图书情报工作
第57卷 第14期 2013年7月

关联数据互联技术研究综述:应用、方法与框架*

■ 沈志宏 黎建辉 张晓林

[摘 要] 以关联数据互联技术为研究对象,分别从应用实例、方法和算法、框架三个方面阐述其研究现状。指出关联数据的互联技术还需要进一步的深入,主要包括:关联数据互联应用领域需要进一步拓展,关联算法需要考虑来自大关联数据的需求,关联发现框架需要面向任务支持更丰富的任务类型,并提供流水线机制和全局的规划以及流程控制能力。

[关键词] 关联数据 互联 关联发现框架
[分类号] TP393 G350
DOI:10.7536/j.issn.0252-3116.2013.14.021

1 关联数据的概念与应用

1.1 关联数据的概念

T. Berners-Lee 在2006年提出了"关联数据(linked data)"的概念[1],并制订了关联数据关于内容描述的"四大基本原则":①使用URI来标识事物(use URIs as names for things);②使用HTTP URI使人们可以访问到这些标识(use HTTP URIs so that people can look up those names);③当有人访问到标识时,提供有用的信息(when someone looks up a name, provide useful information);④尽可能提供关联的URI,以使人们可以发现更多的事物(include links to other URIs so that they can discover more things)。

从技术角度来讲,关联数据可以理解成一组最佳实践的集合[2],它采用RDF(resource description framework)数据模型,利用URI(统一资源标识符)命名数据实体,来发布和部署实例数据和类数据,从而可以通过HTTP协议揭示并获取这些数据,同时强调数据的相互关联、相互联系以及有益于人机理解的语境信息。Wikipedia认为,关联数据用来指代一组采用URI和RDF来实现语义网上的数据、信息及知识的公开、共享与联结的最佳实践。

2007年5月,W3C的关联开放数据(linking open data,LOD)运动正式启动[3],该运动提倡将Web上的开放数据源以RDF的方式发布出来,同时生成数据源之间的RDF链接,以供关联数据浏览器、搜索引擎以及更高级的应用程序使用。很快关联数据概念就流传开来,现在已成为互联网的热门研究领域,从2008年起在年度互联网大会(WWW Conference)上都举行关于Linked Data on the Web(LDOW)的专门会议。另外在ISWC(International Semantic web Conference)、DIST(Data Integration through Semantic Technology)大会上也经常召开专门的会议。2012年4月在法国里昂召开的WWW2012大会上,就包含了LDOW2012工作组会议,讨论的话题涉及关联数据的自动关联、分布式异构性以及互操作。

在LOD项目启动后短短的三年中,越来越多的数据拥有者将他们的数据以关联数据的形式发布到Web上。截至2011年9月,LOD已收录295个数据集(见图1)。按数据涉及的主题领域[4]可分为:多媒体(如BBC、CNET、ThomsonReuters);文献出版物(如DBLP、CiteSeer、EPrints、SWC);生命科学(如UniPort、PubMed、CAS、Bio2RDF);地理数据(如GeoNames、LinkedGeoData);社交网络(如Flickr、FaceBook)以及跨领域的数据(如DBPedia、Freebase、YAGO、UMBEL、OpenCyc)。

1.2 关联数据的互联成为首要问题

从关联数据"四项基本原则"可以看出,关联数据的四大基本要素为:URI、HTTP URI、RDF以及RDF链

*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"(项目编号:XXH12504-01-02)和国家自然科学基金重点资助项目"面向非常规突发事件应急管理的云服务体系和关键技术"(项目编号:91224006)研究成果之一。
[作者简介] 沈志宏,中国科学院计算机网络信息中心高级工程师,博士,E-mail:bluejoe@cnic.cn;黎建辉,中国科学院计算机网络信息中心正高级医师,博士生导师;张晓林,中国科学院国家科学图书馆研究员,馆长。
收稿日期:2013-06-03 修回日期:2013-07-02 本文起止页码:125-133 本文责任编辑:王传清

