DOI: CNKI: 11-2746/G2. 20130730. 1538. 001

面向 LOD 的关联发现过程的定位、目标与复杂性分析*

沈志宏　黎建辉　张晓林

摘　要　本文以关联数据应用过程中的关联发现过程为研究对象，分析了面向关联开放数据(LOD)的关联发现过程的定位、目标与复杂性。本文认为，关联发现过程处于关联数据应用过程三阶段(数据发布、数据互联与数据消费)中的第二阶段。关联发现过程的整体目标是构建多类资源之间的关联数据网络，该过程的本质就是关联数据网络不断演变的过程。关联发现的过程具有多任务、多路径、多步骤等复杂性特征。目前流行的关联发现框架还存在缺乏对整个网络演变过程的支持、任务类型单一、缺乏流水线机制等不足。因此，关联发现技术的研究急需新的面向整个关联数据网络的、支持完整演变过程的、支持多任务集成的理论、方法与框架。图5。表1。参考文献17。

关键词　关联数据　关联发现　数据互联

分类号　G350　TP393

Insights into Link Discovery Process for Linked Open Data: Positioning, Goals and Complexity

Shen Zhihong, Li Jianhui & Zhang Xiaolin

ABSTRACT　This paper studies link discovery process for Linking Open Data (LOD), and analyzes the positioning, goals and complexity of the process. The study finds that link discovery process is located in the second phase of the three stages of linked data application process, namely linked data publishing, interlinking and consuming. The overall goal of link discovery process is to build a linked data network for different LOD resources, and the process is essentially a continuous evolving process of the entire network. It is also found that the link discovery process is a very complex process with multiple tasks, approaches and steps. This paper argues that current frameworks of link discovery are inefficient due to insufficient support for network evolution, the lack of variation in task types, and streamline mechanisms. As a result, there is an urgent need for new theories, methods and frameworks for link discovery, which should be designed for the entire linked data network that can support the full evolution process, and performance of multi-tasks and integration. 5 figs. 1 tab. 17 refs.

KEY WORDS　Linked data.　Link discovery.　Data interlinking.

1　背景

关联数据自2006年被提出以来[1]，随着关联开放数据(Linking Open Data，LOD)运动的开展，在媒体、地学、政府、出版、生命科学等领域得到了广泛的应用[2]。然而与关联开放数据资源飞速发展极不协调的是，跨数据集的数据关联仅占有所有

*　本文系中国科学院“十二五”信息化专项“科技数据资源整合与共享工程”课题“科学数据管理与共享云服务平台”(XXH12504-01-02)、国家自然科学基金重点项目资助“面向非常规突发事件应急管理的云服务体系和关键技术”(91224006)的研究成果之一。

通讯作者：沈志宏，Email：bluejoe@cnic.cn

资源量的1.6%[5]。

在以上背景下,关联数据的互联(interlinking),即建立跨数据集的数据关联,其方法和技术的研究成为近年来的热点话题。在LDOW2010(WWW 2010 workshop on Linked Data on the Web)中,数据互联就是会议的一大专题(其他话题包括关联数据发布、基础设施与架构、关联数据应用等)。另外,由COLD2010(International Workshop on Consuming Linked Data)发布的几大开放问题[7],居于首位的就是关联数据的互联算法(interlinking algorithm)。

LOD资源的互联研究,其进展主要体现在三个方面:应用、互联算法与方法,以及互联框架。应用方面的典型如瑞典联合目录(LIBRIS,http://libris.kb.se)与国会图书馆主题词表数据的关联、LinkedMDB[5]等。典型的互联算法如基于实体的文本映射、基于RDF图形相似度计算的映射等。互联框架,有的时候又称作关联发现框架(link discovery framework),在近几年内得到了较好的发展,典型的代表如SILK[6-7]、LinQuer(Linkage Query Writer)[8]、LIMES[9]、RDF-AI[10]、LDIF[11]等。

