数据库支持的模糊OWL本体管理

SHUJUKU ZHICHI DE MOHU OWL BENTI GUANLI

吕艳辉 著

国防工业出版社
National Defense Industry Press

数据库支持的模糊 OWL 本体管理

吕艳辉　著

·北京·

图书在版编目(CIP)数据

数据库支持的模糊 OWL 本体管理 / 吕艳辉著. —北京:国防工业出版社,2011.5

ISBN 978-7-118-07451-2

Ⅰ.① 数... Ⅱ. ① 吕... Ⅲ. ① 数据库管理系统—研究 Ⅳ. ①TP311.13

中国版本图书馆 CIP 数据核字(2011)第 075649 号

※

国防工业出版社出版发行

(北京市海淀区紫竹院南路 23 号 邮政编码 100048)

天利华印刷装订有限公司印刷

新华书店经售

*

开本 850×1168 1/32 印张 4¾ 字数 120 千字

2011 年 5 月第 1 版第 1 次印刷 印数 1—3000 册 定价 28.00 元

(本书如有印装错误,我社负责调换)

国防书店:(010)68428422 发行邮购:(010)68414474

发行传真:(010)68411535 发行业务:(010)68472764

前　言

语义 Web 是当前 Web 的扩展，它赋予 Web 资源信息机器可理解的语义，从而便于人和计算机之间的交互与协作。为了让机器能够理解 Web 资源信息并做推理，需要建立本体，并使用本体语言进行描述。语义 Web 本体语言的标准是 OWL，它建立在数据表示语言 RDF 和 RDFS（常合称为 RDF(S)）之上，它们一起构成了当今语义 Web 的描述语言基础。

由于很多语义 Web 应用需要处理大量的模糊知识，而现有本体不能直接用于模糊知识的表示和处理，因此，对本体进行模糊扩展以满足模糊知识管理的需要逐渐成为一个研究热点，这一点与数据库技术为表示和处理现实世界中的模糊数据而产生模糊数据库模型的情况相一致。

作为 Web 时代模糊信息表示和处理的两个重要技术方法，模糊数据库模型和模糊本体之间存在着密切的关联关系。一方面，从构建的角度，模糊数据库模型可以作为构建模糊本体的数据源，使模糊本体充分利用现有的模糊数据库模型中的信息；另一方面，从存储的角度，利用模糊关系模型亦即模糊关系数据库在模糊数据存储和处理等方面的优势，能够对语义 Web 上的模糊信息进行有效的管理。事实上，模糊本体构建和存储是模糊本体管理中的两个重要问题。目前，有关模糊本体存储以及利用结构化模糊数据进行模糊本体构建的研究成果还很少。

为有效表示和处理大量的模糊知识、实现模糊语义 Web 本体的管理，本书在对语义 Web 数据层语言 RDF(S) 及本体层语言 OWL 模糊扩展的基础上，展开数据库支持的模糊 OWL 本体管理的研究，目标在于形成一个有关模糊 OWL 本体从表示到构建、存

储的完整理论框架。本书第 1 章阐述了 OWL 本体模糊扩展以及数据库支持的模糊 OWL 本体构建与存储的研究背景和研究动机,分析了国内外相关工作的研究现状。第 2 章介绍了本书的背景知识,包括本体、本体的描述语言基础 RDF(S)与 OWL、描述逻辑,以及模糊集理论和现实应用中模糊信息的分类和表示方法。第 3 章研究了语义 Web 数据层语言 RDF(S)和本体层语言 OWL 的模糊扩展并给出了模糊 OWL 本体的形式化定义。第 4 章研究了利用模糊 EER 模型构建模糊 OWL 本体的方法。第 5 章研究了利用模糊关系数据库构建模糊 OWL 本体的方法。第 6 章研究了模糊 OWL 本体的数据库存储方法。第 7 章对本书所做的工作进行了总结,并对后续研究工作进行了展望。

本书在参考国内外有关文献的基础上,结合作者的科研成果,系统地研究了数据库支持的模糊 OWL 本体管理中的关键技术。本书内容深入浅出,全面地展示了国内外大量最新的科学研究内容和发展动向,具有一定的前瞻性和学术参考价值。

在本书的撰写过程中,得到了东北大学博士生导师马宗民教授点石成金的指导,值此书出版之际表示衷心的感谢。此外,感谢沈阳理工大学信息科学与工程学院的领导对作者研究工作的支持和鼓励。最后,感谢国防工业出版社有关工作人员的帮助。