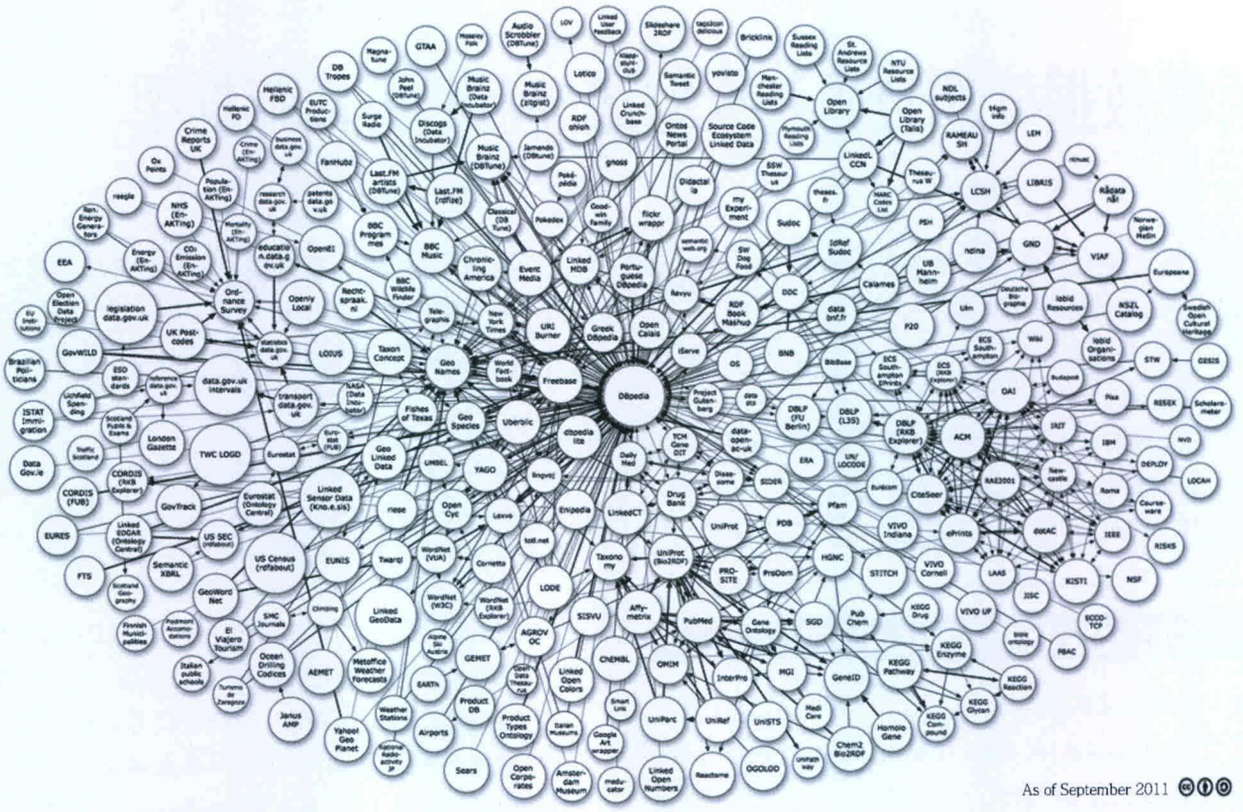

图1　LOD 数据云图[5]

接(RDF link)。RDF 链接在 Web 数据的发现中发挥了重要的作用。根据梅特卡夫定律(Metcalfe's law)[6]:网络结点之间的连接越多,网络的价值则会越大,因此这些关联本身就能产生价值。作为关联数据的实践项目,W3C 的关联开放数据运动也强调了 RDF 链接的重要作用:RDF 链接可以让用户通过语义网浏览器(semantic Web browser)从一个数据源中的某个数据项导航到另外一个数据源中相关的数据项中。RDF 链接同时也可以为语义网搜索引擎所使用,从而为抓取到的数据提供完善的检索和查询功能。另外,由于查询结果是结构化的数据而非指向 HTML 页面的链接,它们能够被其他的应用程序使用。

然而,从 2011 年 9 月 LOD 收录的情况来看[7],295 个数据集中包含了 310 亿条 RDF 三元组(RDF triples),其中包含有 5 亿条 RDF 链接(一个 RDF 链接本身就是一个 RDF 三元组)。表 1 显示出 LOD 数据集在不同领域中的分布情况。从表 1 中可以看出,RDF 链接在 RDF 三元组中所占的比重还不到1.6%,这种弱关联性远远不足以支持高级的知识关联发现。

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

表1　LOD 数据集在不同领域中的分布情况

领域	数据集	三元组	外向链接	链接%
媒体	25	1,841,852,061	50,440,705	2.74
地学	31	6,145,532,484	35,812,328	0.58
政府	49	13,315,009,400	19,343,519	0.15
出版	87	2,950,720,693	139,925,218	4.74
跨领域	41	4,184,635,715	63,183,065	1.51
生命科学	41	3,036,336,004	191,844,090	6.32
用户应用	20	134,127,413	3,449,143	2.57
合计	295	31,634,213,770	503,998,829	1.59

近年来热点话题。在 LDOW2010(WWW 2010 Workshop on Linked Data on the Web)会上,数据互联成为会议的一大专题(其他的专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 COLD 2010(International Workshop on Consuming Linked Data)发布的几大开放问题中[8],居于首位的就是关联数据的互联算法(interlinking algorithm),其次是溯源与信任、数据集动力学、用户界面、分布式查询、评估(见图 2)。

此外,在 LDOW 2012 会议的 16 个报告中,有三篇文献都涉及到关联发现的话题,分别介绍有声电台档案的自动互联、智能城市数据的互联、基于互联视角的数据网络的交互技术、基于 VOID 对互联数据集的查询等。在 COLD 2011 会议中,A. Schultz 等推出了一款

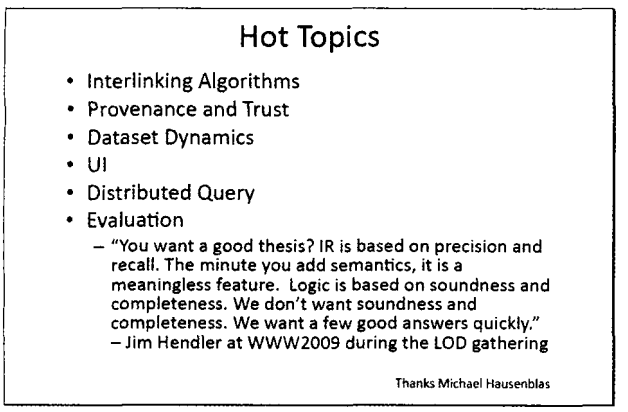

图 2 COLD 2010 发布的关联数据热点话题

集成化关联发现与集成框架 LDIF(Linked Data Integration Framework)[9]。在 WWW 2012 会议上,A. Schultz 又介绍了 LDIF 的新版本[10]。

与国际上关联数据的研究进展相比,从见诸专业刊物的文章来看,关联数据在国内的研究尚不普及[11],并且仅局限于图书馆情报领域,还未引起计算机领域、数据库领域的广泛重视。基于此背景,本文以关联数据互联技术为研究对象,分别从应用实例、方法、框架三个方面阐述其研究现状。

