基于关联数据的知识聚合与发现研究进展

贯 君 毕 强 赵夷平 (吉林大学管理学院 长春 130022)

摘要 关联数据极大地推动了网络信息的关联建立,也为知识聚合提供了新的强有力的方法,由此引发从知识聚合到知识发现的拓展,关联数据的运用也迅速成为知识发现的有效途径和方法。文章对基于关联数据在知识聚合和知识发现领域的研究进行认真调研,并对关联数据知识聚合和发现的关系和前景进行分析和探讨。

关键词 关联数据 知识聚合 知识发现

Linked Data- based Knowledge Aggregation and Discovery Research Progress

Guan Jun Bi Qiang Zhao Yiping (School of Management, Jilin University, Changchun, 130022)

Abstract Linked data not only brought new vigor for the construction of information on the web, but also changed the knowledge aggregation. As knowledge discovery is expanded, the use of the linked data quickly became a new efficient and powerful mean for knowledge discovery. In this paper, the methods of knowledge discovery based on linked data were investigated, and the application of knowledge discovery in various disciplines were analyzed and discussed.

Keywords linked data, knowledge aggregation, knowledge discovery

关联数据作为新的数据交换和表示的语义网技术标准,在提供标准的数据集存取和互操作性的基础上,提高了半结构化数据和结构化数据的易用性,为知识聚合与发现提供了比传统"本体"语义技术更高效的途径和方法。关联数据网络蕴含着辅助知识聚合的更深层次、更重要的信息;知识发现重在发现知识,两者的结合将对信息发掘产生深远影响。文章选取近年来公开发表的期刊会议论文,对以关联数据为基础的知识聚合与发现相关研究进行了较为详尽的考察,以期为未来关联数据知识发现研究提供有益的建议与参考。

1 关联数据的知识聚合与发现

关联数据、知识聚合和知识发现不是同时代诞生,但却有着组织信息发现新的知识的共同目标。通过概念和关系可以从根本上认识它们,了解关联数据的知识发现。

1.1 关联数据、知识聚合与知识发现

2006 年,Berners- Lee Tim 提出:RDF 文档以统一资源定位符(URI)为名称;URI 必须符合超文本传输协议(HTTP);URI 指向的信息必须以标准格式(RDF,SPARQL)提供;发布信息必须包含 URI 4 条规则,形成了以语义网为基础结构的"关联数据"(Linked Data)概念[1]。其意义在于充分利用已有的轻量级、易组织分布数据集的框架,使用标准的知识表示方式与查询语言,通过链接扩展实现知识对象网络动态关联。就这个意义而言,关联数据实质是一种高度规范的"本体"。

知识聚合是近年来出现在国内学术研究中的新概念,通过统计分析、数据挖掘、人工智能等方法对可能存在隐性关联的知识单元进行凝聚,以提取知识单元间的内在关联为手段,构建多维多层又互相关联的知识体系。知识聚合注重信息资源语义及其关联,同时也能够通过计量方法实现。

知识发现自 1989 年第十一届国际联合人工智能学术会议兴起,最初定位为数据库的知识发现,其定义一直沿用至今,即从大量数据中识别出可信的、新颖的、潜在有用的以及最终可理解的模式的高级处理过程[2]。它包括数据预处理、数据挖掘、后处理这一系列转换步骤。

本文系国家自然科学基金项目"语义网络环境下数字图书馆资源多维度聚合与可视化展示研究"(编号:71273111)的研究成果之一。

从它们的概念上看，知识聚合依据数据资源中的脉络对已有的知识进行有机组织，而知识发现则从现有数据资源中发现特异的新知识。知识聚合是知识发现的基础，知识发现是知识聚合的最高目标。在研究方法上相互交叉，而应用方面知识聚合的结果是为知识发现提供素材，知识发现的结果也可以成为另一层次上知识聚合的因子。

1.2 关联数据与知识发现的关系与影响

知识发现的重要步骤——数据挖掘发展较为完备，它结合数据分析方法和复杂算法，能够适应新类型数据分析任务。关联开放数据云包含极大量的相互关联实例与丰富的待检知识。然而由于关联数据云的庞大和差异，人工学习耗时且描述特定类别的实例的重要属性困难。分类聚类、关联分析等很多重要的思想和方法都可以用来指导发现关联数据源之间、不同数据源的数据之间的新关联或者进一步的关联，从而提升关联数据的核心价值。此外，降低本体的异质性并检索各类中使用的核心属性也可以用知识发现的方法对关联和概念进行检验。

