中文链接数据构建
实施路径研究

王 汀　徐天晟　著

科学出版社

中文链接数据构建实施路径研究

王　汀　徐天晟　著

科学出版社

北　京

内 容 简 介

本书从如何构建中文语义链接数据应用的目标出发，着力解决三个问题：一是解决大规模领域本体库的自动化构建问题；二是面对目前的研究主要关注从英文网络百科数据源进行海量知识的抽取，而对中文或其他语言描述的数据源进行知识抽取的研究非常少的研究现状，探索如何构建开放域的中文海量知识库；三是面对目前关联数据的研究工作主要集中在实例级别上展开，而对于框架级别的关联数据构建则易被忽视的现状，探索解决框架级别上的中文大规模关联数据构建的问题.

本书适合从事语义 Web、知识工程和中文自然语言处理等相关领域的研究人员参考阅读.

图书在版编目(CIP)数据

中文链接数据构建实施路径研究/王汀，徐天晟著. —北京：科学出版社，2015

ISBN 978-7-03-044693-0

Ⅰ.①中… Ⅱ.①王…②徐… Ⅲ.①数据处理-研究 Ⅳ.①TP274

中国版本图书馆 CIP 数据核字(2015)第 124284 号

责任编辑：李静科/ 责任校对：彭 涛

责任印制：徐晓晨/ 封面设计：耕者工作室

科 学 出 版 社出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

北京虎彩印刷科技有限公司 印刷

科学出版社发行 各地新华书店经销

*

2015 年 6 月第 一 版 开本：720×1000 B5

2015 年 6 月第一次印刷 印张：7 1/8

字数：130 000

POD定价： 48.00元

(如有印装质量问题，我社负责调换)

前　言

语义 Web 技术使互联网用户可以更好地获取多种信息和相关服务，通过为互联网信息增加语义将使“更多的信息变得更有用”. 语义 Web 的远景和目标是实现数据之网(Web of Data)，数据以多种方式进行描述并以相应的语义链接为基础构成上下文，而本体(Ontology)和知识库(Knowledge Base, KB)的构建工作则被视为语义网发展的前提条件和基石.

关联数据(Linked Open Data, LOD)项目作为语义网发展的重要组成部分，其目标是将 Web 上已经发布的语义数据集之间进行最大限度的关联，以使得各自孤立的语义知识点可以互相关联进而最终形成大规模知识网络，从而使得知识共享和语义互操作成为可能，同时，如果将分散的数据源链接起来形成一个互联的数据之网，则形成的知识网络就会具有更高的价值，由此便可催生不同领域的网络新应用出现. 自从万维网之父 Tim Berners-Lee 提出关联数据的概念至今，相关研究已经越来越受到学者的关注. 本书重点关注关联数据构建的关键技术研究，包括以下三个研究点.

(1) 领域本体的自动化构建.

在进行大规模领域本体的构建时，基于手工方式的构建模式效率较低并且可行性较差. 为解决大规模领域本体库的自动化构建问题，利用中文百科与政务叙词表都具有跨领域覆盖的特点，提出一种领域叙词表与网络百科知识库相融合的两阶段领域本体自动化构建方案.

(2) 面向中文网络百科的知识发现.

从网络百科中自动获取海量知识已经被越来越多的学者所关注. 目前的研究主要关注于从英文网络百科数据源进行海量知识的抽取，而对中文或其他语言描述的数据源进行知识抽取的研究非常少. 因此，为解决中文大规模知识库的自动化构建问题，提出一种基于中文网络

百科架构的大规模知识库自动构建方案.

(3) 框架级(Schema-Ievel)的中文大规模关联数据系统研究.

目前关联数据的研究工作主要集中在实例级别上展开,而对于框架级别的关联数据构建则易被忽视. 同时,本体映射被视为框架级别关联数据构建的典型场景. 特别地,中文知识是网络开放知识库的重要组成部分,但现有的中文本体映射系统在面对大规模本体映射任务时,显得效率较低且可用性不高,目前仍缺乏针对中文大规模本体映射的相关系统. 为解决框架级别上的中文大规模关联数据构建问题,提出一种基于同义词词林的大规模中文关联数据构建模型.