然而,尽管在应用、互联算法、互联框架这三个方面目前都有一些好的实例不断出现,但关于关联发现过程的定位、目标与复杂性特征,一直缺乏深入的分析和统一的结论,由于认识和理解的差异性,各种应用、互联算法、互联框架所偏重的角度也有所差异。在这种背景下,本文试图对关联发现过程的定位、实质性目标与复杂性特征进行较为系统的分析,以期为该研究方向下一步的工作提供一些实际的建议。

2 关联发现的定位

关联数据自其概念提出以来,得到了广泛的应用。本文认为,从软件层次上来讲,关联数据的应用过程可以分成三个阶段:

(1) Publishing linked data:关联数据的发布

关联数据的发布是指将本地的数据发布成符合关联数据"四大基本原则"的Web数据集,即LOD数据集。根据"四大基本原则",关联数据的核心思想是建立Web数据结构化的、关联的RDF表示。因此,该阶段会产生由多个分布的LOD数据集构成的数据集群,但它们彼此之间是独立自治的、弱关联的,缺乏互操作接口,因此尚不具备完善的支持知识发现的能力。本文将该阶段产生的关联数据网络称作关联数据网络I代,它基本上是由多个孤立的子网组成。该阶段的输入为本地的(local)、非LOD化的数据(如:关系型数据库,XML文件等),输出为LOD数据集,因此该阶段关注的是RDF发布服务器、non-RDF到RDF的映射规则等技术,如D2R[12]、Triplify[13]、Pubby[14]等。

(2) Interlinking linked data:关联数据的互联

关联数据的互联是指在关联数据网络I代的基础上,通过人工整理或者自动计算的方法,基于分布的LOD数据集,生成新的数据链接(该过程即关联发现,link discovery),本文将该阶段产生的关联数据网络称作关联数据网络II代。该阶段的输入是阶段一生成的关联数据网络I代(弱关联的数据集群),输出是关联数据网络II代。由于该阶段会涉及多个数据集内容的一致性问题,因此该阶段关注的是RDF词表映射、记录去重、实体鉴别,以及对分布式数据集的高效访问和缓存机制等问题。

(3) Consuming linked data:关联数据的消费

关联数据网络II代已具备完善的支持知识发现的能力,因此可以基于该网络建立新的应用程序。如关联搜索与浏览、关联统计、关联路径发现与可视化,以及面向特定需求的e-Science应用程序。关于"关联数据应用程序"(linked data application)这一概念,DERI(Digital Enterprise Research Institute)关联数据研究中心(Linked Data Research Centre,LiDRC)在"关联数据应用程序——关联数据使用的起源与挑战"技术报告[15]中给出了两种不同的解释,第一种含义是指关联数据结合不同领域(包括生物学、统计学、软件工程、多媒体等)中的应用,第二种含义则是指基于关联数据之上构建的Web应用程序,这类Web应用被称为"由关联数据驱动的Web应用程序"(linked data-driven Web applications)。