关联数据的出现简化了以往知识发现只能依靠复杂方法和运算实现的状况，关联数据的链接模式为知识发现提供了简洁高效的新渠道。关联数据将以描述逻辑为基础的语义网技术引入其中，丰富了知识发现中的机器学习方法，增强了传统知识发现中的对半结构化与非结构化文档的知识发现的能力，也为知识发现增强了结果的语义验证能力。最大的变化在于，关联数据的出现将知识发现从过去的以数据库为中心逐渐转变为以网络数据为中心，在数据组织形式发生巨变的前提下研究和实现关联数据知识发现理论、方法和技术，最终实现应用与推广将是未来知识发现新的发展方向。

2 关联数据知识发现研究动态

关联开放数据的知识组织能力从其产生之初便受到学术界的关注，伴随着开放关联数据项目的提出与发展，至今已积累了许多领域的开放关联数据研究成果。它们为以关联数据为基础的知识聚合与发现奠定了较坚实的基础。知识聚合与发现迎来了又一次应用语义网技术的热潮。

2.1 关联数据知识聚合与发现过程

采用关联数据进行知识聚合与发现，主要是利用关联数据在化解语义异构和本体定位问题的优势，构造知识聚合框架实现。在知识聚合领域有数据发布层、数据关联层和数据集成应用层的3层框架结构[3]。图书情报领域中从服务、组件和对象3个功能实体组成的基于SOA的关联数据的高校图书馆知识服务架构角度将其分为数据层、聚合层、组件服务层、应用层[4]。在知识发现领域分为资源层、知识发现处理层、应用层的3层的基于关联数据的知识发现模型[5]。在知识聚合与发现框架中，关联数据发布层或者说数据层是实现知识组织的基础，需要将数据组织工具关联数据化，才能够将知识节点结合成相互关联、易于拓展的整体；数据关联层或者知识发现处理层是实现由基础关联数据与其他类型资源聚合与发现的关键结构；而数据集成应用层或应用层就是包括用户接口在内的一系列用户服务的上层平台。

关联数据的知识发现过程总体分为功能概括型，即对各个环节的功能进行简要概括，如关联数据发布、相关源选择、关联数据整合、关联数据挖掘4个基本阶段，以及方法描述型，即对执行过程每一步的方法进行描述，如过SPARQL获取信息、数据预处理、转换数据格式、关联数据挖掘算法运算、结果的可视化和模式评估6大步骤[6]。从知识聚合与知识发现模式和过程的解读，展示出虽然采用关联数据进行知识聚合和知识发现工作在我国起步较晚，但是其运作模式仍然十分接近传统知识发现的数据收集、数据预处理、数据转换、数据挖掘、模式解释和评价这一一般过程。

2.2 关联数据知识聚合研究进展

基于关联数据的知识聚合可以把本地资源和外部的数据网络相互连接起来，增强和扩展其资源发现平台，更好地保存、管理和利用研究者创建的数据，有效地促进学术交流[7]。具体体现在以下两个方面。

2.2.1 图情领域基于关联数据的知识聚合

知识组织工具的关联数据化是依靠RDF文件中的大量资源链接来有效扩展知识点连接的范围，这些链接不仅决定了数据的语义，也通过“属性”链接到大量存在关联的资源实体。图书馆行业在这方面进行了积极的探索，取得的研究成果包括分类法、主题词表和元数据描述框架的关联数据化、工作流程关联数据化和政府开放数据。

目前，杜威十进制分类法已经部分实现关联数据化，通过对分类法中的数据集、数据、版本、格式、语言等以实体形式表示，并对它们之间的关联进行描述，从而使分类法的功能由基本的查询检索向深层次的分面信息挖掘和跨领域链接发展，增进了其在多领域中的应用[8]。由于现行文献分类法都是由专家和学者人工制定的，在类目划分、子类设置和类别编码等方面存在随机性，不能通过聚类等方法模拟，使得分类法的关联数据难以实现自动生成。

成果最丰富的领域是元数据向关联数据转化方面。其中以文献资料为基础的有：通过对MARC书目数据进行RDF格式转换并URI资源命名实现馆藏元数据的关联数据化，进而借助指向外部数据源的RDF链接构建与面向关联数据网络的开放发布，实现书目数据的语义转换与网络关联[9]。采用条件随机域

(Conditional Random Fields, CRF) 和支持向量机的机器学习方法，从科学文献中自动抽取前端元数据，实现标题、作者、邮件、出版物、出版地等类别到关联数据网络的映射[10]。

