相关知识和信息的结构性认识。系统通过将各类资源不同粒度知识单元基于内容和外部特征进行多重关联与揭示，自动扩展相关资源发现的范围，提高资源发现的动态性与完整性，从而为用户提供更加精准和智能的知识发现服务。

4.2.3 基于深度融合的知识挖掘分析

大数据环境下，知识发现服务系统除应具备强大的统一检索和关联导航功能外，还需以结构化、半结构化及非结构化大数据的知识组织及关联融合为基础，综合应用数据挖掘、机器学习和推理技术，通过可视化技术，动态、直观地展现海量信息资源中潜在的规律和发展趋势[22]。

目前，在农业知识发现服务探索中，基于知识计算、文献计量分析、知识脉络分析等方法，初步实现基于学术论文、专利和科技成果等资源的科研产出分析，包括年度产出趋势、研究主题分布、核心机构、核心作者、基金项目等学术成果统计分析，以及科研人员、机构科研能力变化趋势等挖掘分析，初步建立基于数据和图表的动态分析机制和初级技术产品。这些功能有助于用户最直观、快速地了解某一领域的领军专家和核心机构，也有助于同一领域相关学者追踪领域最新研究进展。更具有实践价值和探索意义的挖掘分析，还包括正在尝试的跨领域数据关联打通和深度融合，与相关领域专家深度合作，综合应用农业资源遥感、监测数据、气象数据、宏观经济统计等多领域数据，开展农业区划、作物空间布局、贸易网络、粮食进出口汇率因素等深度挖掘分析（见图3）。

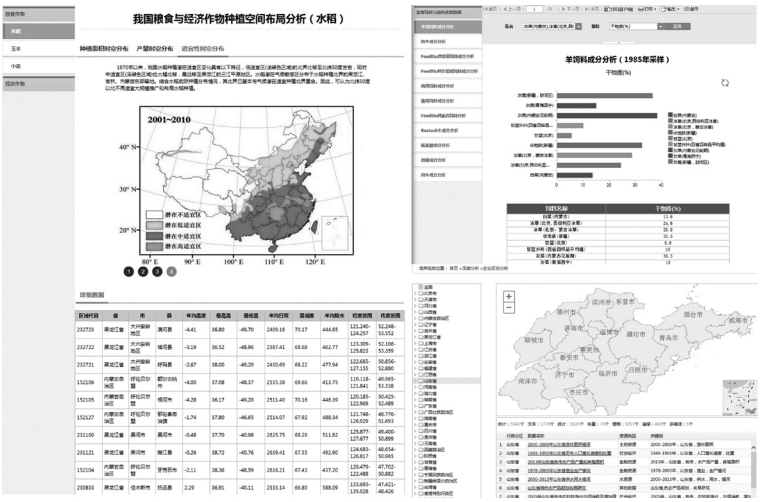


图 3 基于深度融合的知识挖掘分析

4.2.4 面向特定需求的专题知识发现

在数据种类及规模庞大、信息价值密度低的大数据环境下，完全基于系统提供资源和知识发现服务，无法满足不同类型、不同层次用户的个性化需求。因此，图书馆等信息服务机构，须借助于馆场物理空间，基于技术驱动，发挥人类的智慧，三位一体面向特定需求，积极探索提升大数据驱动的知识发现服务在时效性、精确性、全面性和针对性等个性化专题服务方面的能力，这将是信息服务机构应对其他资源发现系统、学术搜索引擎的激烈竞争，体现自身存在价值的有效举措 [23]。

在农业知识发现系统研究实践过程中，本文也在积极探索构建个性化、深层次的知识服务体系，包括面向研究所、课题组和个人的数字化科研信息环境，以及面向学科领域的个性化专题知识服务系统。随着汇聚和挖掘利用多源异构农业大数据资源技术方法的逐步完善，本文开始尝试结合国家重大战略需求和重点领域开展面向特定用户群体的个性化服务。例如，通过选取粮食安全领域，基于大数据资源和农业知识发现系统，经过领域资源动态抽取和技术平台开发，初步构建粮食与食物安全专题服务平台。基于该平台的学科馆员等服务人员可为粮食安全领域的用户提供一站式资源检索与获取、情报分析等深层次知识服务，初步建立“线上+线下”的综合服务模式，并获得用户的充分肯定与好评。

5 结束语

大数据环境下，科研信息环境和科研方式正发生巨大变革，信息过载和选择困惑越来越困扰科技人员，人与大规模数据间的交互已成为重要议题。受限于对各类资源获取的全面性、时效性，以及知识组织和语义关联等基础性、工程化工作的熟练化程度，本文研究构建的农业知识发现系统还有更多工作需要深入推进，包括从语义层面推进农业科技大数据资源的汇聚融合、语义搜索和语义知识发现等技术的应用，实现多源异构农业综合科技资源的全面汇聚、知识组织与关联融合，并紧密结合国家重大战略、农业科技创新和管理决策需求，提供农业综合科技大数据资源保障，以及个性化、深层次、智能化的语义知识发现服务。

参考文献

[1] HEY T,TANSLER S,TOLLE K.The fourth paradigm:Data-intensive scientific discovery [M]. [S.l.]:Microsoft Research, 2009.