由于作者水平所限,加之本书所涉及的内容仍处于不断的发展和变化之中,书中错误和不足之处在所难免,恳请专家、读者批评指正。

作　者

目　录

第1章 绪 论

1.1 研究背景

万维网(World Wide Web)的诞生从根本上改变了人类存储和交换信息的方式,并已影响到人类生活和生产活动的各个方面。同时,万维网的发展也使网络上的信息资源爆炸性地增长,但由于缺乏自动处理网络中海量信息的技术,用户越来越难以有效地检索这些信息。提高 Web 信息检索的质量包括两方面内容:一方面是如何在现有的资源上设计更好的检索技术;另一方面是如何为 Web 上的资源附加上计算机可以理解的内容,便于计算机更好地处理。针对后一种情况,万维网之父 Tim Berners-Lee 于 1998 年首次提出语义 Web(Semantic Web)的概念,并在 2000 年 12 月召开的 XML 2000 会议上发布了语义 Web 体系结构,进一步明确阐述语义 Web 的设想。

语义 Web 广泛吸取人工智能、信息论、哲学、逻辑和计算理论等学科的研究成果,并非是全新的 Web,而是对现有万维网的扩展。语义 Web 的目标是让网络上的信息能够被机器理解,从而实现网络信息的自动处理,以利于人机间的合作与交互,并在此基础上实现各种智能化的应用。

基于上述目标,语义 Web 的架构是一个功能逐层增强的层次化结构,如图 1 - 1 所示。第一层为统一字符编码 Unicode 和统一资源标识符 URI(Uniform Resource Identifier),为 Web 上定义字符和资源提供标准方法。第二层包括可扩展性标识语言 XML (Extensible Markup Language)、XML Schema 以及 XML 命名空间(Name Space,NS)。这两层给出了当前 Web 的基本要素,是语义

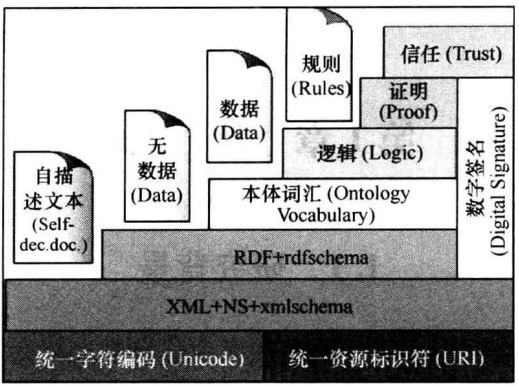

图 1－1　语义 Web 架构

Web 的语法基础，常称之为语义 Web 的文法层。

尽管 XML 规范了 Web 上的数据表示和数据交互，并已被工业界广泛接受，但其存在着公认的缺陷，即 XML 只能定义语法格式，而不能表达形式化语义，这样，不足以用来描述 Web 资源。因此，W3C(World Wide Web Consortium)于 2004 年 2 月发布了一种新的语言，即资源描述框架(Resource Description Framework，RDF)。RDF 用于描述网络上各种资源的信息，它基于 XML 的语法形式，RDF 语义(RDF Semantics)则是通过模型论(Model Theory)方法对 RDF 赋予形式化语义。但是这种说明性的语言没有提供机制来描述属性或属性与其他资源之间的关系，因此需要定义描述中使用的词汇，这就是 RDF 的词汇描述语言，即 RDF Schema(RDFS)。RDF 和 RDFS(常合称为 RDF(S))提供了统一的、形式化的数据表示语言来描述 Web 上资源的含义，二者一起构成了语义 Web 的数据层，即语义 Web 体系结构的第三层。

第四层是语义 Web 体系结构的核心层，即本体层，它借鉴了人工智能领域对知识表示的研究，特别是描述逻辑，引入了更加丰富的表达能力，例如，属性取值约束、基数约束、属性的对称性和传递性等。本体(Ontology)是显式的概念化规范，具有共享性，常用

于描述共同认可的结构化知识。语义 Web 需要形式化规范地说明概念模型,因此,本体适合语义 Web 上的知识表示与推理。语义 Web 本体语言的标准是 OWL(Web Ontology Language),OWL 定义了 RDF(S)描述中使用的词汇的语义,便于 RDF(S)对元数据的处理,是计算机理解 Web 资源的基础。本体支持语义级的数据交换,而不仅仅是语法级,是语义 Web 的核心,具有重要的研究价值。

本体层的上方是逻辑层,目的是用更丰富的逻辑语言表达 Web 上的资源。证明层和信任层偏向于应用,而不仅仅是一种语言层次。从本体层以上,对资源的描述和推理都需要一种通过数字签名实现的信任机制,以保证数据的真实性。目前,这些工作尚未形成标准。