关联数据在不同来源的数据之间建立链接，构成数据网络的特性使各类馆藏资源得以聚合。利用关联数据 URI 复用与 RDF 链接的关联数据聚合机制，通过 URI 命名、关联数据词汇集创建、馆藏语义描述与关联数据发布实现图书馆关联数据集的创建与发布，在图书馆关联数据集与其他数据集之间构建 RDF 链接并进行动态维护，以此为桥梁将实体馆藏与虚拟馆藏聚合成为整体资源空间，实现包括馆藏资源关联数据化与图书馆关联数据链接管理两部分的基于关联数据的馆藏资源聚合模式[11]。利用关联数据搜索引擎 Falcon 对数据集内的各类属性数据进行基于 RDF 三元组的共现检索，实现数据集内各类数据关联发现的目的，进而解决数据集内部关联数据的自动创建问题[13]。

文献传递馆员个人所有的、隐性的，关于业务的流程知识外化成组织的、显性的流程知识，通过实现知识整合帮助部门建立自己的知识库。从流程实现文献传递工作的知识整合和知识发现，使得各种业务流程更加清晰、形象、易于理解，便于共享和交流[13]。任何业务都是一个动态过程，每个过程由若干相互关联的活动组成。这些活动、活动之间的关系和所涉及的相关知识形成业务的流程知识。Li Ding 等学者[14]用支持复杂领域概念建模的 OWL 格式建立描述工作流程的关联起源数据，重用基于开放起源模型的工作流追溯。

2.2.2 知识聚合与获取

知识是由众多结点(知识因子) 和结点间联系(知识关联) 两个要素组成的[13]，从属于不同类别的知识节点间关联的异质性在通过关联数据进行知识组织过程中是一大挑战。通过从互联的实例中检索不同数据集中连接相同实例的类和属性，利用等同(SameAs) 关系实现基于图的本体整合，进而使用机器学习方法找出常用实例对可能遗漏的核心本体类别和属性加以补充，最后将补充的类别和属性相结合形成整合本体[16]是一种有的效应对方法。

在关联数据应用内容中生成可执行映射，即源本体、目标本体和二者间的每个对应产生一个可执行的 SPARQL 映射，对这些映射作待整体转换的源本体数据集的结构和将整体生成的目标本体数据集结构描述，就实现了建立在一致性上的数据交换，从而用来自一个或多个源应用的数据构成目标应用的数据模型。但是这种方法对数据集规格要求较高，一旦发生变化则转化可能出现错误[17]。对关联数据中包含的大量超级链接进行深度利用是知识聚合的又一途径，从关联知识源的链接关系中提取资源元图 (resource meta- graph) 和分类元图(category meta- graph) ，再经过知识源与主题的概念相似度测量实现微博客中的短信息按讨论主题自动分类[18]。

此外，还有采用概率理论与图理论结合对从自然语言文本中提取的实体进行潜在实例匹配和实体链接决策，在此基础上用众包来选出实体和对应的 URI 并再次匹配推荐实例，提升数据整合质量。目的是结合大规模实例自动匹配的高效率和人工标引的准确性，从而实现高质量的关联数据整合[19]。按照关联数据原则对多个第三方数据源进行映射，提取教育工具描述数据并将其匹配到教育工作者可理解的教育词汇表，从而建立教育工具数据集，实现对不同教学设置的教育工具发现、数据集自动维护和供教育领域信息交流工具使用等功能[20]。从庞大的微博客数据集中提取命名实体，通过发现可能参考新实体的推文生成典型微博样本。用排除转发的、无预定义实体的、类似的和冗余的博文同时保留信息内容的方法提升语义分析速度，结果是由细粒度的实体类型结合成的核心实体类型，微博元数据关联数据化为更准确进行主题分析和趋势检测研究提供了有价值的素材[21]。

随着 2009 年政府开放数据运动兴起，外国政府的政府信息关联数据化迅速展开，为此类研究奠定了坚实的开放数据基础。ERMIS 希腊公共管理门户网站[22]就是将信息向关联数据转换，并以此从其他欧洲国家网站找到等价信息的希腊开放政府数据应用。我国则从理论上对电子政务信息资源语义关联组织提出了层次模型[23]，包括电子政务信息资源层、资源描述层、数据发布层、数据关联层和数据利用层。组织电子政务信息资源目录数据以关联数据的形式发布，实现网页浏览、关联数据浏览和 SPARQL 检索等服务。从世界范围来看，目前政府开放关联数据已经有了一定的积累，但是透过这些数据进行知识发现研究尚处于起步阶段。

