基于本体的数据交换系统研究

何 为[1],侯 锋[2],徐东平[1]
(1.武汉理工大学 计算机科学与技术学院,湖北 武汉 430063;
2.国防科技大学 信息系统与管理学院,湖南 长沙 410073)

摘 要:随着 Internet 和 Intranet 技术的迅速发展,异构系统间的信息交换变得日益频繁。当前的方法大都采用对不同的 XML Schema 手动建立映射关系来处理语义异构的问题。当有新的企业加入的时候,该企业的 XML Schema 要与所有其它企业的 XML Schema 进行映射,很显然这种方式不利于系统的扩展。本体对知识的共享表示有助于对内容语义的精确、高效通信。采用 XML 作为中间文件格式,结合目前的信息共享技术,提出了一种基于 XML 和本体的数据交换模型。讨论了该模型的系统架构和关键技术。

关键词:XML;本体;数据交换;异构系统
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2008)06-0047-03

Research on Ontology-Based Data Exchange System

HE Wei[1], HOU Feng[2], XU Dong-ping[1]
(1. School of Computer Sci. & Techn., Wuhan University of Technology, Wuhan 430063, China;
2. School of Info. System & Management, National University of Defense Technology, Changsha 410073, China)

Abstract: Along with the fast development of Internet and Intranet, the information exchange among heterogeneous system becoming more and more frequent. At present, the main method to solve the semantic heterogeneous is constructing mapping relation different XML Schema. However, when a new enterprise join, the new XML Schema must mapping to all the other XML Schemas; therefore, it is difficult to expand the system. The knowledge share representation of ontology can improve the precision and communication of semantics of content. In this paper, an XML and ontology-based data exchange model was proposed, and this data exchange model makes the XML as format. The system architecture and key technologies was discussed.

Key words: XML; ontology; data exchange; heterogeneous system

0 引 言

随着 Internet 和 Intranet 技术的迅速发展,企业间的信息交换变得日益频繁。而企业间的数据大都是异构的,数据的异构性主要表现为系统异构、数据模式异构、语义异构等三类[1]。系统异构主要指数据所依赖的应用系统不同,如数据库管理系统、硬件平台、操作系统、并发控制、访问方式和通信能力的不同。数据模式异构主要指数据在存储模式上的差异;一般的存储模式包括关系模式、对象模式、对象关系模式和文档嵌套模式等几种,其中关系模式为主流存储模式;需要注意的是,即便是同一类存储模式,它们的模式结构可能也存在着差异。语义异构是指信息资源之间存在着语义上的区别,这些语义上的不同可能引起各种冲突,例如从简单的命名冲突(如同名异义,同义异名),到复杂的结构语义冲突(不同的模式表达同样的信息),语义冲突将会使企业间的数据交换变得复杂。

目前企业信息化的基础平台是 Internet,而 XML 已经成为 Internet 环境下数据表达的事实标准,不同格式信息都可以转换成 XML 文档,一些商业数据库如 SQL Server 和 Oracle 也提供了 XML 的导出功能。除此之外,还可以用 XML Schema 对 XML 文档的格式进行限定和验证。因此,XML 的出现很好地解决了系统异构以及数据模式异构的问题。对于语义方面的异构,一些企业或商业的电子交换软件(如 Biztalk)大都采用了对不同的 XML Schema 手动建立映射关系来处理语义异构的问题。这种方式遵循不同 XML Schema 的 XML 文件的数据转换,都需要建立它们之间的转换脚本。当有新的企业加入的时候,该企业的 XML Schema 要与所有其它企业的 XML Schema 进行映射,很显然这种方式不利于系统的扩展。为此,有必要引

收稿日期:2007-09-04
作者介绍:何 为(1980-),男,湖北仙桃人,硕士研究生,研究方向为多媒体技术;徐东平,博士,教授,研究方向为多媒体技术。

人人工智能领域中的本体(ontology)技术。本体是不同领域、不同应用系统之间进行交流、协定并可以共享理解的知识表示。这种协定有助于对内容意义的精确、高效通信;同时又反过来促使系统的交互式操作、重用和共享等一系列的性能得以提高。