本书主要包括以下研究内容和创新点.

(1) 提出一种领域叙词表与网络百科知识库相融合的两阶段领域本体自动化构建方法. 在第一阶段,进行叙词表至本体的粗映射,形成领域粗糙本体. 在第二阶段,将网络开放百科中的结构化知识与粗糙本体进行自动融合、自适应调整和扩充,进而形成含有丰富语义信息的、良构的领域本体库. 基于提出的两阶段方法,以中国电子政务领域为例,自动化地构建一个大规模中文本体框架,进而验证该方法的可行性和有效性.

(2) 提出一种面向中文网络百科非结构化信息的知识获取方法. 大规模开放域知识库的构建方案由以下两个步骤构成. 首先,对知识三元组中的主语和宾语之间的语义关系进行自扩展学习,即新词发现. 其次,基于条件随机场和支持向量机协同分类器,对标注出的属性和属性值实体之间的语义关系进行预测. 基于该方案,自动化地构建大规模开放域中文知识库.

(3) 提出一种新的面向中文大规模本体映射模型的总体框架. 随着本体规模的扩大,如何保证本体映射的效率就成为亟待解决的问题. 本研究提出的本体映射总体框架由三大功能模块组成,分别是本体概念初始相似度计算、本体压缩和确定性映射. 首先,采用基于编辑距离方法来计算待映射本体之间的概念初始关联度. 其次,基于概念初始相似度对待映射本体的规模进行压缩. 最后,根据中文概念特有的语义特征,提出一种基于同义词词林(扩展版)的中文本体概念等价关系确定

性映射策略.

(4) 提出一种对大规模本体映射规模进行压缩约简的新方法. 传统的本体映射系统和方法往往只注重映射结果,而忽视映射效率. 在面对大规模本体映射任务时,传统方法显得实用性不强. 本研究在对中文大规模本体进行等价关系的确定性映射前,提出将时间复杂度控制在可接受的范围内.

本书得以顺利出版要感谢科学出版社的大力支持,感谢首都经济贸易大学、北京工业大学及信息学院有关领导的指导和帮助;感谢北京工业大学邸瑞华教授和首都经济贸易大学计算机系徐天晟老师的辛勤付出. 同时本书的出版得到了首都经济贸易大学重点项目(No. 2014XJG022)、北京市哲学社会科学项目(No. 13SHB015)、北京市属高等学校青年拔尖人才培育计划、首都经济贸易大学 2015 年度科研基金(No. 00791554410264)资助以及 2015 北京市教委科研水平提高经费的资助,在此表示衷心的感谢.

本书在编写过程中参考了相关领域的著作、文献,在此向有关作者致以谢忱.

由于编者的知识、时间及水平有限,书中疏漏之处在所难免,衷心希望专家、学者及广大读者给予批评指正.

王　汀　徐天晟

2015 年 4 月

于首都经济贸易大学

目　　录

第1章 绪 论

知识处理能力标志着人类文明发展的程度. 计算机和互联网的发展使人类知识处理的能力显著提高,语义网(Semantic Web)的出现则使这种能力趋向新的高峰. 在语义网背景下,知识被表示为基于资源描述框架(Resource Description Framework, RDF)的语义数据,支持机器阅读和理解,支持自动推理,从而使计算机能够理解人类知识,处理数据的方式也更加智能化[1].

目前,语义 Web 正在蓬勃发展,许多知识库都被以语义开放数据集的形式发布到 Web 上,同时在数据集内部和不同领域的语义数据集之间建立起了丰富的关联,形成一个巨大的人类知识库,并且其规模正在急速增加. 根据关联数据项目(Linked Open Data, LOD)给出的统计数据显示,截止到 2011 年 9 月,链接并发布到 LOD 上的语义数据总量已经超过 300 亿条三元组(Triples)[2].

自然语言处理、机器学习等数据挖掘和知识抽取技术极大地促进了语义 Web 的发展,使得我们可以充分获取海量数据源中所蕴藏的知识,从而为高效地构建大规模知识库提供了可能. 关联数据网的构建和形成,不仅可以极大地发掘语义数据集之间所蕴涵的隐式知识和潜在价值,同时还可以以此为基础而催生与其相关的新应用的诞生,从而推进语义 Web 向着真正的数据之网(Web of Data)的目标迈进.