知识聚合成果最直接的获取方式是查询和搜索。数据查询是从众多数据资源中选出目标数据的过程。关联数据对数据资源选择方法有 3 种，都是围绕统一资源定位符查询进行的：实时探索法是直接对 URI 进行递归查找实现结果的增量发现；索引法是用已索引的数据进行查询的方法，URI 在其中起到指针的作用；混合法结合实时探索法和索引法，根据索引检出 URI 作为查询种子的基础上进行实时探索更新检索结果。关联数据查询处理同样有 3 类：将数据集中到中心化数据仓库进行统一检索的数据仓库法；将查询分别发送到关联数据集端点执行的联合查询处理方法；以及链接遍历查询处理方法[24]。将 Bio2RDF 关联数据的类型和关联属性向语义科学集成本体(Semantic Science Integrated Ontology) 映射，采用联合查询处理方法实现 SPARQL 查询[25]、Peter Ansell[26]用在一系列终端执行查询后再将结果转化成具有标准化

URI 的 RDF 描述形式的分布式查询方法，形成跨越不同位置数据集的查询模型，以此为基础实现包含命名空间、查询类型和规则的网络应用原型，吸引了大量用户使用。

探索式搜索（Exploratory Search）是指通过用户自己对搜索内容的认识按需引导搜索空间的搜索方式。基于关联数据的探索式检索是探索式语义搜索（Exploratory Semantic Search）的一种，它以本体知识库和建立在其上的关联属性分级诱导方法为核心，用它测定搜索对象的重要程度。yovisto.com[27]就是一款集可选地理信息、分面过滤和探索型搜索导航为一体的视频搜索引擎，利用 DBpedia 关联数据进行查询到实体的映射，采取启发式实体分级发现重要关联实体，将关联实体和索引检索结果一并提供给用户。将描述资源的 RDF 三元组转换为数字形式组织数据摘要索引，使关联数据源与查询的关联程度的抽象模式匹配变为三维空间坐标的显性方法提取描述最相符的数据[28]，进而通过属性重叠量和结合次序判断提升系统效率和准确程度。

在这方面研究有 3 大主要问题，建立能够利用不同元数据模式的基于本体的中介体系结构显得十分重要，通过它与数据聚合器配合工作将更利于检索最新的三元组；对多个数据仓库执行联合检索设计通用数据接口以提高数据提取效率和质量；目前对如何实现和如何评价探索性搜索尚没有普遍接受的最佳检验方法。

2.3 关联数据知识发现研究进展

从知识发现基本任务包括数据描述与辨别、数据分类聚类、关联和相关性、衰退与预报、时序分析以及顺序发现[29]。目前关联数据知识发现研究则集中于数据描述辨别、数据分类聚类、关联性方面，能够完成大部分知识发现任务。

2.3.1 联系的发现与展示

在基于关联数据的知识发现中，关联与相关性集中体现在联系发现中。联系发现也称连接发现（Link Discovery），是从大量数据中挖掘潜在的模式，发现其中隐藏的联系并抽取其中有用的知识的技术[30]。沈志宏等[31]分析了面向关联开放数据的关联发现过程处于数据互联阶段，以构建多类资源关联数据网为目标，具有多任务、多路径、多步骤等复杂特性。司徒俊峰等[32]学者将基于关联数据的知识链接划分为等同链接、相关性链接、词汇性链接三种类型。采用关联数据技术构建知识链接的方法同样有 3 种：知识对象的 URI 标识，创建 RDF 链接，以及知识链接的发布与维护。从应用角度建立包括描述层、链接层、整合层、应用层的 4 层关联数据知识链接应用框架。链接发现方法可分为 3 类[33]：从多维空间的一点或多点对各个实例进行映射来获取链接的多维方法、从相似度空间的单个维度生成必要的约束，通过约束抽取备选链接的单维方法和用降低运行时间的方法生成备选数据，将备选数据合并生成最终的链接列表的混合方法，实质是一种支持向量机方法。用实例间存在的“一致性”（SameAs）关系推理辅以“不同于”（DifferentFrom）关系判断自动生成高准确度训练集[34]，在对比训练集中的实例内容过程中发现特征属性对，通过测量特征类对和消费层领域对的差异，聚集高差异性属性对作为连接同类同领域中的实例的依据。最终为分属不同属性的实例建立链接。

