商品目录

语义集成与智能服务理论研究

陈冬林　聂规划　徐尚英　著

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

商品目录语义集成与智能服务理论研究

陈冬林　聂规划　徐尚英 著

電子工業出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

在线商品数量的增加使得商品目录的作用越来越重要。本书针对现有的以单源商品目录为主的电子商务网站模式缺乏个性化服务的问题进行了研究。首先界定了商品目录的概念并设计了商品目录本体模型；然后研究商品目录本体规则推理技术，设计了个性化商品目录本体模型；最后对开发原型系统进行实证研究。本书在推进网络环境下基于本体的商品目录标准制定，促进供应链管理中商品目录集成，满足个性化商品目录服务等方面，具有较强的理论意义和实用价值。

本书可供从事电子商务、计算机、管理科学科研、教学或学习的高、中级人员，研究生，以及相关专业的电子商务工作者阅读参考。

图书在版编目（CIP）数据

商品目录语义集成与智能服务理论研究/陈冬林，聂规划，徐尚英著. —北京：电子工业出版社，2014.1
ISBN 978-7-121-21795-1
Ⅰ. ①商… Ⅱ. ①陈… ②聂… ③徐… Ⅲ. ①商品目录—语义分析—研究 Ⅳ. ①F760.1
中国版本图书馆 CIP 数据核字（2013）第 261658 号

责任编辑：赵 娜　　特约编辑：王 纲
印　　刷：三河市双峰印刷装订有限公司
装　　订：三河市双峰印刷装订有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　邮编 100036
开　　本：720×1000　印张：13.75　字数：308 千字
印　　次：2014 年 1 月第 1 次印刷
定　　价：42.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

服务热线：（010）88258888。

作者简介

陈冬林

武汉理工大学经济学院教授、管理学博士、博士生导师，电子商务与智能服务研究中心副主任，电子商务系主任，湖北省系统工程学会理事，武汉市系统工程学会常务理事。

曾任职于武汉天喻信息产业股份有限公司，具备丰富的电子商务和电子政务设计、开发经验。美国伊利诺依大学香槟分校电子商务中心访问学者，专注于现代服务业、语义网、本体论及智能推荐技术。

聂规划

武汉理工大学经济学院教授、博士生导师、电子商务学科带头人，中国管理科学与工程学会常务理事，全国计算机模拟与信息技术学会理事，湖北省电子商务学会副理事长，湖北省系统工程学会理事，武汉市系统工程学会常务理事。

专注于语义网、智能服务、个性化推荐和电子商务的研究与实践，以及相关技术咨询和培训工作。

徐尚英

武汉理工大学产业经济学在读博士，加拿大渥太华大学 Telfer 管理学院访问学生。主要研究智能服务、语义网、知识管理及智能推荐技术。

前言

电子商务的迅猛发展使在线商品数量呈爆炸式增长，不同电子商务网站的商品目录分类标准不统一且过于繁杂严重影响了客户体验，因此友好的商品目录变得尤为重要。据统计，90%以上的用户只对特定的有需求的商品目录感兴趣。多标准源和多数据源环境下存在概念众多、国际产品分类标准各异、企业产品数据库分散、多源商品目录无法有效集成等问题，使查询和搜索方式无法满足智能化和个性化服务的要求。

本书在国家自然科学基金项目“多源电子目录语义集成与个性化服务理论研究”的资助下，借鉴了现有的个性化服务理论、电子目录标准、本体论和语义 Web 研究，从客户的角度研究多源商品目录环境下实现标准整合、语义集成、智能推理和个性化定制的目录服务理论、方法和体系。本书首先系统分析了商品目录的概念并界定了其内涵和外延，设计了商品目录本体模型；研究了多源环境局部商品目录本体半自动生成、本体规则抽取理论、基于产品分类的标准集成方法；研究了结构化多层次商品目录本体集成方法（概念合并、属性集成、实例消重及规则集成）。其次设计了个性化商品目录本体模型，研究了商品目录本体规则推理技术，基于推理结果集对个性化商品目录进行了语义扩展，以构建按需定制的客户端商品目录。最后对开发原型系统进行了实证研究。本研究成果在推进网络环境下基于本体的商品目录标准制定，促进供应链管理中商品目录集成，满足个性化商品目录服务等方面具有较强的理论意义和实用价值。