1.1 研究背景与意义

随着信息技术的快速发展,互联网成为人类有史以来构建的规模最大的信息库,也已经成为人们获取信息的主要渠道之一,对人类社会生活的各个方面都产生了深远的影响. 而随着 Web 上的数据和内容快速增长,人们准确、快速、全面获取到信息也变得越来越困难. 搜索引擎

的出现和发展，在一定程度上缓解了这个难题. 目前，基于关键词匹配的搜索引擎已经成为人们使用 Web 的主要工具. 毫无疑问，这些搜索引擎为互联网今天的成就做出了巨大的贡献，但是也存在一系列严重的问题，具体如下.

(1) 高匹配、低精度：在大多数情况下，搜索引擎即使能搜索到主要的相关页面，但大量不相关或者低相关的页面往往会与主要相关页面混在一起，大大影响了检索的效果.

(2) 检索结果对词汇高度敏感，导致搜索结果的低匹配或无匹配：搜索引擎只返回包含关键词的页面，而那些包含与关键词近似或近义的术语的文档则被忽略，这导致用户的搜索漏掉一些重要的相关页面或没有任何搜索结果.

(3) 检索结果是单一的网页：用户必须自己浏览搜索到的文档，花费大量时间和精力从中提取所需的信息. 此外，当用户需要多方面相关信息时，必须给出多个关键词进行多次查询来收集相关的页面，然后自己提取页面中的信息并进行组织和分析.

造成这些问题的主要原因是目前 Web 上的内容主要是提供给人来浏览和理解的，尽管网页中含有一些链接和特殊的信息使得计算机能定位相应的页面并以特定的方式显示文档，但网页中没有提供任何信息帮助机器理解网页的内容，没有采用形式化的表示方法，缺乏明确的语义信息，计算机只能将 Web 内容作为二进制数据进行处理，而无法理解网页内容的语义，无法实现网页内容的自动处理.

解决这些问题的一个途径就是用一种更容易被机器处理的表示方法来描述 Web 内容，并采用智能技术来利用这种表示方法所描述的数据，从而使计算机能够理解和自动处理网页内容，准确地从海量网页中查找出所需要的内容. 这种方法就是 Tim Berners-Lee 于 1998 年提出的语义网(Semantic Web). 语义 Web 技术通过对现有 Web 增加语义支持以使得互联网信息能够在一定程度上被机器所理解，从而使高效的信息共享、信息集成和机器智能协同成为可能[3].

分布在 Web 上的不同数据源是由不同的组织、单位在不同的时间创建并且采用了不同的数据存储格式、数据库，这就使得数据的重用十

分困难甚至是无法重用并带来维护成本巨大的问题. 语义知识表示语言 RDF① 与网络本体语言 OWL(Web Ontology Language)②通过统一的标准,可以解决数据格式不统一、重用困难的难题. 信息源的语义异构问题是目前 Web 信息集成的主要瓶颈之一,本体(Ontology)作为"共享概念模型明确的形式化规范说明"[4],因此特别适合于解决语义异构问题. 语义 Web 技术采用统一的知识表示方法、本体、语义标注、专用的形式化描述方法、唯一的资源标识机制对数据进行描述,可以将异构数据源的数据进行良好的整合.

采用本体技术可以实现信息的语义检索、系统的语义集成及互操作等功能,国内外专家已经在这个方面做了许多工作并取得了一定的成绩. 但是由于互联网的分布性,不可能要求所有的知识源都使用统一的本体模式和命名模式,由此造成的数据异构性会成为限制计算机自动处理数据能力的瓶颈,进而影响互联网上数据和知识共享的效率.

与此同时,由不同组合或个体发布的开放数据可能是不完备的,并且与其他本体库或者知识数据集中的数据可能存在某种形式的关联,这种关联可能使数据的重用、分析产生更大的价值. 关联数据(Linked Data)[5]技术则非常适合将不同个体、系统发布的知识源进行良好的集成,甚至是可以与一些公共数据集(如交通数据、地理数据,乃至国外相关部门的相关数据)进行集成.