可视化是直观展示数据及其之间联系的方法，能够为非专业人员提供全部组成成分概览并测定其关联情况[35]。交互界面、易用性和对不同数据的适应性是可视化成功实现的基础，借助关联开放数据发掘科研数据中的潜在内容，为科研工作发现新的机会和方法是可视化的终极目标。在研究信息关联开放数据（RILOD）基础架构上用关联开放数据可视化套件（LOD/ VizSuite）实现了点对点的可视化解决方案，经过查询转换将研究网络、实践社区、研究员关系、时间线等检索结果形象化地展示在用户面前[36]。洪娜等[37]学者借助 RelFinder 用户主导的可视化过程，制作了基于生物医学关联数据的关联发现系统，有效发现跨数据集的复杂关系，为生物医学潜在知识关联的发现做初步探索。Maulik R Kamdar 等[38]展示了用于癌症研究的关联数据实时可视化探测器与聚合器（ReVeaLD），该平台以关联生物医药数据资源（LBDS）为基础，可查询 in silico 实验数据，蛋白质建模和基因示性。使用经生物医药专家认定并与外部资源要素目录进行映射的领域专业化语言为基础的公式化查询，并提供以用户为中心的可视化分析平台实现分布资源的直观交互。

上述研究探索了关联发现方法并进行了初步实验，形成了以结构化信息联合聚合方法、词汇匹配策略、关联一致性匹配方法在网络本体和 LOD 数据集中寻找等价实体[39]的工作流程。今后利用相似度匹配方法、集成的本体结合迭代反馈方法发现失踪链接关联数据集，能够应对不同的数据类型的关联发现算法，大规模数据环境下关联发现过程的时效性是这一领域的重要发展方向。现阶段可视化主要依靠开放的工具实现，受工具性能制约，对链接与关联发现效率还比较低。未来在高效算法的研究与改进方面具备很大潜力。采用基于关联数据的知识组织、知识表示、知识聚合和可视化方法组织和展示数字资源，能够形成具备关联的知识集。实现更具专业性的深层次知识服务，更好地满足创新主体对知识信息的深层次需求，从而增进知识理解、促进知识扩散、推进知识创新。

2.3.2 领域新知识的发现

自然科学领域方面，采用地表温度数据集和气温

关联数据集，用并行处理方法在依据假设证明需求获取增量提取气温数据的同时将数据转化为关联数据的三元组形式，每次增加数据后采取计量经济学的时间序列分析方法实现整合的新数据集上对气候变化假设的重复证明[40]。Sahana 亚洲系统[41]是支持早期灾难预警的灾难数据管理系统，采取众包开发方法聚合数据流和功能，满足用户对数据多变的需求并处理关联数据的各个分支，将不同的灾难描述数据源转换为 RDF 三元组，并与 LOD 云中的地理领域关联数据连接，建立灾难模式知识库以识别灾害的形势。在此基础上建立决策规则，在提前识别灾难发生形势时，预测受到波及的地域与设施、向用户发布不同级别警告等信息，对灾害管理过程提供重要参考。

在商务与金融领域，公开采购过程的合约归档应用[42]也可以利用关联数据实现。透过对公开采购门户网站中缔约双方发布的信息进行本体建模和公开的合约本体的重用，平行再现买方视角和投标人视角的采购工作流程；运用以关联数据为基础的自动合同匹配过程向买卖双方提供最适宜的招投标建议，提供中介服务，并支持双方灵活决策。可扩展商务报告语言（XBRL）结合关联开放数据联合建立全局金融生态系统是本领域中又一成功范例[43]，该应用以关联数据直接支持数据整合的特性整合金融数据，将 XBRL 的分类和实例转换为 RDF 格式连结分布式开放数据集，形成图结构化的金融商务信息数据生态系统。

生物医药是关联数据知识发现应用最多的领域，将应用程序接口（API）应用于药理学关联数据集成系统中[44]，围绕 API 的定义实现方法与基础数据的访问，采取 API 中介模块的方式增强数据准备的灵活性并实现 API 要求向 SPARQL 查询的转换，为开发者使用关联数据应用架构带来更多选择性。无独有偶，开放自我药物治疗[45]也是一款以 API 连结关联数据源和知识库的药物推介网络应用，采取了独特的 PHP 结合开源 RDF 三元组数据库 Sesame 提高通过超文本传输协议存取关联数据集的可靠性，以药物临床表现、成分等非关联数据信息作为药剂学坐标，针对症状为用户推荐适宜的药品。