文中采用 XML 作为中间文件格式,结合目前的信息共享技术,提出了一种基于 XML 和本体的数据交换模型。该模型考虑到分布式环境下各个数据源是异构的,新开发系统与遗留系统并存。

1 系统框架

本体是一种新型的元数据,其目标是捕获相关领域的知识,提供对该领域知识的共同理解,确定该领域内共同认可的词汇,并从不同层次的形式化模式上给出这些词汇和词汇间相互关系的明确定义,由此实现知识重用。为此,文中的研究思路如下:根据各个企业数据库的 XML Schema 生成局部本体,然后把局部本体与全局本体进行映射生成映射表。根据映射表的对应关系进行数据交换。系统结构如图 1 所示,具体流程说明如下:当企业用户第一次进入交换系统时,通过转换器生成该数据源的 XML Schema 并将 XML Schema 提交给数据交换中心;数据交换中心将 XML Schema 通过局部本体生成器生成该企业的局部本体;然后以局部本体与全局本体作为映射器的输入,映射器输出映射表。企业再次登录,与其它企业交换数据时,把要交换的数据通过转换器转换成 XML 文档并提交到数据交换中心;数据交换中心收到 XML 文档后利用文档转换器根据映射表转换成目标企业能识别的 XML 文档。

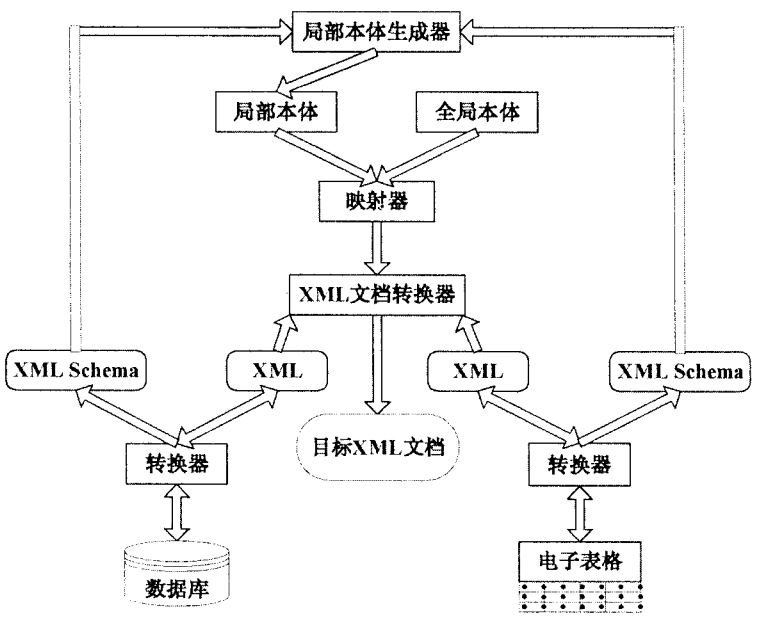

图 1 基于 XML 和本体的数据交换模型流程图

(1)转换器:转换器完成其它数据源的数据(各种数据库、电子表格等)与 XML 文档的双向转换并生成相应的 XML Schema。

XML 与关系模式的映射分为两类:基于模板驱动的映射和基于模型驱动的映射。基于模板驱动的映射是在一个模板中嵌入数据处理命令,再用数据处理中间件解析执行。这种方法以 XML 内嵌的 SQL 执行的数据集为依据,不涉及数据库赖以存在的关系模式或对象模式。基于模板驱动的映射是一种浅层映射,但其比较灵活,甚至可以加入程序逻辑。基于模型驱动的映射是指用一个具体的模型来关联 XML 文档和关系数据。常见模型有表格模型和数据专用对象模型。表格模型直接把关系数据库中的表结构映射成 XML 文档。数据专用对象模型把一个 XML 文档表示为由数据对象,每一个元素类型和对象相对应,XML 中的内容模型、属性和 PCDATA 则对应对象的属性。目前几个流行的数据库产品都支持 XML 数据导出功能,文中使用的 Microsoft SQL Server2000 属于模板映射。