由公众、企业以及其他组织在网络上发布的开放数据所提供的数据格式目前尚缺乏统一的规范和标准,但是这部分数据中却蕴涵着海量的知识,其中比较典型的就是网络开放百科系统[6-8]. 对于 Web 上的海量知识发现和获取、大规模知识库的自动构建,以及不同数据源之间如何进行良好的数据关联与集成也是语义 Web 发展中亟待解决的关键问题.

网络上的开放数据源之间需要通过定义标准的本体并通过本体对不同的资源进行关联才能形成知识网络. 2006 年,Web 的发明人 Tim Berners-Lee 提出了一种 URI 规范,使得人们可以通过 HTTP URI 机

① http://www.w3c.org/TR/rdf-primer/

② http://www.w3.org/TR/2004/REC-owl-features-20040210/

制，直接获得数字资源(Thing)[5]，关联数据(Linked Data)的概念由此产生[9]，其基本特征包括：

(1) 使用 URI 作为任何事物的标识名称；

(2) 使用 HTTP URI 使任何人都可以参引(Dereference)全局唯一名称；

(3) 当有人访问名称时，以 RDF 形式提供有用的信息；

(4) 尽可能提供相关链接.

关联数据的核心是构建数据集之间的“链接”(Interlinking)，即支持结构化数据的任意关联，并使得“链接”变为“链”(Chains)而最终形成大规模知识网络；而关联关系则携带大量而有价值的语义，从而支持基于链接关系的检索和用户浏览. 通过 RDF 链接实现不同数据集之间的关联，可以体现关联数据技术的巨大应用价值. RDF 链接可以使用户通过关联数据浏览器从一个数据源中的数据游历到另一个数据源，从而获得更多、更全面的信息；RDF 链接还可以供搜索引擎和网络爬虫追踪，爬行下来的数据可以进行更复杂的查询和检索.

随着语义网的发展，大规模的中文语言描述的知识资源也越来越多的被构建和共享出来. 关联关系的构建是关联数据实现、发布和扩展的前提. 互联网上和传统数据系统中的结构化数据总量非常庞大，因此，研究语义关联关系的发现具有非常重要的意义[10]. 但是，目前发布在 Web 上的开放中文大规模本体仍然较少，且存在较大的异构性. 中文海量知识库的自动构建仍处于起步阶段[11]，而良构和完备的领域知识库又是关联数据构建的基础和前提，相应地，目前更缺乏面向中文的大规模关联数据自动化构建的成熟系统和框架.

总之，关联数据具有良好的前景，但也正面临着众多难题和挑战. 只有这些挑战被克服，关联数据才能在充分发挥万维网功能并向语义网进军的道路上迈出革命性的一步.

1.2 国内外研究现状

1.2.1 语义 Web 相关研究进展

互联网技术发展至今，其发布的内容仍然只能为人所理解和阅读

且网页上的链接和其他信息只能支持浏览器对其所对应的 Web 页面的显示、解析以使机器可以查找到. 但是由于缺乏形式化的语义描述方式和明确的语义信息,Web 页面上的信息无法帮助计算机理解其所表示的网页内容,这就导致计算机只能提供信息的发布和显示功能,而无法理解信息的含义并进行智能处理.

为了解决该问题,Tim Berners-Lee 爵士在 1998 年首次提出语义网(Semantic Web)框架[12],该框架推荐的核心标准即是本体(Ontology). 通过本体可以对传统 Web 增加语义支撑,其目标是使机器在一定程度上理解信息的语义,从而使得高效信息共享和机器智能协同成为可能. 语义网被视为下一代 Web 技术,其通过对传统 Web 信息的语义化来对现有 Web 框架进行扩展和语义增强,从而使机器具备可以与人进行智能协作的能力[12-14].

