基于Ontology的非结构化信息语义表示机制研究*)

顾晋广[1,2] 陈和平[3] 陈莘萌[1]

(武汉大学计算机学院 武汉430072)[1] (武汉科技大学计算机科学与技术学院)[2]

(武汉科技大学 信息科学与工程学院 武汉430081)[3]

摘 要 考虑到目前非结构化信息表示机制的不足,本文结合XML Schema和Ontology的各自优势,提出一个用于在分布式环境下进行语义信息处理的体系结构OBSA,解决了非结构化信息表示机制中信息源异构性及语义不确定性等问题。重点介绍了OBSA体系结构中基于F-Logic的语义信息表示机制以及一个在分布式环境下处理异质信息的语义适配器框架。

关键词 本体,F-Logic,语义适配器

An Ontology-Based Representation Architecture of Unstructured Information

GU Jin-Guang[1,2] CHEN He-Ping[3] CHEN Xin-Meng[1]

(College of Computer, Wuhan University, Wuhan 430072)[1]

(College of Computer, Wuhan University of Science and Technology, Wuhan 430081)[2]

(College of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan 430081)[3]

Abstract Integrating with the respective advantages of XML Schema and Ontology, this paper puts forward a semantic information processing architecture-OBSA to solve the problem of heterogeneity of information sources and uncertainty of semantic. It introduces an F-Logic based semantic information presentation mechanism and an adapter framework for accessing distributed and heterogeneous information.

Keywords Ontology, F-Logic, Semantic adapter

1 引言

信息技术的飞速发展,促进了非结构化信息如电子邮件、工作流等在信息系统中的广泛应用,然而由于非结构化信息源的异构性以及非结构化信息的语义的多样性,很难找到一种统一的方法来表示这些非结构化信息。XML被认为是一种最可行的表示非结构化信息的方法,XML是一种数据模式的表示方法,有着丰富的内容和关系、语法和语义的分离、内容和表现的分离等特性。尽管它具有这些优势,但需要指出的是XML毕竟只是一种定义文档结构的描述性语言,并且具有语法的多样性。Ontology为特定领域内应用系统的设计提供共享的概念体系,能够减少或消除概念及术语上的混乱,使计算机对特定领域的知识处理更为精确、更为便捷[1]。如果将Ontology与XML文档结构相关联,势必可以较好地提高XML表达非结构化信息的语义丰富性和XML文档中各元素联系的准确性。

众多系统如Ontobroker[2]等在这方面进行了有益的尝试,但没有提出一个完整的基于Ontology的语义信息处理框架。笔者设计了一个基于本体的语义信息处理体系框架,尝试提供一个统一的解决方案用于异质信息源的提取和处理,笔者将这个体系结构框架命名为OBSA。本文首先介绍了OBSA的总体框架,然后介绍了基于F-Logic的语义信息表示。最后介绍了一个用于从异质数据源获取语义信息的语义适配器模型。

2 OBSA的总体结构

如图1所示,OBSA从逻辑上可以划分为五层,分别是数据源层、中间数据层、语义数据存储层、语义访问接口层和应用层。其中,数据源层可以是各种分布自治系统,如企业内部应用系统、Web服务器、文件服务器以及关系型或对象型数据库等,语义转换适配器依照特定的机制完成非结构化信息的表示工作,形成中间数据。语义数据集成层则负责将中间数据层的语义数据按照一定的方式组织存储,为接口层提供语义访问的基础,进而为应用层用户提供

*)本文受湖北省教育厅科学研究计划(项目编号:2003A012)和湖北省自然科学基金项目计划(项目编号:2003ABA049)资助。顾晋广 在读博士研究生,讲师,主要从事分布式系统与知识工程方面的研究。陈和平 教授,硕士生导师,主要从事计算机网络与数据库应用技术方面的研究。陈莘萌 博士生导师,主要从事分布式与智能系统方面的研究。

语义级的统一服务平台。

非结构化信息源存在于异构的分布式环境中，通常具有不同的数据类型和数据操作；另一方面，每个信息源具有相对稳定的语言环境、相对稳定的模式，不同信息源通常反映现实世界的一个侧面，它们之间在语法和语义上相互不能兼容。为了有效共享这些信息，实现它们之间的互操作，必须给用户提供一个全局的、一致的语义视图，以克服各个信息源之间在语义上的差异。