2 关联数据互联的应用实例

作为全球第一个关联数据化的联合目录,瑞典联合目录(LIBRIS,http://libris.kb.se)可以作为图书馆届关联数据互联的示范应用。LIBRIS 自 2008 年起,发布了来自于 170 余个成员馆的 600 万条以上书目记录与 25 万条规范文档记录的 RDF 记录,此外还发布了规范记录与书目记录以及规范记录之间的关联[12],甚至包含国会图书馆主题词表数据的关联,见图 3(图片来源:http://blog.libris.kb.se/semweb/)。

此前,在 DC-2008 年会上,P. Miller 的 Keynotes[13] 探讨了图书馆界在语义 Web 中可以承担的角色,并专门介绍了美国国会图书馆将其主题表(LCSH)以 SKOS 编码的项目(http://lcsh.info)。从 LIBRIS 到 LCSH.info,是一个典型的跨数据集的数据互联应用。

在媒体领域,LinkedMDB[14] 是一个比较实际的跨数据集数据互联的应用实例,LinkedMDB 实现了与其他 LOD 数据集的互联,包括 DBpedia/YAGO、Geonames、FlickrWrapper、RDF Book Mashup、Musicbrainz、Revyu.com 等。这种关联效果见图 4。

LinkedMDB 包含实体 233 103 项,指向其他 LOD 数据的关联数为 162 199 项,关联数目统计见表 2。

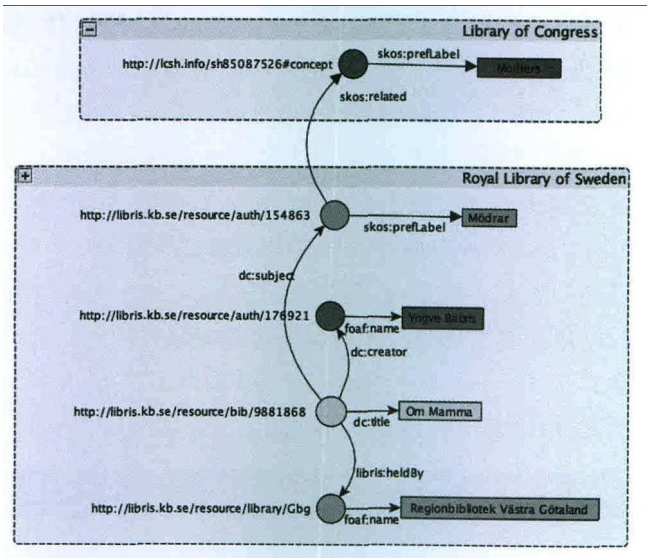

图 3 瑞典联合目录联与国会图书馆主题词表的关联

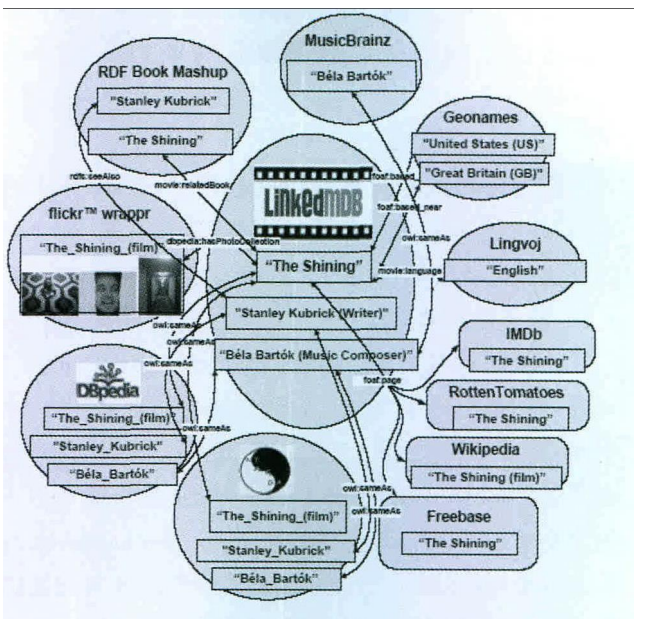

图 4 LinkedMDB 与其他 LOD 数据集之间的关联[12]

表 2 LinkedMDB 自动发现的关联数目

目标	类型	数目(条)
DBpedia	owl:sameAs	30 354
YAGO	owl:sameAs	30 354
flickr wrappr	DBpedia:hasPhotoCollection	30 354
RDF Book Mashup (Books)	movie:relatedBook	700
RDF Book Mashup (Authors)	rdfs:seeAlso	12 990
MusicBrainz	owl:sameAs	2 207
GeoNames	foaf:based near	27 272
GeoNames	owl:sameAs	272
Lingvoj	movie:language	28 253
IMDb	foaf:page	271 671

在生命科学领域,数据互联的应用较多。如:Diseasome Map[15] 应用整合了不同生命科学的数据源,

以关联数据的形式发布了4 300个紊乱和疾病的基因,并构建了一个紊乱基因关系网络。Linked Life Data[16]整合了UniPort、PubMed、Entrez Gene等20余个数据源,并据此提供了关联式的检索和浏览服务。印第安纳大学的Dong Xiao和Ding Ying等[17]开发了Chem2Bio2RDF Dashboard,该系统集成了化学、生物、药物领域的关联数据,用以发现两个实体或概念之间的路径。美国马里兰大学和委内瑞拉西蒙玻利瓦尔大学的M. E. Vidal和L. Raschid等共同开发了BioNav[18]框架,用以发现药物和疾病之间的关系。