本书由陈冬林教授主笔，负责全书写作大纲的拟定并定稿。撰写分工如下：第 1、2 章由聂规划教授执笔，第 3、6、7 章由陈冬林教授执笔；第 4、5 章由徐尚英博士执笔。

在本书的撰写过程中，我们得到了有关院校的同行和相关部门的大力支持与帮助，特别是武汉理工大学电子商务专业的老师和学生们，他们对本书的撰写提出了许多宝贵的建议，在此一并致谢。

作者

2013 年 9 月

反侵权盗版声明

目录

第1章

国内外相关基础理论研究

1.1 商品目录国内外研究现状

1.1.1 语义网

1. 语义网的起源及概念

第一代 Web，即 WWW，又称万维网，是构建在 Internet 上的，采用 B/S 网络计算模式，访问遍布在 Internet 上的所有计算机上的链接文件。这时的 Web 以 HTML 语言、URL、HTTP 等技术为标志，用静态页面的平台形式来展现信息。

第二代 Web 以动态 HTML 语言、Java Script、VB script、ActiveX、API、CGI 等技术为标志。它允许用户通过交互查询数据库，并将数据库中符合要求的结果动态地生成页面，然后展示给用户。

第三代 Web 是一个庞大的知识库，Web 信息无法被自动处理，计算机在其中扮演了展现信息的作用，而没有理解和处理 Web 信息的能力；Web 信息无法被有效利用，基于传统技术的搜索引擎已经无法应对 Web 这个日益庞大的知识库。由于计算机无法精确识别 Web 上的内容，当前搜索引擎

返回的结果中存在着大量的垃圾信息，搜索结果和质量并不令人满意。事实上人们真正关心的是信息的内容，只有对信息的内容的含义进行描述，才能实现智能化的 Web 服务，为此 Berers-Lee 在 2000 年又提出了语义网。

当前对语义网的概念还没有形成统一的定义，对语义网的理解表述不一。主流的定义主要有：

（1）语义网是第三代 Web，其目标是实现机器自动处理信息，它提供诸如信息代理、搜索代理、信息过滤等智能服务。

（2）语义网不同于现存的万维网，其数据主要供人类使用，新一代 WWW 中将提供也能被计算机所处理的数据，这将使得大量的智能服务成为可能。

（3）语义网研究活动的目标是开发一系列计算机可理解和处理的表达语义信息的语言和技术，以支持网络环境下广泛有效的自动推理。

（4）语义网的创始人 Tim Berners-Lee 对语义网的定义如下：语义网是一个网，它包含了文档或文档的一部分，描述了事物间的明显关系，且包含语义信息，以利于机器的自动处理。

尽管对语义网的理解与描述不同，但仍能从这些描述与理解中看出语义网的一些基本特征：

（1）语义网不同于 WWW，它是现有 WWW 的扩展与延伸；

（2）现有的 WWW 面向文档，而语义网则面向文档所表示的数据；

（3）语义网将更利于计算机“理解与处理”，并将具有一定的判断、推理能力。

只有当数据不仅可以被人而且可以被机器自动共享和处理的时候，Web 的潜力才发挥到极致。

语义 Web 最大的优点是可让计算机具有对网络空间所存储的数据进行智能评估的能力，这样计算机就可以像人脑一样“理解”信息的含义，完成“智能代理”的功能。使用语义 Web 搜索引擎搜索的结果比 Web 更为精确。