本体的研究对于语义 Web 的实现具有重要意义。语义 Web 中的信息以结构化形式表示,需要用本体来描述其中的语义,即对现有的 Web 信息进行标注。当信息用本体标注后,其内容就成为机器可识别和处理的数据,软件代理就能够理解其含义,进而自动完成互联网上的信息收集和集成,所以,本体是使 Web 具有语义的关键技术,语义 Web 的实现很大程度上依赖于本体的建立。一个典型的本体有一个层次分类,定义了类、类之间的关系以及具有推理能力的一组推理规则。近年来,越来越多的研究致力于本体构建,同时创建的本体也被广泛地应用到不同领域,如信息检索、机器翻译、知识管理、电子商务和信息集成等。

随着本体在各类信息系统中的大量应用,将本体的语义与关系数据库的语义进行关联逐渐成为数据库和语义 Web 领域的一个研究分支,主要包括从关系数据库中提取本体、本体在关系数据库中的存储、关系数据库到给定本体的映射以及基于本体的关系数据库的整合等方面,其中前两个方面涵盖了本体管理中两个主要的任务,即本体的构建与存储。

但是,随着本体应用的不断拓展与深入,人们逐渐意识到本体存在着一些不足,其中之一就是不支持不精确和不确定信息的表

示与处理。而在很多应用领域中，信息通常是不精确或不确定的，需要处理一些没有明确外延界限的模糊概念，如好、坏、年轻、年老等。这些模糊概念和人类、动物、男性等明确概念有明显的区别，此外也存在一些模糊关系，如“朋友”、“喜欢”等，它们都可以具有程度上的区别。事实上，现实世界中存在着大量不精确和不确定的知识与信息，这些与模糊概念和模糊关系相关的知识称为模糊知识。

描述模糊知识最常用的工具是模糊集合理论。模糊集合理论自提出以来，几乎对所有的传统数学分支都进行了扩展，其应用遍及各个领域。而在计算机科学领域，模糊扩展工作主要集中在数据库方面，形成模糊数据库模型，根据模型应用的不同目的，可以将模糊数据库模型划分为两类，一类是模糊概念模型，其中以模糊 ER/EER 模型为代表；另一类是模糊数据模型，其中以模糊关系模型为代表，这两类模糊模型分别属于两个不同层次。

近年来，随着语义 Web 的发展和本体应用的不断深入，为克服本体在模糊知识表示和处理方面存在的不足，模糊集合理论用于语义 Web 描述语言以及本体的模糊扩展逐渐成为一个研究热点，同时，有大量研究致力于模糊本体的构建。作为 Web 时代模糊数据表示和处理的两个重要技术方法，模糊数据库模型和模糊本体之间存在着密切的关联关系，本书在对语义 Web 描述语言模糊扩展的基础上，着重研究模糊数据库模型支持的模糊本体管理中的关键技术。

将模糊数据库模型引入模糊本体的研究鉴于两方面的需求：一方面，从构建的角度，模糊数据库模型可以作为构建模糊本体的数据源，使模糊本体充分利用现有的模糊数据库模型中的信息，丰富模糊本体的知识表达能力，并将模糊数据库模型中的信息通过模糊本体在 Web 上发布以实现共享；另一方面，从存储的角度，为更好地管理和使用模糊本体，模糊本体需要合理、有效地存储起来。利用模糊关系数据库在模糊数据存储和处理等方面的优势，能够实现对语义 Web 上的模糊信息的更好管理。事实上，模糊本

体的构建与存储是模糊本体管理中的两个主要问题,因此,将模糊数据库模型引入到模糊本体的研究中是一个很有意义的研究课题。

1.2 国内外相关研究的现状与分析

鉴于本体的构建与存储方法还远没有成为一种工程性的活动,在介绍模糊本体相关研究工作之前,首先介绍基于关系数据库的本体构建与存储方法的研究工作。

1.2.1 基于关系数据库的本体研究

基于关系数据库的本体研究是数据库和语义 Web 研究领域的一个研究分支,从广义上讲,可以分为两大类,即从关系数据库中提取本体以及将本体存储在关系数据库中,这两类分别基于本体的构建和本体的存储两方面内容,是本体管理中两大关键技术。

本体的构建方法几乎都从具体的本体建设项目中产生,因为领域各自特点不同,所以构建本体的方法也不相同,目前还没有出现成熟的统一标准支持本体建议。本体的构建方法主要支持的是手工构建方式。手工构建方式能够全面地创建特定领域的本体,但需要领域专家的参与,人力、物力资源花费较多,使得本体的构建成为一项艰巨的任务。因此,利用相关技术自动或半自动地从现有数据资源获取期望的本体是构建本体的有效途径。