3　关联数据互联方法研究

在关联数据互联方法和算法方面,有一部分研究关注于如何在关联开放数据环境下通过一些自动和半自动的方法来创建数据之间的关联。白海燕等[19]将关联数据之间的关联构建概括为映射关联(owl:sameAs)和非映射关联,并以书目数据关系为例介绍基于规则的关联构建方法。

作为关联数据发布的权威教程,C. Bizer等[20]提出,关联的创建可以采用两种算法:基于模式(pattern)的算法以及基于属性的复杂算法。

3.1　基于模式的算法

该算法比较简单,适用于拥有唯一标识符的资源对象,如DBPedia包含有《哈利·波特与混血王子》图书的RDF描述,并记录该图书的ISBN编号为0747581088,同时由于RDF Book Mashup[21]采用了形如http://www4. wiwiss. fu - berlin. de/bookmashup/books/{ISBN number}的URI模式,那么就可以创建DBPedia与RDF Book Mashup之间的关联,如:

< http://dbpedia. org/resource/Harry_Potter_and_the_Half-Blood_Prince >

owl:sameAs

< http://www4. wiwiss. fu -berlin. de/bookmashup/books/0747581088 >

3.2　基于属性的算法

该算法适用于没有唯一标识符的资源对象,如针对某个地理位置,可以采取涉及该位置的文章标题、经纬度、国家、行政区划、人口等属性信息,建立Dbpedia与Geonames之间的地理位置的映射。

Y. Raimond等[22]结合音乐数据集,介绍了自动创建关联的两种途径:基于实体的文本映射(又分为简单文本查找、扩展文本查找)及基于RDF图形相似度计算的映射。

3.2.1　基于实体的文本映射　该方法又分成简单文本查找和扩展文本查找方法。简单文本查找可通过遍历式匹配或者SPARQL WHERE语句来实现,扩展文本查找则借用分类、分面、类型特征、属性特征等进一步对资源进行限定。

作为例子,以下代码表示通过SPARQL语句查询LIBRIS的规范文档数据集,来获取关于William Gibson的规范记录[17]:

```
DESCRIBE ? s WHERE
{
? s a foaf:Person.
? s foaf:name "William Gibson".
}
```

3.2.2　基于RDF图形相似度计算的映射　RDF图形相似度计算相对比较复杂,以图5为例,计算RDF图形的相似度可以分解成如下三个步骤:

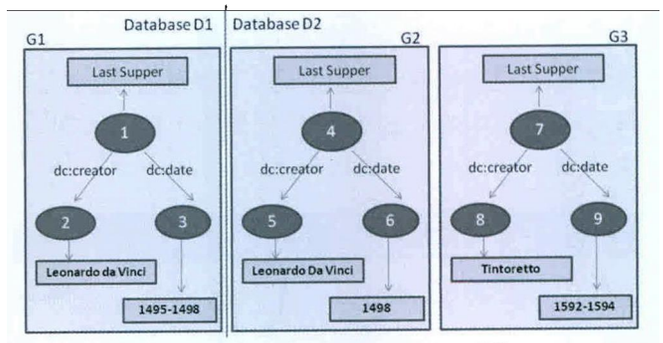

图5　三个待比对的RDF图形[20]

- 标注RDF图形中所有的结点,如表3所示:

表3　标注RDF图形中所有的结点

图	主语(S)	谓词(P)	对象(O)
G1	1:Last Supper	dc:creator	2:Leonardo da Vinci
	1:Last Supper	dc:date	3:1495 - 1498
G2	4:Last Supper	dc:creator	5:Leonardo Da Vinci
	4:Last Supper	dc:date	6:1498
G3	7:Last Supper	dc:creator	8:Tintoretto
	7:Last Supper	dc:date	9:1592 - 1594

- 计算每个结点的相似度,计算结果如表4所示:

表4　RDF结点的相似度

结点1	结点2	结点相似度
1	4	1
2	5	0.94
3	6	0.44
1	7	1
2	8	0.11
3	9	0.56

• 根据结点的相似度，计算图的相似度，如表5所示：

表5 RDF图形的相似度

图		映射	图相似度
G1	G2	Mg1：Mg2 = {(1,4),(2,5),(3,6)}	0.79
G1	G3	Mg1：Mg3 = {(1,7),(2,8),(3,9)}	0.56

邓兰兰等[23]针对同构和异构模式下的关联关系创建技术作了详细的综述。同构模式下的关联创建方法包括单一方法（属性值相似度算法、图形相似度算法等）、组合方法（基本算法聚合、分类模型等），异构模式下的关联创建的主要策略则是先建立异构数据之间的模式映射[24-25]，然后再运用同构方法来创建实例之间的关联。此外，A. Nikolov等人介绍了一种通过聚类算法推导出模式级别关联的方法[26]。M. Rowe介绍了如何生成Facebook与Twitter和Myspace.com之间的链接[27]。

可以看出，以上方法最终会涉及到属性文本的相似度算法。目前针对属性相似度计算的研究比较成熟：字符串相似度如jaro、jaroWinkler、qGram、概念距离、levenshtein距离、Jaccard、Dice等算法。此外，还有基于TF-IDF及向量空间模型的文档相似度算法、字符串集合相似度算法等。