通过以关联数据为基础的知识聚合研究和基于关联数据的知识发现进行对比，可以发现二者在实现框架、程序安排、方法运用等方面存在众多近似之处。知识发现可以作为基于关联数据知识聚合的一步，用于生成新的链接或关联数据；知识聚合也可以作为以关联数据为基础的知识发现的一步，用来聚集基础信息，从而实现由原有的信息发掘新知识的跨越。从这个意义上来说，以关联数据为基础的知识聚合与知识发现呈现出互相包容、相互促进的螺旋形上升的发展形态。

3 关联数据知识聚合与发现的发展趋势

从中外学者们在关联数据知识聚合与发现的研究成果了解现阶段的问题和未来工作的计划，能够使我们了解我国关联数据知识发现研究面临的挑战，把握未来研究方向。

3.1 关联数据知识聚合与发现的挑战

将关联数据应用于知识发现领域首先面临关联数据的制备问题。虽然关联数据标准简洁宽松，但关联数据内容可读性差、编辑繁琐。为了克服这种问题产生了“精益的语义网”等使用可读性强的编程语言按需实现语义关联的多种分支。在发挥关联数据连接作用的过程中，需要对本地资源针对关联开放数据源进行关联数据化，其间一直存在标准术语派和自由表达派的观念之争。而源自网络的开放关联数据集又普遍存在质量较低、维护成本高等弊端。多方面不利因素制约了关联数据在知识聚合与发现方面的应用与发展。

由于不同语言的语义差异，信息组织方法也各不相同，单纯依靠翻译和移植并不能完全解决问题。我国在语义网技术研究过程中本体发展迟缓的情况延续到关联数据，开放关联数据与知识库匮乏，直接导致我国关联数据知识发现存在多理论研究、少实证研究的现象：即使在以关联数据为基础的知识发现成果较多的图书馆领域也未能推出开放关联数据集，实证研究又多采用英文关联数据与知识库的情况可见一斑。正如国外学者[46]在研究朝鲜语文档进行语义标注时探讨国际化和本地化困境所言：“一方面是语言处理技术对外国语不支持或仅部分支持；另一方面是容量、结构、实体链接、实体语义丰富度和词项化数据的有效性等知识库质量问题。”

在科研领域，关联数据知识发现带来数据融合能力的收益，同时也面临可信度的挑战。由于关联数据核心作用在于向不同的服务系统和知识源提供标准化访问接口和 RDF 数据模型，对概念层次和知识表示存在缺陷，直接导致发布的关联数据不能反映研究方法，也不能体现研究人员的权利和声望。而科研工作又是以有可信来源和完备解释的研究结果发表为基础，通过方法描述支持其再现性。国外已有一些学者[47]提出，建立在关联数据上增加整合研究背景信息等潜在属性的“研究目标”层结构体系，依照严格的原则操守保存数据。但这方面工作仍然任重道远。

在现有资源的关联数据建立发布过程中，由服务方式、分类体系、数据内容、领域本体及专业词表的复杂性造成的同名异义、异名同义、同体异构等情况，数据分面复杂、类型多样，再加上数据资源的大数据量，给实体识别与自动互联算法的准确性、高效性提出了较高的要求。

3.2 关联数据知识聚合与发现的前景

从国内研究机构来看，图书馆在关联数据知识聚合与发现研究方面首当其冲。对大量馆藏信息深层次开发利用是图书馆引入信息资源聚合、语义服务等技术的内在动力。拥有较强直接整合数据能力和平台无关性的关联数据是图书馆扩展资源发现平台、推进知识服务的一个有效方案。未来一段时期图书馆将凭借先发优势和资源储量继续引领我国的关联数据知识发现研究。其他学科对关联数据知识发现应用较少，受制于关联数据资源的情况明显，但随着开放共享氛围逐步形成，这些领域的关联数据知识发现研究增长潜力十分巨大。

从国外研究情况来看，已经出现了一些处于起步试验阶段的基于关联数据的知识聚合与发现应用案例，医药和生命科学正是关联数据知识发现飞速发展的强势学科。其关联数据储备丰厚、研发前景广阔，但是也存在应用层次及应用框架各异，致使知识链接数据难以共享重用的问题。故采用统一的应用框架将是今后推广和应用基于关联数据的知识链接的关键所在。政府机构纷纷加入关联开放数据运动使各类经济、民生等社会信息关联数据化，极大提高了数据易用性，但是建立在这些数据上的知识发现研究还较为鲜见。今后对政府发布的关联数据从行业等领域进行知识发现也将成为分析国家、地区社会发展的重要途径。

关联数据进行知识发现具备先天的结构和操作优势。它能够仅以 URI 直接关联解决很多异构数据库需要大量时间精力整理数据和程序才能实现的知识整合工作。根据不同用户在进行知识发现活动时的偏好和习惯推荐最适合的机器学习算法，充分利用关联数据获得有价值的知识，是提升此类系统易用性应考虑的问题。融合统计分析和人工智能等先进技术，关联数据在从多种来源的知识发现新的知识方面必然有更佳的表现。

参考文献

[1] Berners-Lee T. Linked Data-design Issues[EB/OL].[2009-06-18]. http://www.w3.org/DesignIssues/LinkedData.html.