(2)局部本体生成器:局部本体生成器根据各企业提交的 XML Schema 生成局部本体。

(3)映射器:映射器根据输入的局部本体和全局本体,输出两者之间的映射表。

(4)文档转换器:根据输入的 XML 文档生成符合目标 XML Schema 的 XML 文档。

2 关键技术

2.1 全局本体的建立

这里的全局本体是系统的全局视图的语义化描述,是企业业务的领域本体。比如各化工企业之间的信息交换需要的是化工领域本体。全局本体构建了待集成领域的知识模型,并为数据集成提供了公共的语义描述。Gruber 在 1995 年提出了构建本体 5 条规则:明确性和客观性、完全性、一致性、最大单调可扩展性和最小承诺。目前没有一个标准的本体构造方法。MikeUshold[2]等人提出了构建本体的骨架(skeletal)法,该方法的步骤包括:

1)识别目的和范围:这个阶段需要弄清楚为什么要建立本体、建好后的用途有哪些、使用该本体的用户范围是什么等问题。

2)建设本体:包括本体捕获、本体编码,与现有的本体集成。本体捕获识别相关领域中关键的概念和关系;给出这些概

念和关系的精确无二义的文本定义;识别能表达这些概念和关系的术语。本体编码是利用某种形式化语言显式地表现上个阶段的概念化成果。

3)评价:并没有提出自己的评价方法,引用了Gomez Perez关于评价的定义。

4)文档化:目前很多知识库和本体缺少文档也是一种知识共享的障碍,这些文档应该包括本体中定义的主要概念、meta-ontology等。

5)每阶段的指导方针:把设计本体的初始的指导方针总结为以下设计标准:清楚(Clarity)、一致(Coherence)、可扩展性(Extensibility)、最小本体的承诺(Minimal ontological commitment)、最小的编码偏差(Minimal encoding bias)。

用户使用本体描述语言为领域模型编写清晰的、形式化的概念描述。根据Gruber的规则,它应该满足以下要求:良好定义的语法、良好定义的语义、有效的推理支持、充分的表达能力、表达的方便性。大量的研究工作者活跃在该领域,因此诞生了许多种本体描述语言:和web有关的语言有RDF和RDF-S、OIL、DAML、OWL、XOL,这几种语言之间又有着密切的联系,且都是基于XML的;只在相关项目中使用的有Ontolingua、CycL和Loom;KIF已经是美国国家标准,但是它并没有被广泛应用于互联网,作为一种交换格式更多地应用于企业。

2.2 从XML Schema中学习局部本体

局部本体生成器是从XML Schema自动构建局部本体。因为XML Schema描述了XML数据的层次结构,可以认为它是XML的逻辑模型。对于这种数据可以采用映射技术构建本体,即利用一些映射规则将其中的一些元素映射到本体。这种方法的关键是映射规则的发现,现有的方法可以分为两类:一类是基于学习的方法,即利用一些机器学习的手段自动获取;一类是基于预定义规则,即根据用户预先定义的一些规则,例如,Doan[3]等人和Mello[4]等人使用预定义的规则,从DTD中提取语义信息生成相应的概念模式,然后对这些概念模式进行语义集成得到本体。但是,由于DTD和XML Schema在语法上的差异,需要使用不同的映射规则。为此,Volz等人提出将这些半结构化数据映射成一棵语法树,然后使用一些规则将这些非终结符集和终结符集中的元素映射为本体中的概念和关系。所以这种方法克服了现有模式语言在语法上的差别。文中采用预定义规则来实现映射。

OntoLiFT[5]是德国卡尔斯鲁厄大学开发的一个基于HMarfra的开源本体学习工具,它可以从半结构化数据(XML schema,DTD)和结构化数据(关系数据库)中获取本体(包括概念及其关系)。HMarfra能够实现从XML Schema到本体的映射,OntoLiFT开发了一个从DTD到XML Schema映射的中间工具。这样,将这两个工具合并起来,就实现了从XML Schema和DTD中获取本体。从关系数据库中获取本体的部分是基于Java JDBC标准提供的接口,然后按照一定的命名规范将数据库中的表名和属性名等信息,按照映射规则转换为本体中的元素。对于这两种类型的数据源,它都采用基于映射规则的方法来获取本体。在系统实现中,为了考虑以后支持DTD,在OntoLiFT的基础上,对其源代码进行了修改,使其支持中文的本体学习。