图1展示了如上三个阶段之间的递进过程。

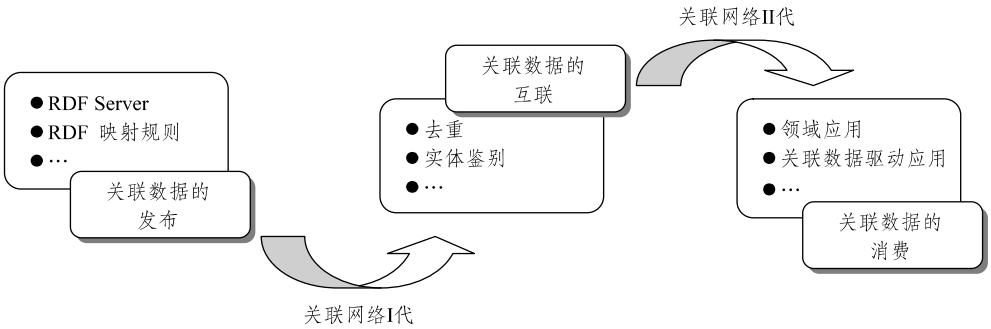


图 1　关联数据应用的三阶段

可以看出，第一、第二阶段都会创建 Web 资源之间的关联，但这两个阶段存在着本质的区别：关联数据的发布阶段通常创建的是本地关联，大部分是通过指定的模式（pattern）直接映射生成。而关联数据的互联阶段，其输入是阶段一生成的关联数据网络 I 代（弱关联的数据集群），输出是关联数据网络 II 代。关联发现过程处于关联数据的互联阶段，即关联数据应用过程的第二阶段，因此关联发现的前提必须是：Web 上已具备多个 LOD 化的数据集，它们都已具有各自的数据模型和关联词表，并且开放了标准化的数据访问接口（HTTP GET 命令或者 SPARQL 协议）。

关联发现有在线计算和离线计算之分。在离线方式下，通过发现得到的数据链接，会保存在 RDF 库中，以供后续的消费程序使用。而在在线方式下，仅仅在获取数据链接的时候才会触发关联发现的操作。在线计算的优势是数据关联的及时性有保障，劣势是对分布式 LOD 数据集的在线访问以及在线关联计算的性能要求非常高，会损失一些准确性。离线计算的优势是可以通过完整的计算步骤甚至人工甄别等工作来提高数据关联的质量，劣势则是需要采取额外的机制对数据集之间的关联进行维护，即在源内容对象和目标内容对象发生变化时，通过某种机制保持关联信息的及时更新[10]。

3　关联发现的目标与本质

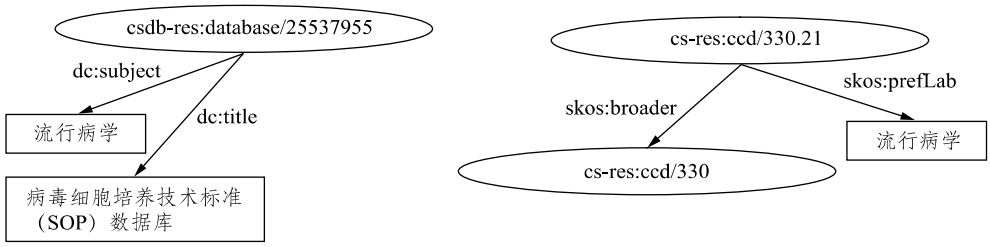
关联发现的局部目标是建立起两个或多个资源对象之间的直接关联。从技术角度来看，关联发现的目标即为已知 RDF 资源添加一条指向其他 RDF 资源的 RDF 链接（RDF link）。

图 2 以“科学数据库元数据——分类条目”的关联为例，展示建立 RDF 链接的过程。在关联发现之前，尽管数据库记录 csdb-res: database/25537955 中包含一个值为“流行病学”文本的 dc: subject 属性，但此时数据库记录 csdb-res: database/25537955 与分类条目 cs-res: ccd/330.21 之间互相孤立（disconnected），没有任何的关联关系。通过关联发现的过程，创建了一条由数据库记录 csdb-res: database/25537955 指向分类条目 cs-res: ccd/330.21 的 RDF 链接 dc: subject（虚线所示），此时，csdb-res: database/25537955 与分类条目 cs-res: ccd/330.21 之间就关联（connected）上了。