4 关联发现框架研究

4.1 关联发现框架

由于数据集互联的过程极其复杂，人们开始研究适用于关联数据互联的框架，从而实现关联发现、关联集成任务的自动化和流程化。

首先值得一提的是基于规则的关联发现框架SILK[28-29]，SILK允许用户制定SILK-LSL（SILK Link Specification Language）规则文件，并藉此自动生成出不同数据集之间的实例级的链接。SILK不仅能够生成数据之间的owl:sameAs关联，也可以生成其他类型的关联，如：DBPedia电影与LinkedMDB导演之间的dbpedia:director关联。原理上，SILK主要通过给定的两个数据集中数据的属性相似度来计算它们之间的关联关系。SILK-LSL支持的相似度算法包括Jaro距离、Jaro-Winkler、Levenshtein算法、q-grams文本相似性计算、文本等价性、数值距离、日期距离等算法。最新版本的SILK推出了SILK workbench，允许用户通过图形化的界面在线完成SILK-LSL配置的定义，并启动关联发现的任务。SILK workbench中的Linkage Rule Editor（关联规则编辑器）如图6所示：

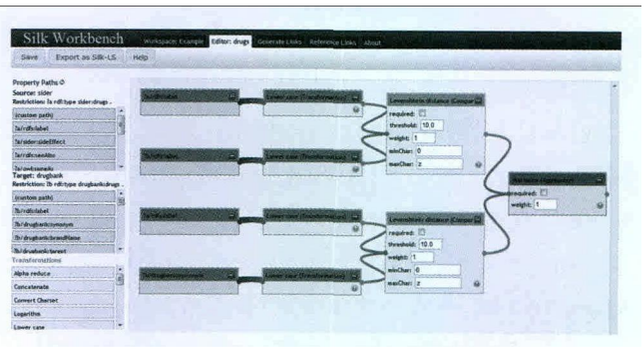

图6 SILK的关联规则编辑器

类似地，O. Hassanzadeh等提出了一款完全针对关系型数据的语义连接发现框架LinQuer（Linkage Query Writer）[30]，并且同时提出了一套声明式语言LinQL。

A. C. N. Ngomo等[31]根据三角形不等式，提出了两种高效的近似距离计算方法，并集成多种比对算法构建了关联框架LIMES（Link Discovery Framework for Metric Spaces）。

RDF-AI[32]是另外一款基于用户配置的RDF数据集的融合和互联框架。RDF-AI接受两个RDF数据集作为输入，产生由两个数据集融合形成的新数据集，或者产生它们比对的结果数据集。如图7所示，RDF-AI的架构包括5个相对独立的模块：预处理、匹配、融合、互联、后处理。RDF-AI提供了灵活的配置接口，允许用户定义融合任务的输入和输出。通过对AKT Eprints数据集中的314条记录和Rexa数据集中的2 103条记录进行关联，证明通过采取合适的预处理，RDF-AI关联的正确率高达95.9%，这比KnoFuss[33]针对同样的数据集得到的正确率（92%）还要理想。

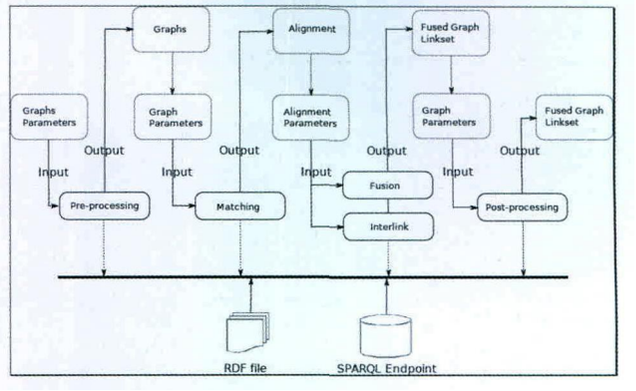

图7 RDF-AI系统架构

数据互联往往不是一个孤立的过程，它通常会伴随着数据转换、数据融合等任务。基于该思路，LDIF[9]为用户提供了一个关联数据集成处理框架，该框架有效集成了如LDSpider、R2R[34]、SILK、Sieve[35]等关联数据的应用工具，可以帮助用户构建完整的流水线，包括数据采

集、Schema 映射、标识识别(identity resolution)、质量评估与数据融合、输出这 5 个步骤。LDIF 提供了三种使用模式:单机 In-Memory 版本、单机 RDF 库版本以及集群 Hadoop 版本。图 8(图片来源 http://www4.wiwiss.fu-berlin.de/bizer/ldif/)显示了 LDIF 给出的关联数据应用程序的层次架构,从中可以看出,LDIF 处于应用层与关联数据层之间,负责数据的访问、集成和存储。

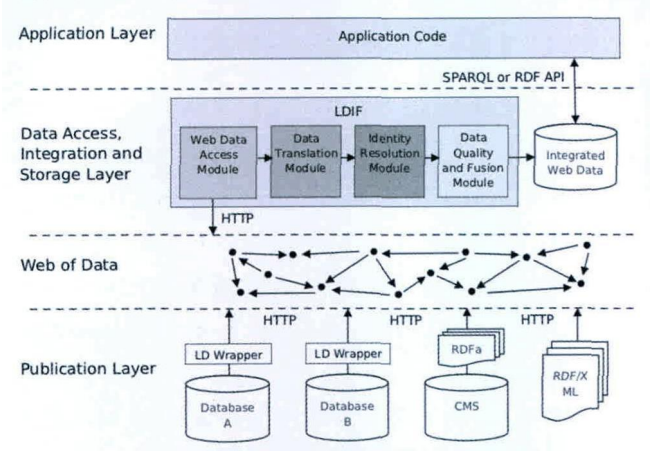

图 8　LDIF 给出的关联数据应用程序架构

由于中文处理的特殊要求和资源语义化现状的限制,上述通用的发现框架和算法无法在国内的项目中直接应用,但它们对关联开放数据环境下中文语种的资源间的关联发现研究无疑有着极大的借鉴意义。如陶俊等[36]重点研究了 RDF 映射语言 R2R,并基于 R2R 提出了面向 LOD 的关联系统 VocaR。