[2] Fayyad U, Shapiro G P, Smyth P. From data mining to knowledge discovery in databases[J]. AI Magazine, 1996, 17(3): 37-54.

[3] 马费成，赵红斌，万燕玲，等. 基于关联数据的网络信息资源集成[J]. 情报杂志，2011(2): 167-170, 175.

[4] 赵卫军. 基于SOA的关联数据的高校图书馆知识服务架构[J]. 图书馆学刊，2013(6): 103-105.

[5] 李 楠，张学福. 基于关联数据的知识发现模型研究[J]. 图书馆学研究，2013(1): 73-77, 67.

[6] 李 俊，黄春毅. 关联数据的知识发现研究[J]. 情报科学，2013，31(3): 76-81.

[7] 丁 楠，潘有能. 基于关联数据的图书馆信息聚合研究[J]. 图书与情报，2011(6): 50-53.

[8] 贾君枝，赵 洁. DDC关联数据实现研究[J]. 中国图书馆学报，2014，40(4): 76-82.

[9] 游 毅. 面向馆藏聚合的书目关联数据实现[J]. 情报理论与实践，2014，37(8): 105-110.

[10] Tudor Groza, Gunnar Aastrand Grimnes, Siegfried Handschuh, et al. From raw publications to linked data[J]. Knowledge Information System, 2013(34): 1-21.

[11] 游 毅，成 全. 试论基于关联数据的馆藏资源聚合模式[J]. 情报理论与实践，2013，36(1): 109-114.

[12] 王忠义，夏立新，郑 路，等. 数据集内关联数据自动创建方法研究[J]. 情报杂志，2014，33(1): 152-156.

[13] 丁 楠，潘有能. 基于流程的文献传递知识整合与知识发现——以浙江大学图书馆为例[J]. 图书馆学研究，2011(3): 67-72.

[14] Li Ding, James Michaelis, Jim McCusker, et al. Linked provenance data: A semantic Web-based approach to interoperable workflow traces[J]. Future Generation Computer Systems, 2011(27): 797-805.

[15] 王知津. 文献演化及其级别划分——从知识组织角度进行探讨[J]. 图书情报工作，1998(1): 4-7.

[16] Lihua Zhao, Ryutaro Ichise. Ontology integration for linked data[J]. Journal on Data Semantics, 2014(5): 1-18.

[17] Carlos R Rivero, Inma Hernández, David Ruiz, et al. Exchanging data amongst linked data applications[J]. Knowledge Information System, 2013(37): 693-729.

[18] Andrea Varga, Amparo Elizabeth Cano Basave, Matthew Rowe. Linked knowledge sources for topic classification of microposts: A semantic graph-based approach[C]. Web Semantics: Science, Services and Agents on the World Wide Web, 2014: 36-57.

[19] Gianluca Demartini, Djellel Eddine Difallah, Philippe Cudré-Mauroux. Large-scale linked data integration using probabilistic reasoning and crowdsourcing[J]. The VLDB Journal, 2013(22): 665-687.

[20] Adolfo Ruiz-Calleja, Guillermo Vega-Gorgojo, Juan I Asensio-Pérez, et al. A linked data approach for the discovery of educational ICT tools in the Web of data[J]. Computers & Education, 2012(59): 952-962.

[21] Oluwaseyi Feyisetan, Elena Simperl, Ramine Tinat, et al. Quick-and-clean extraction of linked data entities frommicroblogs[C]. SEM14, Leipzig, AA, Germany, 2014.

[22] Eleni Galiotoua, Pavlina Fragkou. Applying linked data technologies to greek open government data: A case study[C]. Procedia - Social and Behavioral Sciences, 2013: 479-486.

[23] 吕元智. 基于关联数据的电子政务信息资源语义组织研究[J]. 图书情报工作，2012，56(21): 143-146, 130.

[24] Olaf Hartig. An overview on execution strategies for linked data queries[J]. Datenbank Spektrum, 2013(13): 89-99.

[25] Alison Callahan, José Cruz-Toledo, Michel Dumontier. Ontology-based querying with Bio2RDF's linked open data[J]. Journal of Biomedical Semantics, 2013, 4(Suppl 1): S1.