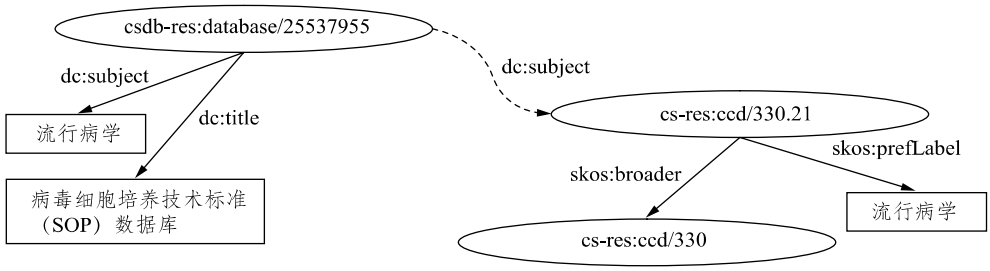
以上关联发现的过程，实质上产生了一条 RDF 三元组（主语、谓语、宾语），使用 Turtle/N3 表达为：csdb-res: database/25537955 dc: subject cs-res: ccd/330.21。

同时也可以发现，RDF 的描述机制为关联发现提供了技术上的便利。一旦发现了两个资源之间的关联，就可以采用一个三元组补充到 RDF 库中，这是一个“添加”而非“修改”的操作，这一增量式的特性在关联发现的迭代式过程中无疑会体现出巨大的技术优势。

广义上讲，两个资源之间的关联，应该不局限于点对点的直接关联，还应该包括借助于中间结点的间接关联。如图 3 所示，数据资源 D 与文献资源



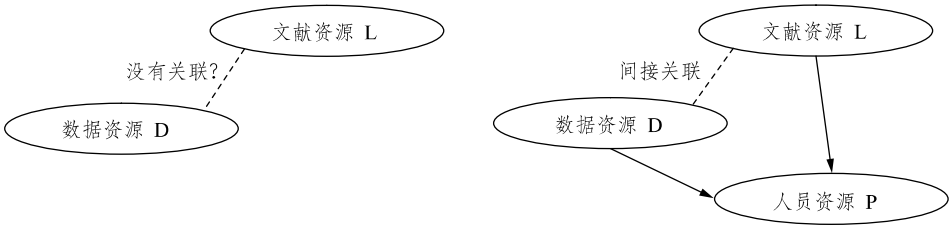
（a）关联发现之前



（b）关联发现之后

图2 关联发现过程示意图

L 没有直接关联，但它们通过共同关联的人员资源 P 的引入产生了间接关联（虚链接），间接关联又可称作桥式连接，其中的中间资源 P 常常被称作桥（bridge）。在知识发现的过程中，间接关联往往比直接关联更有意义。



（a）没有直接关联　　（b）存在间接关联

图3 直接关联与间接关联示意图

图4 展示了一个更为复杂的例子，随着越来越多的、不同类型的中间资源的引入，可以归纳得出结论：关联发现的整体目标实质上就是构建多个资源之间的关联数据网络（linked data network），关联的程度越高，则可以认为是关联发现的效果越好。

