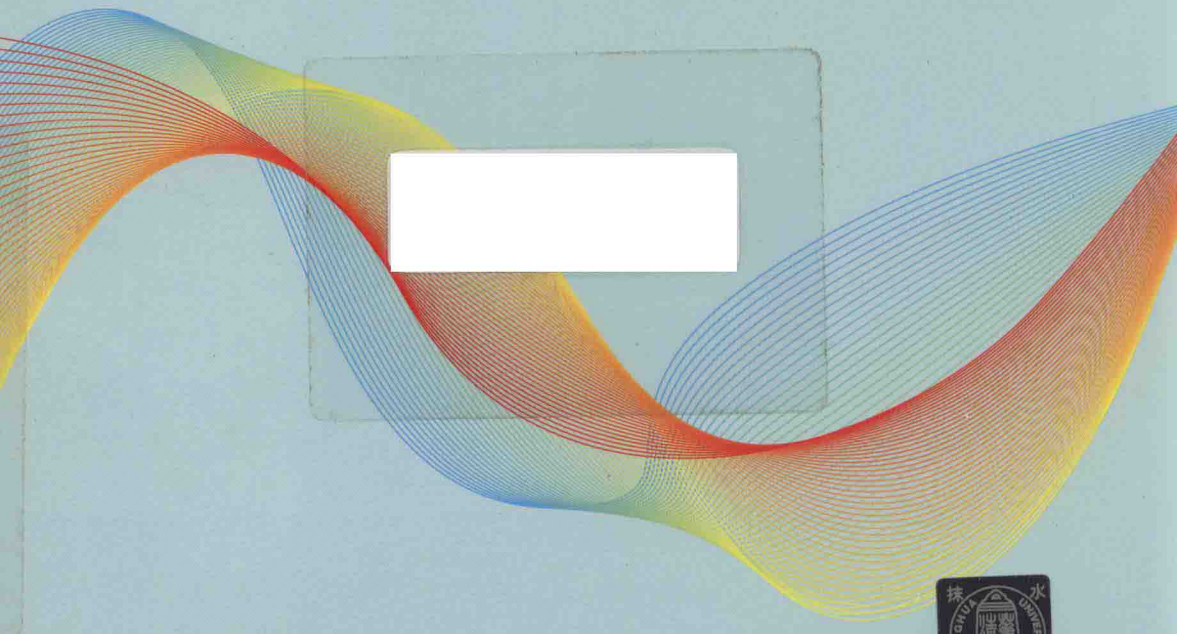
Ontology Model and Semantic Web Knowledge Discovery

本体建模与语义Web知识发现

阎红灿 著



清华大学出版社

Ontology Model and Semantic Web Knowledge Discovery

本体建模与语义Web知识发现

阎红灿 著

清华大学出版社
北京

内 容 简 介

本书是作者多年来科研工作的梳理和升华，内容包括：XML 文档管理和分类技术、知识资源描述语言和发布技术、本体建模和知识推理技术、基于知识库的知识发现关键技术和模型框架，同时给出了基于知识库的知识发现的典型应用。

本专著阐述明晰，内容新颖，力求深入浅出，可以用作高等院校有关专业的研究生和高年级本科生的信息检索、知识发现等课程教材，也可供从事数据挖掘、语义检索和知识管理的科技人员阅读参考。

图书在版编目(CIP)数据

本体建模与语义 Web 知识发现/阎红灿著. —北京：清华大学出版社，2015
ISBN 978-7-302-41627-2

Ⅰ. ①本…　Ⅱ. ①阎…　Ⅲ. ①知识工程　Ⅳ. ①TP182

中国版本图书馆 CIP 数据核字(2015)第 228380 号

责任编辑：付弘宇　柴文强
封面设计：何凤霞
责任校对：梁　毅
责任印制：李红英

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社总机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课　件　下　载：http://www.tup.com.cn，010-62795954
印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　　**印　　张**：18　　**字　　数**：287 千字
版　　次：2015 年 12 月第 1 版　　**印　　次**：2015 年 12 月第1 次印刷
印　　数：1～1000
定　　价：69.00 元

产品编号：063882-01

前　言

知识发现是从数据集中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的非平凡过程。随着语义 Web 的发展，基于知识库的知识发现成为知识发现领域的一个研究目标。语义 Web 下经过 XML 标注的自然语言文本、资源描述框架 RDF 和本体建模为知识发现提供了有力工具，数据资源的描述和存储、索引，实现语义检索的知识推理，都是语义 Web 下基于知识库的知识发现的关键技术支撑。