语义网框架的提出是整个互联网技术的发展和继承的结果. 自从 Tim Berners-Lee 在 1990 年将超文本技术与互联网进行融合并首次提出了万维网(World Wide Web, WWW)框架以来,面向互联网的知识表示和信息交互技术被越来越多的学者所研究. 一般认为,互联网的发展经历了文件之网(Web of Files)、社会网络(Web of People)以及数据之网(Web of Data)[15-17]三代. 其中,文件之网可被视为 Web 1.0,其特点是网站所有人基于 HTML 标准来发布信息以供普通互联网用户浏览,该模式的缺点是其缺乏与用户的互操作且用户的主动参与较少;近年来,社会之网已经有了长足的发展,其框架是以用户为核心的,并突出信息共享与无中心化的特点,这就使得普通互联网用户逐步成为信息发布与共享的参与者和核心力量,同时也使万维网上的数据量呈现出极度增长的趋势;而数据之网即是未来的 Web 3.0,但是目前该技术尚未成熟,相关的应用及其适应场景仍处于研究和探索的阶段,这是因为信息爆炸使得信息消费能力不足的问题凸显,而为使机器可以自动理解和处理互联网上提供的海量信息,需要研究面向海量 Web 信息的知识发现、表示以及共享技术.

可以看到,Web 1.0 技术早已成熟和普及. 近年来,随着 Web 2.0 技术的逐步实施,其典型应用也已成为万维网的主流并为广大用户所

采用和接受. 而 Web 3.0 技术可被视为未来互联网的发展方向,相关的研究也正在如火如荼地进行中.

随着整个互联网技术的发展和演进,Web 应用系统的模式和架构也在发生改变. 例如,第一代 Web 页面显示规范即为超文本传输协议(Hyper Text Markup Language, HTML),其主要为用户浏览器所解析和显示的 Web 页面提供了一整套标签和原语,从而使得网络资源特别是多媒体信息的显示得以规范化. 但是 HTML 所蕴涵的语义信息极为匮乏,因此其只能用于机器来解析和显示网页内容的格式,但是无法实现网络上跨平台式的异构数据交互和智能推理,因此 HTML 可以被认为是机器阅读和理解的初级阶段,而距离语义网愿景的实现仍有很长的路要走. HTML 规范使得 Web 网页上的数据和显示被混杂在一起,并且可扩展性较差,这就使得对于海量 Web 信息的知识抽取、共享和利用以及在此基础上所要进行的智能应用的开发变得极为困难.

因此,新的 XML 标准便应运而生. XML 具备极好的自描述能力,用户可以根据自身应用和需求的特点来对 XML 的标签库进行自主定义和扩充,并且 XML 较传统的 HTML 标准对数据的传输和显示进行了剥离,这就使其具备一定的语法约束力和良构性,并向着 Web 知识表示的目标又迈进了一步,因此 XML 技术也被视为语义网架构的重要基础[3]. XML 标准的采用极大地提升了网络信息和异构平台数据交互的效率,但是单纯的 XML 文档只能够携带和交互数据,其标签不带有机器可读的语义信息,那么万维网数据仍无法被计算机理解[18]. 因此在 XML 的基础上,语义网架构的语义层便出现了支持机器语义处理的技术和标准.

语义层上所包含的相关标准对整个语义 Web 来讲至关重要,其中最核心的标准为资源描述框架(Resource Description Framework, RDF). 通过 RDF 标准使得语义网具备了对于各种异构网络资源的标准化描述模式和知识表示方式,它是实现语义网愿景的基石. 该标准中包含一系列用以描述资源的术语,如将资源通过 RDF/RDFs 词汇表进行刻画,然后通过 RDF/XML 或者 N-Triple 格式进行表达,它们都是 W3C 所推荐的语义 Web 相关重要规范[20-21]. RDF 标准采用图模式