[26] Peter Ansell. Model and prototype for querying multiple linked scientific datasets[J]. Future Generation Computer Systems, 2011(27): 329-333.

[27] Jorg Waitelonis, Harald Sack. Towards exploratory video search using linked data[J]. Multimed Tools Appl, 2012(59): 645-672.

[28] Jürgen Umbrich, Katja Hose, Marcel Karnstedt, et al. Comparing data summaries for processing live queries over linked data[C]. World Wide Web, 2011: 495- 544.
[29] 孙吉红,焦玉英. 知识发现及其发展趋势研究[J]. 情报理论与实践,2006,29(5): 528- 530, 527.
[30] 陈 飞,商 琳,骆 斌,等. 联系发现:一种新的数据挖掘方法综述[J]. 计算机科学,2006,33(11): 123- 127, 131.
[31] 沈志宏,黎建辉,张晓林. 面向 LOD 的关联发现过程的定位、目标与复杂性分析[J]. 中国图书馆学报,2013,39(6): 101- 108.
[32] 司徒俊峰,曹树金,谢 莉. 论基于关联数据的知识链接构建与应用[J]. 图书情报工作,2013,57(16): 123- 129.
[33] Axel- Cyrille Ngonga Ngomo. On link discovery using a hybrid approach[J]. Journal on Data Semantics, 2012(1): 203- 217.
[34] Wei Hua, Rui Yang, Yuzhong Qu. Automatically generating data linkages using class- based discriminative properties[J]. Data & Knowledge Engineering, 2014(91): 34- 51.
[35] Joachim Baumeister, Martina Freiberg. Knowledge visualization for evaluation tasks[J]. Knowledge Information System, 2011(29): 349- 378.
[36] Anastasia Dimou, Laurens De Vocht, Geert Van Grootel, et al. Visualizing the information of a linked open data enabled research information system[J]. Procedia Computer Science, 2014(33): 245- 252.
[37] 洪 娜,钱 庆,范 炜,等. 关联数据中关系发现的可视化实践[J]. 现代图书情报技术,2013,230(2): 11- 17.
[38] Maulik R Kamdara, Dimitris Zeginis, Ali Hasnaina, et al. ReVeaLD: A user- driven domain- specific interactive search platform for biomedical research[J]. Journal of Biomedical Informatics, 2014(47): 112- 130.
[39] Wang Zhichun, Li Juanzi, Zhao Yue. A unified approach to matching semantic data on the Web[J]. Knowledge- Based Systems, 2013(39): 173- 184.
[40] Jaakko Lappalainen, Miguel- ángel Sicilia, Bernabé Hernández. Automatic hypothesis checking using eScience research infrastructures, ontologies, and linked data: A case study in climate change research[C]. Procedia Computer Science, 2013: 1172- 1178.
[41] Thushari Silva, Vilas Wuwongse, Hitesh Nidhi Sharma. Disaster mitigation and preparedness using linked open data[J]. Ambient Intell Human Comput, 2013(4): 591- 602.
[42] Martin Necasky, Jakub Klimek, Jindrich Mynarzb, et al. Linked data support for filing public contracts[J]. Computers in Industry, 2014(65): 862- 877.
[43] Seán Ó Riain, Edward Curry, Andreas Harth. XBRL and open data for global financial ecosystems: A linked data approach[J]. International Journal of Accounting Information Systems, 2012(13): 141- 162.
[44] Paul Grotha, Antonis Loizoua, Alasdair J G Grayd, et al. API- centric linked data integration: The open PHACTS discovery platform case study[C]. Web Semantics: Science, Services and Agents on the World Wide Web, 2014: 1- 7.
[45] Olivier Curé. On the design of a self- medication web application built on linked open data[C]. Web Semantics: Science, Services and Agents on the World Wide Web, 2014: 27- 32.
[46] David Müller, Mun Yong Yi. Annotating Korean text documents with linked data resources[J]. Multimed Tools Appl, 2014(68): 413- 427.
[47] Sean Bechhofer, Iain Buchanb, David De Roure, et al. Why linked data is not enough for scientists[J]. Future Generation Computer Systems, 2013(29): 599- 611.

[作者简介] 贯 君,男,1982 年生,吉林大学管理学院博士研究生。
毕 强,男,1954 年生,吉林大学管理学院教授,博士生导师。
赵夷平,男,1981 年生,吉林大学管理学院博士研究生。

收稿日期:2015- 02- 17