近几年来,利用关系数据库这种结构化数据源构建语义 Web 本体的技术得到了广泛的研究,主要集中在对关系模式进行语义分析,获取构建本体所需的概念和关系。

文献[29-35]主要通过给出关系数据库模式到 OWL 本体的转换规则来创建本体。文献[29-31]通过分析关系模式和实例数据,提取关系数据库的语义,进而构建本体。文献[32-34]通过分析关系模式的主键、属性、引用关系和完整性约束等信息,给出了一组从关系数据库模式到 OWL 本体的通用转换规则,并利

用关系数据库中的部分数据来建立本体,形成对信息的集成和分类。文献[35]提出了一种从关系数据库向本体转换的方法,并以 FLogic 作为描述语言对本体进行了描述。

文献[36,37]主要借助中间表现形式将关系数据库模式转换成本体。文献[36]由 WonderWeb 项目组开发,是原型工具 OntoLiFT(集成到 KAON Work Bench)的一部分。它利用 F - 谓词逻辑和公理作为中间表现形式,旨在从源数据(即 XML Schema 或关系模式)中提取轻量级本体,它的不足之处在于只能提取出轻量级本体。文献[37]定义了一种从关系数据库提取本体的框架和一种描述本体的语言,利用数据库数据的概念视图描述数据模式与本体之间的语义映射,进而生成本体。

文献[38 - 40]主要建立了本体的结构而没有考虑本体实例的生成。文献[38]定义了从关系数据库生成 OWL 本体的算法,但是不产生本体实例,而是通过 R2O 文档将本体查询转换为 SQL 查询,从而获取对应的数据实例。文献[39,40]是一种半自动化本体提取方法,该方法假设关系模式符合 3NF,在此基础上提供了若干规则,分别用于获取目标 OWL 本体的类、属性、概念/属性的层次、基数和实例。该方法可以半自动化地生成本体及其实例,在生成本体时并不利用数据实例所提供的知识。

此外,还有一些研究将 ER 模型或扩展的 ER 模型转换为 OWL 本体。文献[41,42]提出一种由 ER CASE 工具建立的 ER 模式转换为 OWL 本体的自动方法。该方法还提供了从 OWL 抽象语法到交互语法的自动转化工具,使得生成的本体可直接在 Web 上发布。文献[43]提出从扩展的 ER 模型提取本体的算法并实现了相应的原型系统 Eronto,该算法中对扩展 ER 模型中实体、属性、二元联系、多元联系、单继承、多继承等情况的 OWL 表示方法进行了定义,其中使用了 OWL Full 的特性。

以上方法的主要目的都是利用关系数据库提取本体,即给定一个关系数据库,根据一定转换规则及算法,构建相应的本体。这不同于关系数据库到本体的映射,后者是假定关系数据库和本体

已经存在,在关系数据库和本体之间建立一组语义映射关系,例如文献[44-48]所涉及到的研究内容。

近几年来,随着语义 Web 的发展,有大量研究致力于本体的构建,并取得了一定的研究成果,但是若将本体应用到实际系统中,必须选择合适的方法将之存储起来,同时,本体的有效存储是对本体进行管理和使用的前提,所以本体有存储的需求。本体的存储方法主要分为基于纯文本存储和基于关系数据库存储。基于纯文本存储方法是将本体库以文件形式存储在本地文件系统中,这种存储形式的缺点在于不适合较大规模的本体库,因为它每次都需要读入内存操作,受到内存大小的限制。对于本体海量数据的存储和管理,利用关系数据库是一种较好的选择。关系数据库技术相对成熟,本体数据和传统的结构化数据可以共存,适合大规模本体数据的存储,并且易管理、便于查找。与利用关系数据库构建本体的研究工作相比,基于关系数据库的本体存储方法的研究相对较少。

文献[49-52]主要通过给出本体到关系数据库的语义映射关系,实现本体的存储。文献[49]提出一种将 OWL 文档映射到关系数据库的算法,旨在利用关系数据库来存储和查询 Web 上大量的数据,通过比较查询性能证明了该方法的有效性。文献[50,51]在分析了 OWL 本体和关系数据库模式之间概念对应的基础上,针对本体的类、对象属性、数据类型属性以及限制分别给出了本体在关系数据库中存储的算法,并以 Wine 本体为例描述了算法的执行过程。文献[52]首先分析了本体到关系数据库模式的转换在理论上的可行性,进而给出了 OWL 本体和关系数据库模式的形式化定义,定义了将 OWL 本体存储到关系数据库的转换规则,并基于 J2SE 平台对本体存储算法进行了实现。