4.2　关联发现框架比较

值得一提的是,S. Wölger 等[37]针对目前流行的一些关联发现工具和框架进行了综述与对比,调研对象涉及 RDF-AI、SILK、Knofuss、LinQuer、Interlinking Distributed Social Graphs[27]、Guess What[38]、PoolParty[39]、Grapple User Modeling Framework (GUMPF)[40]、CaMiCatzee、ExpLOD、GNAT 这 11 个工具或系统,重点针对自动化程度、人工参与、适用领域、匹配方法、采用本体、输入形式、输出形式、后处理、数据访问方式这 9 个方面进行了比较,本文予以摘录并整理,形成表 6。

可以看出,各种关联发现框架都实现了自动、半自动的发现方式,在进行 RDF 资源匹配的时候也分别采取了不同的匹配方法。作为补充,本文从支持理论、任务覆盖面、可配置性以及任务流水线这 4 个方面针对 SILK、RDF-AI、LinQuer、LIMES、LDIF 等 5 类流行的框架做出了二次比较(见表 7)。

表 6　常见互联工具与框架比较(1)

工具	自动化程度	匹配方法	输入	输出	数据访问
SILK	半自动	字符串匹配,相似度算法	SILK-LSL 配置文件	linkset	SPARQL
RDF-AI	半自动	字符串匹配,词关系匹配	两个数据集和参数	合并得到的数据集或者包含 sameAs 关系的实体对应表	Jena API
LinQuer	半自动	同义词,下位词,字符串匹配,关联表达式	SQL 数据库	SQL 数据库	SQL 查询
Knofuss	自动	字符串匹配,自适应学习	源、目标知识库,融合本体	RDF 陈述,合并得到的集合	本地拷贝
PoolParty	半自动	概率字符串匹配	RDF 三元组,Web 界面	包含 sameAs 关系的 RDF 三元组,包含 RDFa 的 HTML 页面	Web 界面

表 7　常见互联工具与框架比较(2)

工具	支持理论	任务覆盖面	可配置性	任务流水线
SILK	针对比对任务给出了概念体系:数据源、关联规则语言、黄金集合(gloden set)	仅提供了比对任务,SILK-LSL 仅提供了 Compare 和 Aggregate 配置	SILK-LSL 文件	单一任务、固定流程:数据源、分块、关联生成、过滤、输出
RDF-AI	比对任务包括 5 个步骤:预处理、匹配、融合、互联、后处理。RDF-AI 以数学形式描述了每个步骤	仅提供了比对任务	无配置语言	固定流程:预处理、匹配、融合、互联、后处理
LinQuer	无	互联	通过程序提交 LinQL 查询	固定流程
LIMES	三角形不等式,计量空间(metric space)	仅提供了比对任务	无配置语言	固定流程:样本计算、过滤、相似度计算、序列化
LDIF	数据互联往往不是一个孤立的过程,它通常会伴随着数据转换、数据融合等任务	包括数据采集、Schema 映射、标识识别(identity resolution)、质量评估与数据融合、输出	需要单独配置 LDSpider、R2R、SILK、Sieve 的配置文件,并给出集成任务的配置文件	目前是固定流程,以后会考虑工作流

基于以上对比,本文认为,目前流行的这些框架尚存在着如下不足:

- 支持的任务类型单一。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(matching),关注于其中的比对规则的表达以及比对算法的优化策略。
- 缺乏流水线机制,不支持多次、多路关联发现

过程的串接。以根据科研人员发现科学数据和科技文献的关联路径为例，它需要分解成建立“人员——机构”、“文献——人员”、“数据——人员”等多个任务，这些任务之间如何衔接？关联数据集成框架 LDIF 显然注意到了这一点，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。

• 缺乏全局的规划和控制能力。在一个完整的关联数据网络中，发现两个资源之间的关联，从时间上往往需要历经很长的计算过程，从空间上往往需要涉及到多类资源结点，甚至覆盖至整个关联数据网络。因此，关联发现往往需要提前进行多步、多路发现路径的规划，而目前的框架明显缺乏全局规划（手动或者自动的）的能力。另外对运行过程中发生的耗时、失败等情况，目前的框架也缺乏有效的控制能力。

5 结 语

综上所述，关联数据的互联无论在应用、方法，还是在关联发现框架研究上，都得到了较为充分的发展，应用丰富、方法多样，而且各框架稳定可靠。

然而，本文认为，在应用、方法以及框架方面，关联数据的互联还存在着如下一些问题与不足：

• 关联数据的互联应用目前主要局限在媒体、出版与生命科学领域，尚缺乏在其他领域中的成熟范例。以中国科学院科学数据库项目（http://www.csdb.cn）中的青海湖数据库（http://www.qinghailake.csdb.cn）为例，其数据内容涉及地学、生态学、生物学等多种数据，一旦围绕该专题建立起完整的、跨学科的关联数据网络，对高级 e-Science 应用的开发就很有意义。另外，国内外目前有很多研究关注于科学数据与科技文献的集成与关联服务[41-42]，根据《第四范式：数据密集型科学发现》的观点，科学数据与科技文献的互操作具有较大的意义[43]。但根据目前调研的情况来看，还没有在关联数据环境（linked data context）下实现二者关联服务的先例。