OBSA 采用了自顶向下的语义方法将非结构化信息收集起来，统一存储在语义数据集成层，其过程如下：

(1)Ontology 建立：在领域专家的参与帮助下，建立相关领域的 Ontology 作为 OBSA 模型的全局语义视图统一底层各信息源的语义。从图1中可以看到，Ontology 是 OBSA 的语义核心，它不仅支持集成层中 XML 文档的语义表示，还为访问接口层中的查询、推理等过程提供语义指导。

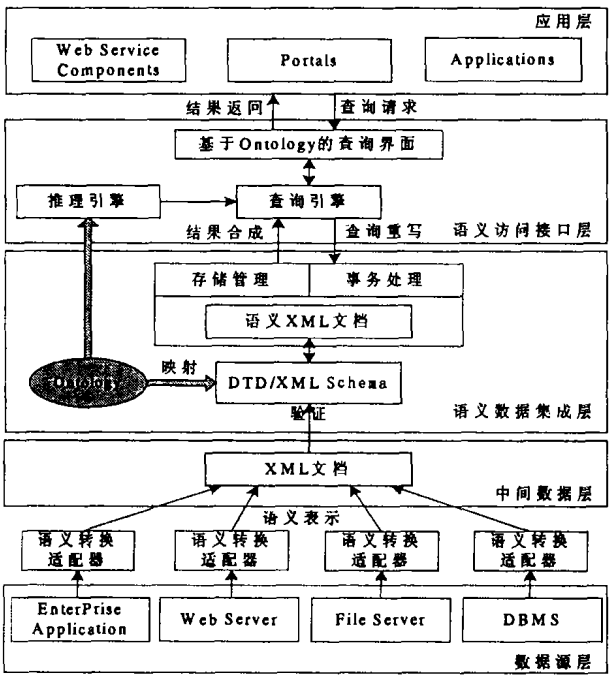

图1 OBSA 体系结构

(2)信息收集、组织与存储：收集各种非结构化信息，并参照已建立的 Ontology 对它们进行语义表示，最后统一存储在公共数据仓库中。

(3)查询处理：对用户查询界面获取的查询请求，按照 Ontology 提供的语义约束转换成一定的形式，在 Ontology 的协助下从公共数据仓库中匹配出符合条件的数据集合。OBSA 采用 F-Logic 描述语言来表示 Ontology，并利用 F-Logic 的逻辑推理能力来完成智能化的语义信息检索。

(4)检索结果处理：检索的结果经过定制合成处理后，返回给用户。

3 OBSA 的语义处理

Ontology 和 DTD/XML Schema 都提供了描述信息的词汇集和结构，不同的是前者用来指定领域知识，侧重于概念级别的语义表达，后者则是提供信息资源完整性约束的一种手段，侧重于 XML 文档结构的语法描述[3]。从 Ontology 中导出 DTD/XML Schema 就是把 Ontology 中的概念实体、属性、概念之间的关系语义保持地映射为 XML 文档的结构及其标记集合，形成新的模式规范，则满足这种 XML 模式规范语法验证的 XML 文档就能与 Ontology 相兼容，从而可在一定程度上达到语义验证的目的。由于 XML 的元素、属性标记及标记之间的相互关系能够表现明确的领域知识，对信息的描述将从原有的语法表示级上升到概念及其间关系的抽象级，实现了 XML 表示非结构化信息的一致性，避免了语义异构冲突，同时，使用基于 Ontology 的标准化语义级标记来描述文档的内容，能帮助计算机自动从文档中抽取信息的语义。因此，对于 OBSA 的语义处理，重点要做好两个方面的工作，首是要寻找一种合适的表示语言来表示本体信息，另外一个方面是设计一个算法，实现从本体到 XML DTD/Schema 的映射。

OBSA 采用描述语言 F-Logic[4] 来表示 Ontology。F-Logic 是一种面向对象、基于框架的语言，它提供了定义 ontology 所需的基本建模元语。使用 F-Logic 定义的 Ontology 通常包括三个组成部分[2]：

· 概念类的层次定义：定义不同概念类之间的子类关系；

· 属性定义：定义概念类的属性并声明属性值的有效类型；

· 规则集合：定义不同概念类和属性之间的关系。

转换算法以 Frame-Logic 表示的 Ontology 作为输入，将 Ontology 的特定部分映射为 DTD/XML Schema 中的相应结构并输出结果 DTD/XML Schema，基本思路为[5,6]：