[2] 王宁.浅析大数据背景下的数字图书馆知识发现系统[J].图书馆工作与研究,2016(4):58-61.

[3] CICCONE K,VICKERY J.Summon,EBSCO discovery service, and Google Scholar:a comparison of search performance using user queries[J].Evidence Based Library & Information Practice,2015, 10(1):34-49.

[4] GEWIRTZ S R, NOVAK M, PARSONS J. Evaluating the intersection between WorldCat Local and student research[J].Journal of Web Librarianship, 2014, 8(2):113-124.

[5] EBSCOhost Research Databases.Free databases from EBSCO[EB/OL].[2016-08-02].https://www.ebscohost.com/.

[6] Empowering libraries to shape the discovery experience

[EB/OL].[2016-08-02].http://www.exlibrisgroup.com/files/Primo_Brochure-2016.pdf.

[7] 苏建华.图书馆选择资源发现系统的策略分析——以资源发现系统与学术搜索引擎的比较为视角[J].情报科学,2015(6):91-94,105.

[8] 谢奇,李立立,毕玉侠.五大学术搜索引擎比较[J].情报探索,2015(11):42-46.

[9] The Hague DECLARATION.The Hague declaration on knowledge discovery in the digital age[EB/OL].[2015-06-07].http://thehaguedeclaration.com/the-hague-declaration-on-knowledge-discovery-in-the-digitalage/.

[10] 盛东方,孙建军.基于语义搜索引擎的学科知识服务研究——以GoPubMed为例[J].图书情报知识,2015(4):113-120.

[11] 覃燕梅.百度学术搜索与超星发现系统比较分析及评价[J].现代情报,2016(3):48-52,60.

[12] 王悦辰.国内四大中文知识发现系统比较分析[J].图书馆工作与研究,2015(9):42-45.

[13] 曾建勋.资源发现系统的颠覆性[J].数字图书馆论坛,2016(2):1.

[14] 刘江玲.面向大数据的知识发现系统研究[J].情报科学,2014(3):90-92,101.

[15] 夏立新,陈晨,王忠义.基于多维度聚合的网络资源知识发现框架研究[J].情报科学,2016(5): 3-8.

[16] VOHRA D. Pro Docker[M].Berkeley:Berkeley CA Apress,2015:195-218.

[17] 朱明瀚. 基于数据仓库的数据搜索引擎设计与实现[D].上海:华东理工大学,2015.

[18] CARACCIOLO C,STELLATO A,MORSHED A,et al.The AGROVOC linked dataset[EB/OL].[2016-08-02].http://eprints.rclis.org/17010/1/AGROVOC%20Dataset_vFinal_Preprint.pdf.

[19] 习近平.在第二届世界互联网大会开幕式上的讲话[J].中国信息安全,2016(1):24-27.

[20] 刘文远,李少雄,王晓敏,等.大数据知识发现[J].燕山大学学报, 2014(5):377-380.

[21] 张艳新,杨瑜.中文检索平台知识发现功能比较研究[J].情报探索,2016(1):80-84,89.

[22] 王峰,刘燕,王学光.论知识服务中海量数据的知识挖掘与发现[J].情报探索,2013 (8):43-45,49.

[23] 杨亮,雷智雁.大数据环境下图书馆个性化服务研究[J].现代情报,2014,34(4):74-77.

作者简介

赵瑞雪，女，1968年生，研究员，博士生导师，研究方向：信息管理与信息系统、信息资源管理、知识组织及数字图书馆，E-mail：zhaoruixue@caas.cn。
鲜国建，男，1982年生，博士，副研究馆员，研究方向：知识组织、关联数据，通讯作者，E-mail：xianguojian@caas.cn。

Study on Agricultural Knowledge Discovery Service in Big Data Environment

ZHAO RuiXue, XIAN GuoJian, KOU YuanTao, GU LiangLiang, ZHU Liang
(Agricultural Information Institute of CAAS, Beijing 100081, China)

Abstract: In big data environment, research and build a new generation of knowledge discovery system to meet the new needs of users, that is the challenges and also opportunities faced by professional information institutions such as libraries. This paper firstly analyzed the latest progress of several well-known resource and knowledge discovery systems, and designed the architecture of agriculture domain-specified knowledge discovery systems, and then detailed the progress, including the integration and fusion of the multi-source heterogeneous information resources, unified search based on metadata warehouse, multi-dimensional knowledge discovery based on semantic association, data mining based on knowledge fusion and specific requirement oriented personalized service.

Keywords: Big Data; Resource Discovery; Knowledge Discovery; Resource Aggregation; Knowledge Service

（收稿日期：2016-08-29）