语义 Web 提供了一种崭新的信息描述和知识表达的手段，而要在语义层次上实现信息的互操作，就需要对信息含义的理解达成一致。语义 Web 采用了本体的思想，本体描述的是具有共识的、概念化的事物，它对实现语义层次上的知识共享、知识重用发挥着核心作用。

语义 Web 的目标是让 Web 上的信息能够被机器理解，从而实现 Web 信息的自动处理，以适应 Web 信息资源的快速增长，更好地实现任何计算机的交互以及合作。近年来，无论在国际上，还是在国内，人们对语义 Web 机器关键技术和应用的研究如火如荼，语义 Web 的支撑软件与应用开发日益受到重视，语义 Web 被看成新一代的信息基础设施，被人们称为第三代 Web。

2．语义网的架构

Berners-Lee 于 2000 年提出了语义网的体系结构（见图 1-1），并对此做了简单的介绍。该体系结构共有七层，自下而上其各层功能逐渐增强。

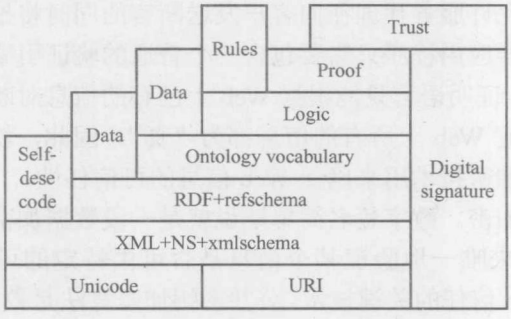

图 1-1　语义网体系结构

第一层，Unicode 和 URI。URI 是 Web 的核心概念之一，它能够唯一地识别 Web 上的任意一个资源，其思想是在需要的时候通过链接引用资源，因此不需要对资源进行复制或集中管理。Unicode 是一种新的字符编码标准，它支持世界上所有的语言，无论在什么平台、什么程序中，每个字符都对应于一个唯一的 Unicode 编码值。

第二层，XML+NS（名称空间）+xmlschema。XML 提供文档结构化的语法，实现了文档结构与文档表现形式的分离，根据不同的目的同一个文档可以有不同的表现形式。XML 名称空间是名称的一个集合，用于文档元素和属性名有效性的验证，由 URI 引用来标示。xmlschema 是约束 XML 文档结构的语言。

第三层，RDF+refschema。XML 实现了文档结构化，但文档信息并不包含任何语义。RDF 数据模型提供简单的定义，RDF 属性可以看成资源的属性，同时又表达了资源之间的关系，因此 RDF 数据模型对应于传统的

属性二值对，又类似于 E-R 图。RDF Schema 为 RDF 模型提供了一个基本的类 ing 系统，其目的就是定义资源的属性，定义被描述为资源的类，并对类和关系的可能组合进行约束，同时提供约束违例的监测机制。

第四层，Ontology vocabulary（本体层）。虽然 RDF（S）能够定义对象的属性和类，还提供了类的泛化等简单定义，但它不能明确表达描述属性或类的术语的含义及术语间的关系。本体层就是要提供一个明确的形式化语言，以准确定义术语的语义及术语间的关系。

第五至七层，Logic（逻辑）、Proof（证明）、Trust（信任）。除了本体层定义的术语关系和推理规则外，还需要一个功能强大的逻辑语言来实现推理。证明语言允许服务代理在向客户发送断言的同时将推理路径也发送给客户代理。这样应用程序只需要包含一个普通的验证引擎就可以确定断言的真假。但是，证明语言只能根据 Web 上已有的信息对断言给出逻辑证明，它并不能保证 Web 上所有的信息都为“真”。因此，软件代理还需要使用数字签名和加密技术用来确保 Web 信息的可信任性。

数字签名和加密。数字签名简单地说就是一段数据加密块，机器和软件代理可以用它来唯一地验证某个信息是否可由特定的可信任的来源提供。它是实现 Web 信任的关键技术。公共密钥加密算法是数字签名的基础。