专著全面而又系统地介绍了 XML 文档分类技术、RDF 知识资源描述语言和本体推理等知识发现的方法和技术，反映了当前本体建模与知识发现研究的最新成果和进展。全书共分 9 章。第 1 章是引言绪论，概述语义网络中智能检索的关键技术和基于知识库的知识发现的重要概念；第 2、3 章讨论 XML 数据库的存储策略，重点论述基于关系存储策略的索引、维护和更新技术，第 4 章论述 XML 网页的频繁模式挖掘和分类技术；第 5 章介绍一种基于 N 层向量空间模型的全文检索方法；第 6 章系统阐述知识表示和本体建模技术；第 7 章综述数据资源的描述和关联数据发布，阐述基于关联数据的知识发现模型；第 8 章讨论本体推理技术及粗逻辑在本体推理中的应用；第 9 章论述基于本体知识库的知识发现框架和关键技术，给出基于知识发现的案例推理应用模型。

书中内容新颖，高度总结和梳理了作者多年的科研成果，取材国内外最新资料，反映了当前该领域的研究水平。论述力求概念清晰，表达准确，算法丰富，突出理论联系实际，富有启发性。

为了丰富专著的内容，书中引用了参考文献中的学术资料，对这些研究成果和做出辛勤劳动的作者表示真诚感谢；专著出版得到了国家自然科学基金、河北省重点学科、华北理工大学以及清华大学出版社的大力支持，华

北理工大学理学院应用数学专业研究生王会芳、张奉等同学在本书的编写和校稿过程中做了一定工作，在此一并表示诚挚的谢意。

因为时间仓促，加上作者的学识浅陋，书中的不足与错误，敬请同行给予批评指正。

阎红灿

2015 年 6 月

目　录

第 1 章　基于语义 Web 的智能检索和知识发现

Web 技术的出现使人类的生存空间得到极大扩展，并逐渐成为人们获取、传播和交换信息的重要途径。随着 Internet 的飞速发展和广泛应用，其缺陷也逐渐暴露出来，如搜索引擎智能程度低，搜索出来的结果往往不是用户真正需要的，检索结果是单一的网页等等。互联网的创始人 Tim Berners-Lee 于 2000 年 12 月 18 日在 XML2000 会议上正式提出语义 Web(Semantic Web)。语义 Web 的目标是使 Web 上的信息具有计算机可理解的语义，满足智能软件代理对万维网上异构和分布式信息的有效访问和搜索。语义 Web 不是另外一个 Web，它是现有 Web 的延伸，其中信息被赋予了良定义[1]。语义 Web 将在更加微小的信息之间建立直接的连接，例如一条街道的地址与一份地图等，用户可以将两个毫不相干的东西连接在一起，比如说银行报账单和日历。用户可以将银行报账单拖到日历上，也可以将日历拖到银行报账单上，这样就可以知道何时进行支付。语义 Web 将呈现给人们一个所有数据“无缝”式连接的网络。在语义 Web 技术破土而出之后，人们对 Facebook 和 MySpace 等社交网站的“痴迷”终将被“无所不连”的网络所取代。

1.1　语义 Web 下的信息检索

语义网的实现需要 3 大关键技术支持：XML、RDF 和 Ontology。Tim Berners-Lee 提出的语义网层次结构[2]如图 1-1 所示。

1) Unicode 和 URI 层：Unicode 和 URI 层是整个语义万维网的基础，

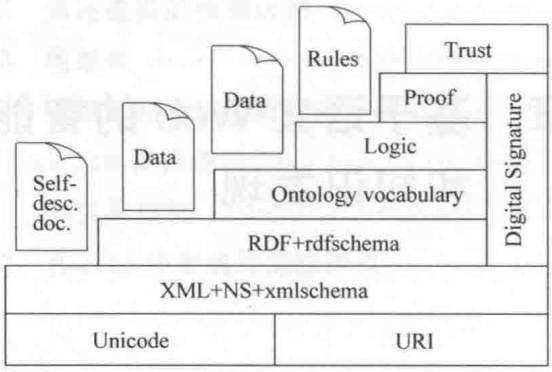


图 1-1 语义 Web 的层次模型

其中 Unicode(统一资源编码)处理资源的编码,保证使用的是国际通用字符集,实现网上信息的统一编码。URI 是 URL(Universal Resource Locator)的超集,URI 支持语义网上的对象和资源的精细标识,从而使精确信息检索成为可能。这一层是语义万维网的基石。