在关联发现的过程中，关联数据网络不断发生着变化，具体体现在：

（1）新的资源节点会不断增加或减少；

（2）资源的类别会发生变化，如会有新的资源类别被引入；

（3）既有资源节点之间的链接不断增加或减少；

如果将原始网络定义为关联数据网络 I 代，输出网络定义为关联数据网络 II 代，那么其中的每一个演变的版本可以视为 1.1 代、1.2 代等。

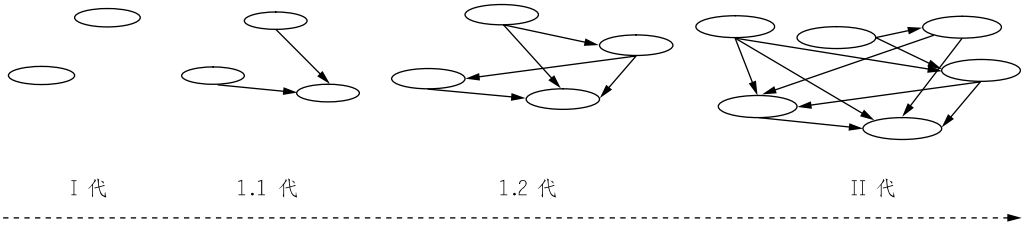


图 4　关联发现的整体目标

因此，关联发现的过程，其本质就是关联数据网络不断演变的过程。

在关联数据网络的演化过程中，有一个方向性的问题，即：网络可以按照什么样的方向演化？为此，需要在构建网络的时候给出其结构的描述模型，该模型通常可以采取 RDF 词表（RDF vocabulary）进行形式化描述。此外，在关联数据网络的演变过程中，还可以借助社会网络分析学（Social Network Analysis，SNA）的一些量化指标来衡量关系网络的演变效果，譬如通过社会联结的密度、强度、对称性、规模等来说明特定的行为和过程。另外，还可以采取一些可视化分析的软件对关联数据网络进行展示。

4　关联发现过程的复杂性特征

在关联开放数据环境下，要实现数据之间的关联发现，还存在着一些挑战。关联发现往往不是一个简单的过程，它具有较强的复杂性，本文将这种复杂性归纳为“多任务、多路径、多步骤”，具体分析如下：

（1）关联发现是一个“多任务”的过程

在最终建立两个资源的链接之前，往往需要更多的处理步骤，不同步骤所采取的发现方法也不尽相同。本文将这些方法初步归纳为五种任务类型，即词表映射、资源比对、资源鉴别、去重、属性 IRI（Internationalized Resource Identifiers）化，如表 1 所示。

由于描述信息普遍存在着不规范性，关联发现过程中需要进行很多的文本处理，不同的文本类型则要求有不同的相似度算法。如：字符串文本可能会用到 Jaro、基于词频统计的内容相似度计算，数值和日期可能会用到数值距离等算法。

（2）关联发现是一个“多路径”的过程

关联发现往往会同时存在着多条发现路径，以发现科学数据和科技文献的关联为例，较常见的思路是以“科学数据——文档特征——科技文献”为途径，即根据“文档相似度”来寻找二者之间的“内容相关性”。但这种采取单一方法发现关联，其效果往往极其有限。为了发现更多的关联，则需要借助于一些外部特征，如：分别根据“科学数据——科研人员——科技文献”、“科学数据——科研项目——科技文献”路径，即根据科学数据的采集者和科技文献的作者、科学数据的隶属项目和科技文献的基金项目，分别发现“科学数据——科技文献”之间的间接关联。图 5 显示出了这种路径的多重性。

（3）关联发现是一个“多步骤”的过程

关联的发现往往需要分解成多个连续的步骤。以上面的“科学数据——科研人员——科技文献”为例，即根据科学数据的采集者和科技文献的作者发现数据和文献之间可能的关联，但由于人员同名情况非常严重，在实际操作的过程中需要先构建“人员——机构”的关联，然后才能基于“人员”及其“机构”信息，来构建“人员——数据”之间的关联。可以看出，在整个过程中，关联的发现被分解成多个步骤，某些步骤甚至是高计算资源消耗的、

表1 关联发现过程中的任务类型

任务类型	含 义	说 明	相关研究
词表映射	用以实现两个异构RDF资源之间的转换	如:可以将科学数据库项目中关于一个联系人的信息csdb:Contact映射成一个人员信息foaf:Person和一个机构信息vcard:Organisation。	如RDF映射语言R2R[17]等
资源比对	用以计算两个RDF资源之间的相似度	如:在文献数据库中,会出现两个同名作者,为了确定他们的身份,需要考虑其他的属性(如所在的工作单位)来计算两条RDF资源的图形相似度。	相似度计算的研究如simmetrics①、SemMF② 等
资源鉴别	根据一个RDF资源的属性,通过与规范记录比对,获得该资源的规范名称(URI)	与资源比对不同,资源鉴别往往需要借助于规范库,如根据人名规范库确定“鲁迅”和“周树人”是同一个作者。	
去重	根据资源比对结果,将两个被认为同指的资源进行合并。	记录去重是数据库领域中研究较多的部分。在语义网领域,RDF则赋予待去重资源以更多的上下文信息。	
属性IRI化	将RDF资源的属性替换成IRI	如:将值为一段文本的dc:subject,更换成一个指向某个分类条目的IRI。	