3．语义网关键技术

要实现语义网，首先要解决信息描述问题，即如何在 Web 页面上添加机器可理解的语义信息。目前，已开发出许多语言，通过采用不同的语法、语义来解决网上知识的标示问题，主要包括 XML、RDF、OIL、DAML+OIL 和 OWL 等语言。语义网的实现需要三大关键技术的支持：XML、RDF 和 Ontology。

1）XML

XML 为可扩展性标记语言，它与 HTML 的固定标记集合所不同的是，允许其制作者创建自己的标记，通过标记来对网页的内容进行注释，并利用标记的层次关系使文档具有结构性。也就是说，XML 是一种可拓展的元指标语言，允许程序开发人员根据其所提供的规则，制定各种各样的指标语言，主要通过数据文档、DTD、样式单三个分离的部分来描述数据。其突出的优点是：良好的可扩展性、内容与形式分离、遵循严格的语法要求，便于不同系统之间信息的传输，有较好的保值性。

但是 XML 的信息是数据层面的，要利用标记来获得语义信息，首先要明白各个标记的含义，如果事先没有对应用的标记名称、组织格式和含义进行一致的约定，计算机就很难表示标记的含义。因而 XML 本身是缺乏予以描述能力的，在语义网体系结构中知识作为语法层而存在。为此，W3C 组织推荐用 RDF 来解决 XML 的语义局限性问题。

2）RDF

RDF 是 W3C 组织推荐使用的用来描述资源及其之间关系的语言规范，具有简单、易扩展、开放性、易交换和易综合等特点。RDF 由三个部分组成：RDF Data Model、RDF Schema 和 RDF Syntax。RDF Data Model 提供了一个简单但功能强大的模型，通过资源、属性及其相应值来描述特定资源。模型定义为：

（1）它包含一系列的节点 N；

（2）它包含一系列属性类 P；

（3）每一属性都有一定的取值 V；

（4）模型是一个三元组，{节点，属性类，节点或原始值 V}；

（5）每一个 Data Model 可以看成由节点和弧构成的有向图。

模型中所有被描述的资源以及用来描述资源的属性值都可以看成“节点”（Node）。由资源节点、属性类和属性值组成的一个三元组叫做 RDF Statement （或 RDF 陈述）。在模型中，陈述既可以作为资源节点，同时也可以作为值节点出现，所以一个模型中的节点有时不止一个。这时，用来描述资源节点的值节点本身还具有属性类和值，并可以继续细化。

为了避免不同词表的名字冲突 ，W3C 还开发了一种轻量级的模式定义语言 RDF Schema，提供了定义在 RDF 之上的抽象的词汇集，引入了一些诸如类、属性等面向对象设计概念。RDF Schema 定义了如下内容。

三个核心类：rdf：Resource，rdfs：property，rdfs：Class。

五个核心属性：rdf：type，rdfs：subClassOf，rdfs：seeAlso，rdfs：subPropertyOf，rdfs：isDefinedBy。

四个核心约束：rdfs：ConstrantResource，rdfs：range，rdfs：ConstraintProperty，rdfs：domain。

RDF Syntax 构造了一个完整的语法体系以利于计算机的自动处理，它以 XML 为其宿主语言，通过 XML 语法实现对各种元数据的集成。

虽然 RDF 可以用来描述 Web 数据的语义，但在表达能力和逻辑严格性方面却存在着不足，这对于构造一个真正支持丰富语义的 Web 是有影响力的。为此，人们引入基于本体的描述语言——OIL 、DAML+OIL+OWL。

3）OWL

OWL 则是 W3C 小组提出的一种基于本体描述语言，在 DAML+OIL 的基础上提供了附加的词汇表，其中的词汇都有正式语义，可明确表达各词汇的含义及它们之间的相互关系。通过一定的处理机制计算机之间就能够协同工作。OWL 语言能够清楚地表达实体间的术语词汇和联系。