2) XML + Name Space + XML Schema 层

XML 和命名空间层用于表示数据的内容和结构,XML 层具有命名空间(Name Space)和 XML 模式(XML Schema)定义,通过 XML 标记语言可以将网上资源信息的结构、内容与数据的表现形式进行分离。

3) RDF + RDF Schema 层

RDF 和 RDFS 层用于描述万维网上的资源及其类型,为网上资源描述提供通用框架和实现数据集成的元数据解决方案。

4) 本体层

本体层用于描述各种资源之间的联系,采用 OWL 表示。本体(Ontology)揭示了资源以及资源之间复杂和丰富的语义信息,将信息结构和内容分离,对信息做完全形式化的描述,使 Web 信息具有计算机可理解的语义。

5) 逻辑层

逻辑层主要用于提供公理和推理规则,为智能推理提供基础。逻辑层可以进一步增强本体语言的表达能力,并允许创作特定领域和应用的描述性知识。

6）证明层

证明层设计实际的演绎过程以及利用 Web 语言表示证据，对证据进行验证等。证明注重提供认证机制，证明层执行逻辑层的规则，并结合信任层的应用机制来评判是否能够信任给定的证明。

7）信任层

信任层提供信任机制，保证用户 Agent 在 Web 上提供个性化服务，以及彼此之间安全可靠的交互，基于可信 Agent 和其他认证机构，通过使用数字签名和其他知识才能构建信任层。当 Agent 的操作时安全的，而且用户信任 Agent 的操作及其提供的服务时，语义 Web 才能充分发挥其价值。

在 Tim Berners-Lee 的语义网模型中，作为语法层的 XML 层，作为数据层的 RDF 层和作为语义层的 Ontology 层是语义 Web 的关键层。用于 Web 信息的语义，也是现在语义 Web 研究的热点所在。

语义 Web 下的信息检索，不再是一味的基于关键词的匹配，而是基于知识体系或概念网络的智能检索，即语义检索。比如用户查询“计算机”，与“电脑”相关的信息也能检索出来，甚至可以进一步缩小查询范围至“微机”、“服务器”或扩大查询至“信息技术”或查询相关的“电子技术”、“软件”、“计算机应用”等范畴。另外，智能检索还包括歧义信息和检索处理，如“苹果”，究竟是指水果还是电脑品牌，“华人”与“中华人民共和国”的区分，将通过歧义知识描述库、全文索引、用户检索上下文分析以及用户相关性反馈等技术结合处理，高效、准确地反馈给用户最需要的信息。

在信息检索分布化和网络化的趋势下，信息检索系统的开放性和集成性要求越来越高，需要能够检索和整合不同来源和结构的信息，这是异构信息检索技术发展的基点，包括支持各种格式化文件，如 TEXT、HTML、XML、RTF、MS Office、PDF、PS2/PS、MARC、ISO 2709 等处理和检索；支持多语种信息的检索；支持结构化数据、半结构化数据及非结构化数据的统一处理和关系数据库检索的无缝集成以及其他开放检索接口的集成等。所谓“全息检索”的概念就是支持一切格式和方式的检索，从实践来讲，发展到异构信息整合检索的层面，基于自然语言理解的人机交互以及多媒体信息

检索整合等方面尚有待进一步突破。

智能信息检索系统应具有如下的功能：

(1) 能理解自然语言，允许用自然语言提出各种询问；

(2) 具有推理能力，能根据存储的事实，演绎出所需的答案；

(3) 系统具有一定常识性知识，以补充学科范围的专业知识。系统根据这些常识，能演绎出更泛化的一些答案来。

语义 Web 为实现用户语义的智能检索提供技术平台和技术支撑。

1.2 知识和知识发现

知识[3]这一概念有三种比较有代表性的定义：

(1) Feigenbaum：知识是经过消减、塑造、解释、选择和转换的消息；

(2) Bernstein：知识是由特定领域的描述、关系和过程组成；

(3) Heyes-Roth：知识＝事实＋信念＋启发式。知识常常是模糊、不确定或不完全的，而且知识还处在不断地动态变化过程中。

1.2.1 知识描述

通常采用 Heyes-Roth 提出的知识的三维空间来描述任何知识，即知识的范围、知识的目的和知识的有效性。范围由具体到一般，目的从说明到指定，有效性由确定到不确定。知识的三维空间描述如图 1-2 所示。