(Graph)的方式刻画三元组知识,知识即是一组事实性陈述(Set of Statement),我们将其简称为 RDF 图. 其中,RDF 图中的各个节点(Node)也被称为资源(Resource),每个资源通过统一资源描述符(Universal Resource Identifier, URI)进行刻画[20]. 而每组事实性陈述(Statement)可以用三元组(Triple)的形式来表达,即采用〈主语,谓语,宾语〉的形式. 其中,谓词(Predicate)也可被称为属性(Property)或者关系(Relationship),而宾语即是属性值(Property Value). 对于不同的上下文环境,所采用的术语表达可以有所不同. 具体地,若上下文语境主要涉及语义 Web 的理论方面,即描述逻辑,则采用谓语或者谓词这样的表达方式进行叙述,相应的三元组则被更多地称为事实性陈述. 当上下文语境主要对语义网的数据模型进行阐述时,则将知识三元组中的谓语更多地称为属性或者关系,而三元组中的主语和宾语则代表 RDF 图模型中的节点,属性就是连接节点之间的边,更用来刻画主语和宾语之间的语义关系[22]. 海量的知识三元组通过节点之间的语义关联而形成表示人类知识的复杂网络[23],本质上就是通过 RDF 图来刻画网络资源之间的语义关系进而形成机器可理解的数据之网.

通过 RDF 规范可以将 Web 原始数据统一地语义化为网络可访问的三元组格式,同时为了使得机器可以更好地理解每条知识三元组所蕴涵的语义信息,还必须要用更高级的知识表示模型,本体作为"共享概念模型的明确地形式化的规范说明"[24],是由 Studer 于 1998 提出并已被普遍接受. 因此,采用本体规范建立人与机器之间的共享概念模型至关重要. 在语义 Web 范畴中的本体框架可以视为被机器所理解和阅读的领域叙词表,不同于传统的纸质版叙词表,本体还必须满足 URI 可访问、具备属性约束、含有公理和规则、实例填充以及统一的 RDF 表示等基本特征. 本体的构建过程既是对某个领域或者开放域的知识进行规范化和梳理的过程,又是基于某种特定的知识表示语言,如 RDF/RDFs 或者 OWL 等进行数据的语义化并形成领域知识库的过程. 其中,OWL(Ontology Web Language, OWL)语言目前的应用最为广泛且表达和推理能力最强[25-26]. 而 SPARQL(Simple Protocol and RDF Query Language)语言则是 W3C 推荐的语义数据查询规范,支持对知

识三元组的本地查询和基于 SPARQL 协议的远程查询[26—27]. SPARQL 查询通过图匹配的方式进行语义查询和简单推理并将查找到的三元组返回给用户. 语义网的理论基础是描述逻辑. 计算机可以通过定义的规则、公理来进行推理和演算从而使得应用程序更加智能化.

为了实现语义网的愿景,我们不能仅仅是将网络实体资源通过 RDF 三元组或者基于 OWL 规范,以本体的形式表达并发布到万维网上. 真正意义上的数据之网应当是能够将不同知识源链接起来的,通过数据的链接使得数据之间的访问更加灵活,用户可以在不同的数据源之间进行游历,我们将这种知识获取和数据组织方式称为关联数据(Linked Open Data, LOD). 第一代 Web 是面向 HTML 文本的、不携带语义数据的超链接 Web. 而我们将要面对和研究的下一代 Web,就是通过 URI 来描述网络上任何实体资源或者概念,并通过超链接以使其均可被访问到,最终基于 RDF 标准实现数据源之间语义关联的数据之网.

1.2.2 本体自动化构建的研究进展

本体(Ontology)是构建语义 Web 的核心和关键,同时也是大规模关联数据网形成的基础和前提. 本体的概念最早出现在哲学领域,是共享概念模型的明确的、形式化规范说明[28]. 本体作为一种在语义层和知识层上描述信息的概念建模工具,在知识工程、数字图书馆、信息检索与集成等领域得到广泛应用[22].

从知识共享的角度看,本体可以看成是对特定领域概念体系的明确化、规范化的说明,是对客观存在的领域概念和关系的形式化描述. 它将隐含在领域专家头脑中的知识体系以计算机可理解的方式表达出来,从而减少了对领域概念和逻辑关系的误解.

数据源的语义异构问题是目前 Web 数据集成和知识共享的主要瓶颈之一,本体作为"共享概念模型明确的形式化规范说明",在解决语义异构问题方面具有与生俱来的优点. 基于本体的信息集成方法目前已经成为 Web 信息集成的主要方法. 并且在 Web 环境中,海量的领域知识信息是以 HTML 格式表达和发布的,同时,在目前的中文网络百科