4）Ontology 本体

Ontology 本体是语义网的另一项关键技术，将在下一节介绍。

1.1.2 本体

1. 本体的概念

Ontology （本体或本体论），原本是一个哲学上的概念，用于研究客观世界本质。目前 Ontology 已经被广泛应用到包括计算机科学、电子工程、远程教育、电子商务、智能检索、数据挖掘等在内的诸多领域。它是一份正式定义名词之间关系的文档或文件。一般 Web 上的 Ontology 包括分类和一套推理规则。分类，用于定义对象的类别及其之间的关系；推理规则，则提供进一步的功能，完成语义网的关键目标，即“机器可理解”。本体的最终目标是“精确地表示那些隐含（或不明确的）信息”。

当前对本体的理解仍没有形成统一的定义，如本体是共享概念模型的形式化规范说明，通过概念之间的关系来描述概念的语义，本体是对概念化对象的明确表示和描述，本体是关于领域的显式的、形式化的共享概念化规范等。斯坦福大学的 Gruber 给出的定义得到了许多同行的认可，即“本体是概念化的显示规范”。概念化（Conceptualization）被定义为：C=<D,W,Rc>，其中 C 表示概念化对象，D 表示一个域，W 是该领域中相关事物状态的集合，Rc 是域空间上的概念关系的集合。规范（Specification）可形成对领域内概念、知识及概念间关系的统一的认识与理解，以利于共享与重用。

2. 本体描述语言

本体需要某种语言来对概念化进行描述，按照表示和描述的形式化的

程度不同，可以将本体分为完全非形式化本体、半非形式化本体、半形式化本体和严格形式化的本体。有许多语言可用于表示 Ontology，其中一些语言是基于 XML 语法并用于语义网的，如 XOL（Xml- based Ontology exchange Language），SHOE（Simple HTML Ontology Language），OML（Ontology Markup Language）以及由 W3C 组织创建的 RDF 与 RDF Schema（RDFS）。还有建立在 RDF 与 RDFS 之上的、较为完善的 Ontology 语言 DAML（DARPA Agent Markup Language）、OIL 和 DAML+OIL，如图 1-2 所示。

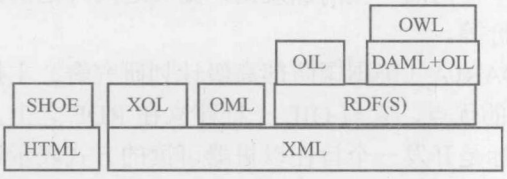

图 1-2　本体描述语言关系图

XOL 是一种基于 XML 语法和 OKBC 语义的本体交换语言。它由美国生物信息学术团体设计，用于其领域的一组异构软件系统间本体定义的交换，它以 Ontolingua 和 OML 作为基础，融合了 OKBC 的高层表达方式和 OML 的语法。当前还没有支持 XOL 本体开发的工具，但由于它采用 XML 语法，可以采用 XML 编辑器来创建 XOL 文件。SHOE 由马里兰大学开发，它将机器可读的语义知识与 HTML 文档或其他 Web 文档相结合，允许直接在 WWW 的基础上设计和应用本体。近来 SHOE 的语法已转向 XML，它使得代理（Agents）能够收集有意义的 Web 页面和文档的信息，改善搜索机制和知识收集。OML 由 Washington 大学开发，部分基于 SHOE。它有四个层次：OML 核心层（与语言的逻辑层相关）、简单 OML（直接映射 RDF 和 RDFS）、简化 OML 和标准 OML。

RDF 是 W3C 推荐的一种信息描述方式，目的是克服 XML 的语义限制，提供一种简单的模式来表示各种类型的资源。在 RDF 的基础上，RDFS 建立了一些基本的模型限制。RDF 具有较强的表达能力，但仍存在一些不足，如 RDF 没有定义推理和公理的机制，它没有说明包含特性以及没有版本控制等。