• 关联数据的互联方法与算法方面，由于关联数据的技术基石是 HTTP 和 RDF，因此现有的方法更多地关注 SPARQL 的自动构造和查询优化以及基于 RDF 图的高效路径匹配。随着大数据（big data）的提出，现在已有学者在提“大关联数据”（big linked data）[44]。对应于大数据的 3V 特征（volume，variety，velocity），大关联数据的互联算法同样需要应对来自于大数据量、语义异构、需要在线快速响应的需求。

• 如 4.2 部分所述，目前的关联发现框架支持的任务类型单一，缺乏流水线机制，而且缺乏全局的规划和控制能力。这种缺陷造成这些关联发现框架只适用于发现路径简单、待关联资源的种类少、关联网络不够复杂的场合，并且需要较多的人工数据整理和任务衔接工作。另外，在规则语言、比对算法模型、鉴别机制、中文信息处理等方面，目前的关联发现框架也存在着改进空间。

综上所述，关联数据的互联应用应进一步扩展至其他领域，构建跨学科领域的关联应用则具有更大的意义。关联数据的互联方法与算法应结合大关联数据的需求，增强对大数据量、语义异构、在线快速响应的支持。新的关联发现框架需要面向任务支持更丰富的任务类型，并提供流水线机制和全局的规划以及流程控制能力。

参考文献：

[1] Berners-Lee T. Design issues: Linked data [EB/OL]. [2013-04-10]. http://www.w3.org/DesignIssues/LinkedData.html.

[2] Linked Data FAQ [EB/OL]. [2013-04-10]. http://structureddynamics.com/linked_data.html.

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

[4] 黄永文. 关联数据在图书馆中的应用研究综述[J]. 现代图书情报技术，2010(5):1-7.

[5] Cyganiak R, Jentzsch A. Linking open data cloud diagram [EB/OL]. [2012-10-01]. http://lod-cloud.net/.

[6] Hendler J, Golbeck J. Metcalfe's law, Web 2.0, and the Semantic Web[J]. Web Semantics: Science, Services and Agents on the World Wide Web, 2008, 6(1):14-20.

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

[8] Linked data: Open research problems. Consuming Linked Data Tutorial [EB/OL]. [2012-10-01]. http://www.slideshare.net/juansequeda/07-openresearchproblems.

[9] Schultz A, Matteini A, Isele R, et al. LDIF - Linked data integration framework[EB/OL]. [2013-06-13]. http://ceur-ws.org/Vol-782/SchultzEtAl_COLD2011.pdf.

[10] Schultz A, Matteini A, Isele R, et al. LDIF - A Framework for Large-Scale Linked Data Integration[C]//21st International World Wide Web Conference. Lyon: Developers Track, 2012.

[11] 刘炜. 关联数据：概念、技术及应用展望[J]. 大学图书馆学报，2011(2):5-12.

[12] Malmsten M. Exposing library data as linked data [EB/OL].

[2011 - 09 - 10]. http://citeseerx. ist. psu. edu/viewdoc/download? doi = 10. 1. 1. 181. 860&rep = rep1&type = pdf.

[13] Miller P. Why the Semantic Web matters [EB/OL]. [2011 - 09 - 10]. http://dc2008. de/wp - content/uploads/2008/09/dcmi20080925 - miller. pdf.

[14] Hassanzadeh O, Consens M. Linked movie data base [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 538/ldow2009_paper12. pdf?.

[15] Goh K I, Cusick M E, Valle D, et al. The human disease network [J]. Proceedings of the National Academy of Sciences, 2007, 104 (21): 8685 - 8690.

[16] Momtchev V, Peychev D, Primov T, et al. Expanding the pathway and interaction knowledge in linked life data [EB/OL]. [2013 - 06 - 13]. http://www. ontotext. com/sites/default/files/publications/LLD_semantic_web_challenge_2009. pdf.

[17] Dong Xiao, Ding Ying, Wang Huijun, et al. Chem2Bio2RDF Dashboard: Ranking semantic associations in systems chemical biology space [EB/OL]. [2013 - 06 - 13]. http://delos. zoo. ox. ac. uk/pub/2010/Proceedings/FWCS2010/03/Paper3. pdf.

[18] Vidal M E, Raschid L, Márquez N, et al. BioNav: An ontology-based framework to discover semantic links in the cloud of linked data[J]. The Semantic Web: Research and Applications Lecture Notes in Computer Science, 2010(6089):441 - 445.

[19] 白海燕,朱礼军. 关联数据的自动关联构建研究[J]. 现代图书情报技术, 2010 (2): 44 - 49.

[20] Bizer C, Cyganiak R, Heath T. How to Publish Linked Data on the Web [EB/OL]. [2012 - 03 - 08]. http://www4. wiwiss. fu - berlin. de/bizer/pub/LinkedDataTutorial/.

[21] Bizer C, Cyganiak R, Gau? T. The RDF book mashup: from web APIs to a web of data [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 248/paper4. pdf.

[22] Raimond Y, Sutton C, Sandler M. Automatic interlinking of music datasets on the semantic web [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 369/paper18. pdf.

[23] 邓兰兰, 李春旺. Web 数据关联创建策略研究[J]. 现代图书情报技术, 2011(5):1 - 6.

[24] Shvaiko P, Euzenat J. A survey of schema - based matching approaches[J]. Journal on Data Semantics IV, 2005(3730):146 - 171.

[25] Rahm E, Bernstein P A. A survey of approaches to automatic schema matching[J]. VLDB Journal, 2001, 10(4):334 - 350.

[26] Nikolov A, Uren V, Motta E. Data linking: Capturing and utilising implicit schema - level relations [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 628/ldow2010_paper07. pdf.