文献[53,54]主要通过分析本体现有存储模式的不足,提出改进的本体存储模式。文献[53]提出一个用于本体存储和推理的系统 Minerva,采用 WSML-DL 语言完成了主要的描述逻辑推理任务,在分析利用关系数据库存储本体的优势后,给出了本体在

关系数据库中的存储模式。文献[54]通过对现有本体存储模式的分析,给出了本体存储模式的设计原则,并基于该原则提出了基于关系数据库的本体存储模式。

上述方法能够按照一定策略将本体组织在关系数据库中,但也存在一些不足,表现在:主要通过在本体与数据模式之间建立映射关系,来给出本体存储模式,而较少考虑本体实例的存储以及语义是否保持问题。

1.2.2 模糊关系数据库的研究

关系数据库系统是目前使用最为广泛的数据库系统之一,它以二值逻辑和严密的数学理论为基础,擅长表示精确的、有良好结构的数据,但对于现实世界中大量存在的模糊信息,却不易用传统方式表达,解决的方法是利用模糊集合理论扩展关系数据库,形成模糊关系数据库系统。

国际上对模糊关系数据库的研究始于 20 世纪 80 年代初期,旨在克服传统数据库难以表达和处理模糊信息的弱点,进而扩展关系数据库的功能,开拓更新、更广的应用领域。20 多年来,取得了丰硕的理论研究成果。基于关系数据库的模糊扩展主要从三个方向上进行,即模糊数据表示与数据模型、模糊查询和模糊数据依赖与规范化理论。

模糊数据表示所要解决的问题是在传统的关系模型的哪些方面引入模糊性,从而使得现实世界中的某些不确定或不精确信息在数据库中得到反映。相应地,有两种主要的模糊关系数据库模型。第一种模糊关系数据库模型基于模糊关系和类似关系(或接近关系),第二种模糊关系数据库模型基于可能性分布,它又可以进一步分成两类:元组与隶属度相关联、属性值由可能性分布表示。基于上面提到的基本模糊关系数据库模型,还有几种扩展的模糊关系数据库模型。例如,可以把上述两种基于可能性分布的模糊关系数据库模型结合到一起,也可以把可能性分布同类似关系(或接近关系)结合到一起。

应当指出的是,文献中根据模糊数据的表示形式已经提出多种模糊关系数据库模型,但是模糊关系数据库中的模糊性从表现形式上来看,只有两种形式,即属性值上的模糊性和元组上的模糊性。

模糊查询是指查询标准或查询条件具有模糊性,而数据库本身是传统(非模糊)数据库,此类查询的性质是为传统关系数据库提供柔性查询的方式。经典关系数据库中缺少柔性查询,所给定的查询条件和数据库的内容都是精确的。对于经典关系数据库上的模糊查询,一个需要解决的关键问题是如何把模糊查询条件转化成语义相近的精确查询条件,从而依托现有的关系数据库技术实现经典关系数据库的模糊查询。有关模糊关系数据库查询的研究,其内容涉及模糊查询处理方法、模糊查询语言等方面。

模糊数据依赖与规范化理论关注的是模糊关系数据库的设计,旨在获得合理的数据库模式,进而避免可能出现的数据冗余和修改异常。其中主要的工作集中在数据依赖方面的研究,包括模糊函数依赖和模糊多值依赖。模糊函数依赖的研究工作主要有模糊函数依赖的公理化系统和无损连接与分解,对后者的研究构成了模糊关系数据库规范化理论研究的基础。模糊多值依赖的研究主要包括文献[68,69]等。

除了关系数据库的模糊扩展外,还有一些研究工作是针对其他数据库模型进行的模糊扩展,其中以 ER 模型为代表。最先把模糊逻辑引入到 ER 模型的是该模型的创始人 Peter Chen 所在的研究组,该研究组提出了模糊 ER 模型。在此基础上,进一步扩展了传统的 ER 代数,引出了相应的模糊 ER 代数的概念。其他对 ER 模型进行模糊扩展的研究工作,还有文献[71 -73]等。

模糊 ER 模型的提出开辟了模糊概念数据建模一个新的研究领域,特别是文献[70]的工作奠定了模糊概念数据建模基础,其后这方面的研究深受该文献研究工作的影响。但是应当指出的是,由于 ER 模型自身表达能力的限制,模糊 ER 模型不能表达含模糊信息的复杂对象和复杂语义关系。因此,为了表达更多的语