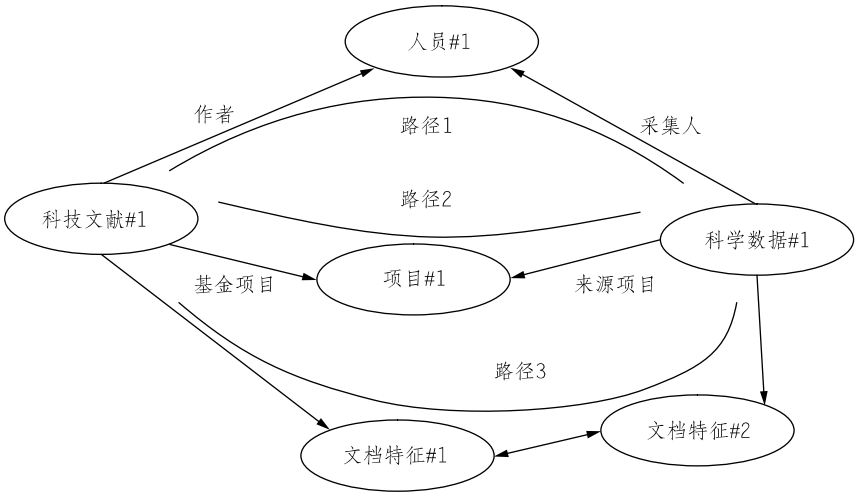


图5 关联发现的“多路径”特性

长时间的,整个关联发现过程是一个贯穿“多步骤”而“螺旋式上升”的过程。

综上所述,关联发现是一个“多任务、多路径、多步骤”、面向整个关联数据网络的、无法一步到位

① http://sourceforge.net/projects/simmetrics/

② http://sites.wiwiss.fu-berlin.de/suhl/radek/semmf/doc/index.html

的过程。由于关联数据网络中不断有新的资源类型和资源实体被引入，关联的发现甚至会成为一个常态化的任务。这种复杂性决定了关联发现必须借助于一种高度可配置的、可扩展的、自动化的框架。

5 结语

通过以上对关联发现过程的定位、目标与复杂性特征进行分析，可以得出如下结论：

(1) 关联数据的应用过程可以分成三个阶段：关联数据的发布、互联与消费。关联发现过程处于其中的第二阶段，其输入是阶段一生成的关联数据网络 I 代(弱关联的数据集群)，输出是关联数据网络 II 代。

(2) 关联发现的局部目标是建立起两个或者多个资源对象之间的直接关联。关联发现的整体目标，实质上就是构建多类资源之间的关联数据网络。关联发现的过程，其本质就是关联数据网络不断演变的过程。

(3) 关联发现是一个高度复杂、“多任务、多路径、多步骤”、面向整个关联数据网络的过程。由于关联数据网络中不断有新的资源类型和资源实体被引入，关联的发现甚至会成为一个常态化的任务。

在以上结论的基础上，可以看出，目前流行的关联发现框架还存在着一些不足：

(1) 缺乏对整个网络演变过程的支持。目前流行的关联发现框架忽略了关联发现是一个面向整个关联数据网络的、多次迭代的、无法一步到位的过程，因此缺乏对整个发现过程的宏观支持。

(2) 支持的任务单一。就目前的调研情况来看，大部分关联发现框架仅关注其中某些任务，如 SILK、RDF-AI，以及 LIMES，都只是关注于如何实现资源比对，关注于其中比对规则的表达，以及比对算法的优化策略。

(3) 缺乏流水线机制。在多任务之间缺乏灵活的连接，虽然关联数据集成框架 LDIF 注意到了这一点，将 LDSpider、R2R、SILK、Sieve 等多个工具引入框架，但由于它没有采用统一的描述理论和配置语言，多个过程之间还存在着明显的隔断，用户仍需要熟悉不同的软件工具，并分别编写遵循不同语法的任务脚本。

因此，关联发现技术的研究急需新的面向整个关联数据网络的、支持完整网络演变过程的、支持“多任务、多路径、多步骤”无缝集成的理论、方法与框架。

参考文献

[1] Berners-Lee T. Design issues: Linked data [EB/OL]. [2012-10-10]. http://www.w3.org/DesignIssues/LinkedData.html.