[27] Rowe M. Interlinking distributed social graphs [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 538/ldow2009_paper5. pdf.

[28] Volz J, Bizer C, Gaedke M, et al.: SILK - A link discovery framework for the web of data [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 538/ldow2009_paper13. pdf.

[29] Isele R, Jentzsch A, Bizer C. SILK server - Adding missing links while consuming linked data [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 665/IseleEtAl_COLD2010. pdf.

[30] 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. New York :ACM, 2009: 1101 - 1102.

[31] 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. Barcelona: AAAI Press, 2011: 2312 - 2317.

[32] Scharffe F, Liu Yanbin, Zhou Chunguang. RDF - AI: An architecture for RDF datasets matching, fusion and interlink. [EB/OL]. [2013 - 06 - 13]. https://www. sti2. at/public_html/pub/RDF - AI - an - Architecture - for - RDF - Datasets - Matching - Fusion - and - Interlink - IR - KR - IJCAI - 2009. pdf.

[33] Nikolov A, Uren V, Motta E, et al. Integration of semantically annotated data by the KnoFuss architecture [M]//Knowledge Engineering: Practice and Patterns. Berlin: Springer, 2008: 265 - 274.

[34] Bizer C, Schultz A. The R2R framework: Publishing and discovering mappings on the Web [EB/OL]. [2013 - 06 - 13]. http://ceur - ws. org/Vol - 665/BizerEtAl_COLD2010. pdf.

[35] Mendes P N, Mühleisen H, Bizer C. Sieve: linked data quality assessment and fusion[C]//Proceedings of the 2012 Joint EDBT/ICDT Workshops. ACM, Lyon , 2012: 116 - 123.

[36] 陶俊, 孙坦, 刘峥. 关联数据映射语言: R2R[J]. 中国图书馆学报, 2012,38(3):100 - 109.

[37] Woelger S, Siorpaes K, Buerger T, et al. A survey on data interlinking methods [EB/OL]. [2013 - 06 - 13]. http://www. insemtives. org/publications/A_Survey_on_Data_Interlinking_Methods. pdf.

[38] Markotschi T, V? lker J. Guess what?! - Human intelligence for mining linked data [C]// Proceedings of the Workshop on Knowledge Injection into and Extraction from Linked Data (KIELD) at the International Conference on Knowledge Engineering and Knowledge Management (EKAW), Lisbon, 2010.

[39] Schandl T, Blumauer A. PoolParty: SKOS thesaurus management utilizing linked data [J]. The Semantic Web: Research and Applications, 2010(6089):421 - 425.

[40] Leonardi E, Abel F, Heckmann D, et al. A flexible rule - based method for interlinking, integrating, and enriching user data[J]. Web Engineering, 2010(6189):322 - 336.

[41] Entrez cross - database search. [EB/OL]. [2010 - 09 - 10]. http://www. ncbi. nlm. nih. gov/Entrez.

[42] ICPSR, Find & Analyze Data. [EB/OL]. [2011 - 09 - 08]. http://www. icpsr. umich. edu/icpsrweb/ICPSR/index. jsp.

[43] Tansley S, Tolle K M. The fourth paradigm: data - intensive scientific discovery[R]. Microsoft Research, 2009.
[44] Hu Bo, Carvalho N, Laera L, et al. Towards big linked data: A large - scale, distributed semantic data storage[C]// Proceedings of the 14th International Conference on Information Integration and Web - based Applications & Services. New York :ACM,2012: 167 - 176.

Research Review on the Interlinking Technology of Linked Data: Applications, Methods and Frameworks

Shen Zhihong[1] Li Jianhui[1] Zhang Xiaolin[2]
[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190
[2]National Science Library, Chinese Academy of Sciences, Beijing 100190

[**Abstract**] This paper investigates the research status of the interlinking technology of linked data, from three perspectives of its application examples, the popular methods and algorithms, and the linked discovery frameworks. It is concluded that: interlinking applications of linked data require further development; the interlinking algorithms need to meet the requirements of the "Big Linked Data"; the linked discovery frameworks need to support richer types of tasks, and provide functionalities of pipelining, overall planning and process control.

[**Keywords**] linked data interlinking linked discovery framework

(上接第55页)

Analysis on Dissimilarities Among the Evaluation Models of Library Readers Satisfaction Degree

Li Zhifang Deng Zhonghua
School of Information Management, Wuhan University, Wuhan 430072

[**Abstract**] Through literatures surveying, we found that the current evaluation models of library readers satisfaction are mainly built based on SERVQUAL、LibQUAL + TM、Rodski Group、AHP and ACSI. This paper makes a comprehensive comparison among these models from their backgrounds, scale set, weight set, applicability as well as advantages and disadvantages. Then we can get the dissimilarities among these evaluation models. In the research, we find that the most obvious dissimilarities are their scales and applicability; Different evaluation models have different focuses in the measurement process and results.

[**Keywords**] readers satisfaction degree evaluation model evaluation scale dissimilarities

(上接第86页)

Study on Library Deprofessionalization Based on Recruitment Data

Huang Yongqin[1] Yao Xinlin[2]
[1]Department of Information Management, Shanghai Campus of Nanjing Political College, Shanghai 200433
[2]School of Management, Fudan University, Shanghai 200433

[**Abstract**] This paper studies the library job ads in the recent five years by using content analysis method. By analyzing the content of jobs, professional threshold and capacity requirements, we summarized six problems of library deprofessionalization problems. Based on above analysis, this paper gives some suggestions, including establishing qualification of professional librarians, transforming the professional education of LIS, and changing the concept of library services.

[**Keywords**] librarianship deprofessionalization job ads content analysis