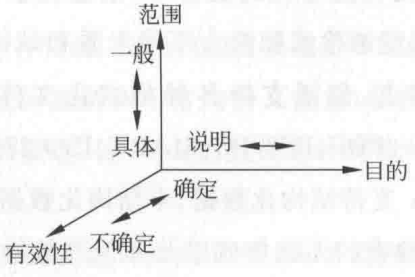


图 1-2 知识的三维空间描述

大量的数据经过加工后才会有价值，经过分析处理的数据形成了信息，信息的作用后有时间和范围的限制。为了使信息在较长的时间内有效，必须进行一系列的内部处理，这个过程叫综合，综合后的信息组成了知识。

图 1-3　数据、信息和知识的层次关系

从计算机科学的观点来看，知识是信息综合处理的结果。在综合过程中，信息传递相互比较，结合成有意义的链接。数据、信息和知识具有层次关系，它们的层次关系如图 1-3 所示。

1.2.2　领域知识和知识库

领域知识主要应用在基于知识的专家系统和自然语言理解以及有关概念的约束的集合。知识工程对领域知识进行了三方面的描述：

(1) 领域知识是一个概念模型，这个概念模型包括概念和概念之间的关系；

(2) 领域知识是概念和概念之间的约束；

(3) 领域知识是陈述如何推导计算出新概念和新概念之间的关系的规则。

领域知识的两个基本概念：

(1) 领域特征概念：这是领域知识的概念化，是各种相关领域内的重要概念的语义描述；

(2) 领域特征属性：这是指某一领域内的概念所具有的特点，领域特征概念可以是词，也可以根据需要扩展成短语甚至词串。

知识库是针对某一领域问题求解的需要，采用某种知识表示方式在计算机存储器中存储、组织、管理和使用的互相联系的知识的集合。

领域知识是指在某一专门领域中重要问题或概念以及概念之间的相互关系的集合。领域知识库这一术语源于人工智能领域。在人工智能领域

中，领域知识主要应用于知识的专家系统和自然语言理解的系统。

1.2.3 基于知识库的知识发现

随着网络技术和通信技术的发展，数据、信息的产生和收集能力已经迅速提高。而更能为人们提供帮助的，潜在这些数据中的信息和知识却相对缺乏。这种数据爆炸而知识缺乏的情况激起人们对新技术和自动工具的需求，数据挖掘技术就是新技术之一。如何从长期积累的、大量的信息中找到对我们有用的知识已成为一个亟待解决的问题，于是知识发现应运而生。多研究者从不同的角度给出了有关知识发现的定义，目前较一致认同的描述性定义是 Fayyad 等人给出的：知识发现是从数据集中识别出有效的、新颖的、潜在有用的以及最终可理解的模式的非平凡过程。

传统的知识发现是指基于数据库的知识发现（KDD：Knowledge Discovery in Databases），是从大量的、不完整的、有噪声的、模糊的和随机的数据中，提取隐含在其中的、人们事先不知道的，但又是可信的、潜在的和有价值的信息和知识的过程。知识发现将信息变为知识，从数据矿山中找到蕴藏的知识金块，将为知识创新和知识经济的发展做出贡献。

知识发现与数据挖掘的关系密不可分。数据挖掘（Data Mining），就是从海量的数据中抽取出隐含的、未知的、具有潜在使用价值信息的过程[4-5]。由于数据挖掘是 KDD 过程中最为关键步骤，在实际应用中两个概念往往不加以区分。一般认为广义的数据挖掘又称数据库中的知识发现，简称知识发现（KDD）；狭义的数据挖掘是一个利用各种分析工具在海量数据中发现模型和数据关系之间关系的过程，是知识发现过程的一个步骤，一个完整的知识发现过程如图 1-4 所示。从图中可见，数据挖掘是知识发现过程中一个发现模式的子过程，并且是最核心的过程。

完成从大型源数据中发现有价值知识的过程可以简单概括为：

首先从数据源中抽取出感兴趣的数据，并把它组织成适合挖掘的数据组织形式；然后调用相应的算法生成所需要的知识；最后对生成的知识模式进行评估，并把有价值的知识集成到企业的智能系统中。