· Ontology 中的每个概念在 DTD/XML Schema 中生成一个元素类型；

· 概念的每个属性在 DTD/XML Schema 中生成此概念对应元素类型的一个子元素(或者生成对应元素类型的一个属性，视具体情况而定)；

· 如果属性表示的是与其它概念的关系，则该属性转换而来的元素的内容模型为相关的概念元素，否则为其原有类型。

4 语义适配器的实现

数据源层的原有系统往往已经在程序和知识

库、知识表达方面固化了自己对领域知识的理解，并且各系统的知识库和知识表达的语言和语法都是不一样的，形成了各自的局部语义模式。

OBSA 采用自顶向下的集成方法，为用户提供了一个基于特定领域 Ontology 的统一的全局语义视图，同时，每个数据源都配备有一个语义转换适配器，负责在各自的局部语义模式与全局语义模式之间建立映射，对参与集成的非结构化信息提取出来进行语义表示，预先转换成全局语义模式，使得语义异构的数据环境在集成层上得到语义的统一。

OBSA 的语义数据集成层为用户提供领域 Ontology 的同时还针对每个领域 Ontology 生成了 DTD/XML Schema，作为用户描述相应领域内非结构化信息的语义模版，用户可以在理解集成模式 Ontology 的基础上参照 DTD/XML Schema，使用 Ontology 提供的概念术语，运用 XML 标记完成各自非结构化信息的语义表示工作，得到的 XML 文档就是符合集成层全局语义模式的语义 XML 文档，可直接存入存储仓库。

至此，OBSA 系统中形成了均以 XML 为表现形式的全局语义模式和局部语义模式：

· 局部语义模式是各资源提供者提交的 XML 文档，准确地说是这类 XML 文档的模式 DTD/XML Schema，因为从目前来看，DTD/XML Schema 给出了关于 XML 文档最多的信息，包括部分语义和丰富的结构信息，从 DTD/XML Schema 中寻找语义信息来构造局部语义模式是一种很自然的想法。

· 全局语义模式则是 Ontology，在此不妨认为是由特定 Ontology 所导出的含真正语义的 DTD/XML Schema。

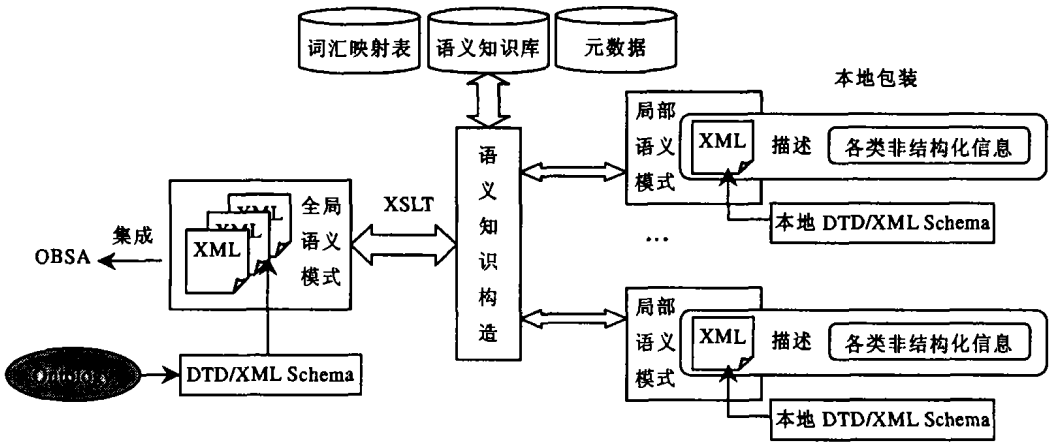

图2 语义映射示意图

局部语义模式与全局语义模式语义之间的映射如图2所示，大体上可分为3个阶段：

(1)局部语义模式抽取：该阶段通过分析数据源包装后提交的 XML 文档及其 DTD/XML Schema，来获取源数据的局部语义模式，具体指抽取 XML 文档中各元素、属性标签，发现它们在描述属性信息和内容信息时的含义和特点。

(2)概念匹配：语义知识的产生以及不同局部模式的集成很大程度上取决于概念匹配的结果。概念匹配不同于关键词匹配，它是根据局部语义模式抽取的结果，从 XML 标签描述属性信息和内容信息的特点进行匹配。该过程的初期需要有人工干预。其中涉及到语义知识的构造，首先利用同义词字典库来构造一个语义知识库，并进行训练，随着训练次数的增加，不断扩充新的语义知识。概念匹配的结果产生词汇映射表。

(3)全局语义 XML 文档形成：根据词汇映射表中的知识，将不同信息源提供的异构 XML 文档根据语义知识库的知识进行翻译，然后严格参照 OBSA 集成层中由既定 Ontology 所导出的 DTD/XML Schema，运用 XSLT 编程实现局部 XML 文档到全局 XML 文档的转换，除参考词汇映射表之外，此过程还需要人工干预识别概念术语之间的对应关系。适配器最后输出的 XML 文档就是符合集成层全局语义模式的语义 XML 文档，可以直接通过集成层的 DTD/XML Schema 验证进入存储仓库。