[2] W3C community projects: Linking open data [EB/OL]. [2010-07-12]. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData.

[3] State of the LOD cloud [EB/OL]. [2011-09-10]. http://www4.wiwiss.fu-berlin.de/lodcloud/state/.

[4] Linked data: Open research problems [C/OL]// World Wide Web Conference 2010. http://www.slideshare.net/juansequeda/07-openresearchproblems.

[5] Hassanzadeh O, Consens M. Linked movie data base [C]// Proceedings of LDOW2009. Madrid, Spain; April 2009.

[6] Volz J, Bizer C, Gaedke M, et al. SILK—A link discovery framework for the Web of data [C]//Proceedings of LDOW 2009. Madrid, Spain; 2009.

[7] Isele R, Jentzsch A, Bizer C. SILK server—Adding missing links while consuming linked data [C]//Proceedings of 1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, China (November 2010).

[8] Hassanzadeh O, Lim L, Kementsietsidis A, et al. A declarative framework for semantic link discovery over relational

data [C] //Proceedings of the 18th international conference on World Wide Web. ACM, 2009: 1101 - 1102.

[9] Ngomo A C N, Auer S. LIMES: A time-efficient approach for large-scale link discovery on the Web of data [C] //Proceedings of the Twenty-Second international joint conference on Artificial Intelligence—Volume Volume Three. AAAI Press, 2011: 2312 - 2317.

[10] Scharffe F, Liu Y B, Zhou C G. RDF-AI: An architecture for RDF datasets matching, fusion and interlink [C] //Proceedings of the IJCAI 2009 workshop on Identity, reference, and knowledge representation (IR-KR), Pasadena, CA US, 2009.

[11] Schultz A, Matteini A, Isele R, et al. LDIF—Linked data integration framework [C] //Proceedings of 2nd International Workshop on Consuming Linked Data (COLD 2011), Bonn, Germany, October 2011.

[12] The D2RQ platform v0.7—Treating non-RDF relational databases as virtual RDF graphs [EB/OL]. [2010 - 09 - 12]. http://www4.wiwiss.fu-berlin.de/bizer/d2rq/spec/.

[13] triplify.org: Overview [EB/OL]. [2010 - 09 - 12]. http://triplify.org/.

[14] Pubby—A linked data frontend for SPARQL endpoints [EB/OL]. [2008 - 07 - 17]. http://www4.wiwiss.fu-berlin.de/pubby/.

[15] Hausenblas M. Linked data applications [EB/OL]. [2013 - 07 - 29]. https://bitbucket.org/mhausenblas/linked-data-applications/src/7c6962007b5c/release/lod-app-tr-2009-07-26.pdf.

[16] Haslhofer B, Popitsch N. DSNotify-detecting and fixing broken links in linked data sets [C] //Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on. IEEE, 2009: 89 - 93.

[17] Bizer C, Schultz A. The R2R framework: Publishing and discovering mappings on the Web [C] //1st International Workshop on Consuming Linked Data (COLD 2010), Shanghai, November 2010.

沈志宏 中国科学院计算机网络信息中心高级工程师。

通讯地址: 北京市海淀区中关村南四街4号。邮编: 100190。

黎建辉 中国科学院计算机网络信息中心正高级工程师,博士生导师。通讯地址同上。

张晓林 中国科学院国家科学图书馆研究员,博士生导师。

通讯地址: 北京中关村北四环西路33号。邮编: 100190。

(收稿日期: 2013 - 02 - 27; 修回日期: 2013 - 05 - 14)