结束语 由于信息源的异构性和语义的不确定性，因此，在分布式环境下表示非结构化语义信息是一个十分复杂的过程。通过语义适配器，采用基于 F-Logic 的语义表示机制以提供一个全局的语义环境是一种可行的方案。同时，语义适配器的这种结构也符合软件工程中的软件设计模式方法，是一种十分有益的尝试。

本文所提出的 OBSA 体系结构框架与大家所熟知的非结构化信息表示技术，如语义网、知识网

基于决策树改进CART算法的决策支持与分析技术

李春鑫[1] 李天伟[2]

(海军大连舰艇学院研究生15队 大连116018)[1] (海军大连舰艇学院航海系 大连116018)[2]

摘 要 针对决策支持与分析技术,提出了基于决策树的改进CART算法。该算法由树生长和树剪枝两部分构成,具有辨识相关输入的能力,由于引入了递归最小二乘估计器,对线性模型可降低计算量,并采用模糊技术处理不连续边界问题。我们给出了该算法的应用实例,由于隐含权值归一化,该算法能够快捷地对自适应神经模糊推理系统进行结构辨识。

关键词 决策树,CART算法,递归最小二乘估计器,自适应神经模糊推理系统

Decision Support and Analysis Technique Based on Improved CART Arithmetic of Decision Trees

LI Chun-Xin[1] LI Tian-Wei[2]

(Dalian Naval Academy, Dalian 116018, China)[1] (Dept Navigation, Dalian Naval Academy)[2]

Abstract Aiming at decision support and analysis technique, improved CART arithmetic of decision trees is put forward. The arithmetic consists of developing trees and cutting trees, because recursive LSE is introduced, it can reduce calculation for linear model, besides, to deal with the question of discrete boundary, fuzzy technique is used, and it also has the ability of identifying relevant input. The application of the arithmetic is also presented. Due to the normalization of connotative weight, the arithmetic can identify the structure of ANFIS conveniently.

Keywords Decision trees, CART, RLSE, ANFIS

1 引言

决策是针对某一问题,根据确定的目标及当时的实际情况制定多个候选方案,然后按一定标准从中选出最佳方案的思维过程。在许多情况下,决策一旦作出就无法挽回,因而如何快速准确地作出决策显得极其重要。基于此我们提出了决策树的改进CART算法。

2 CART算法的原理

决策树是把数据集的输入空间划分为互斥区域,每个区域赋予一个标识、一个值和一个表示该区域内数据点特色的动作。为了构造一个合适的决策树,CART首先基于采样数据集广延地生长树,然后,基于最小复杂性代价准则[1],再回头修剪这棵树,这样得到一系列不同大小的树,最后所选择的那

李春鑫 硕士研究生,主要从事军事航海信息及控制关键技术的研究。

格[7,8]等并不违背,相反,OBSA是上述两种技术的一种具体的实现方式。

参 考 文 献

1 Uschold M, Gruninger M. Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, 1996, 11(2): 93~155

2 Decker S, Erdmann M, Fensel D, et al. ONTOBROKER: Ontology based Access to Distributed and Semi-Structured Information. In: R. Meersman et al. eds. Semantic Issues in Multimedia Systems Boston: Kluwer Academic Publiser, 1999

3 Klein M, Fensel D, Harmelen F, et al. The Relation between Ontologies and XML Schemata. In: Proc. of the (ECAI)'00 Workshop on Applications of Ontologies and Problem-Solving Methods, Berlin, 2000

4 Kifer M, Lausen G, Wu J. Logic Foundations of Object-Oriented and Frame-Based Languages. Journal of the ACM. 1995, 42(4): 741~843

5 Erdmann M, Studer R. Ontologies as Conceptual Models for XML Documents. In: Proc. of the KAW '99 12th Workshop on Knowledge Acquisition, Modelling and Management

6 张维明. 语义信息模型及应用. 电子工业出版社, 2002. 77~79

7 Cannataro M, Talia D. Knowledge Grid—An Architecture for Distributed Knowledge Discovery. CACM, 2003, 46(1): 89~93

8 Goble C A, Roure D D, Shadbolt N R, et al. Enhancing Services and Applications with Knowledge and Semantics. In: The Grid2: Buleprint for a New Computing Infrastructure, New York: Morgan Kaufmann Publishers, 2004. 431~458