2.3 局部本体与全局本体之间的映射

本体映射是以两个本体作为输入,然后为这两个本体中的各个元素(概念、属性或者关系)建立相应的语义关系,从而将源本体的实体(概念、实例、属性等)映射到目标本体实体上。本体映射领域的研究较多,各种定义、表达甚至研究目的都差别较大(如:本体集成、本体合成等)。文献[6]提到:本体映射就是指给定两个本体A和B,对于A上的每一个实体,设法在B上找到与其有相同或相近语义的实体,这些实体包括本体的类、属性以及类的实例。Ehrig[7]给出了一个形式化的本体映射函数:map:$O_1 \rightarrow O_2$。如果本体O_1中的实体e_1与O_2中的实体e_2的语义相似度大于阈值s,则$\mathrm{map}(e_1) = e_2$。Ehrig和Staab总结出本体映射的6个过程:

(1)特征提取:提取可以计算实体语义相似度的特征,如概念、属性的名称等;

(2)选择用于映射的概念对;

(3)进行相似度计算;

(4)相似度整合:通常有多种方法可以衡量本体实体之间的相似度,得出多种相似度值,因此要对各相似度进行综合考虑,从而得到一个整体上的相似度;

(5)优化:第(4)步结束后,已经得到待映射的各个实体之间的初始相似度,这时一般需要人工的干预,利用领域知识,对其进行调节;

(6)迭代第(1)步到第(5)步,直到达到满意结果。

3 结束语

随着Internet和Intranet技术的迅速发展,异构系统间的信息交换变得日益频繁。文中针对企业间异构信息交换问题,提出了一种基于XML和本体的数据交换模型。该模型首先从关系数据库自动构建局部本体,然后与全局领域本体进行语义影射,从而将一方的

(下转第53页)

一的完整过程及其结果如图 2 所示。图 2(a)为待处理的车牌灰度图像;图 2(b)为 Otsu 二值化后的图像;图 2(c)为形态学处理后的图像,形态运算使不同的部分形成不同的连通域,但保留形状特征。原始图像外形上的形状纹路明显不见,图像中只保留了具有区域特征的连通域,原图像目标的形状特征经处理后有小的畸变,但整体形状特征得到了较好的保留;图 2(d)为删除较小连通域后的图像,图像中的连通域个数由图 2(c)的 11 个减少到 5 个,小连通域被剔除,不但减少了背景干扰,而且可大大减少后续处理的运算量;图 2(e)为标记连通域、车牌区域筛选后的图像;图 2(f)为定位切割下的车牌图像,可见被选中、标记的区域正是车牌所在区域,很好地实现了车牌的定位与分割。

图 3 为另外两幅不同背景车牌图像的定位实验。从图 3 也可以看出,车牌切割完整,无残损;背景很少,对车牌识别不会形成干扰,满足车牌定位要求。

(a)原始图像　(b)定位的车牌1

(c)原始图像2　(d)定位的车牌2

图 3　原始车牌图像及其定位结果

4　结束语

介绍了基于形态学的车牌定位算法,充分利用形态学的不同结构元形态运算具有的不同的作用效果和保持物体几何形状的特性,很好地实现了车牌定位。对各种条件下拍摄的 132 幅含有车牌的图像应用该算法,有效定位 130 幅,准确率达到 98.4%。同时,由于算法以二值图像来进行,而形态学运算固有的能将大量的复杂二值图像处理运算转换为基本的逻辑与移位运算的组合来完成,便于并行处理与硬件实现的特性,因而本算法不仅算法简单、很好地实现了车牌定位,而且运算速度快利于实时处理。

参考文献:

[1] 王洪建.基于 HSV 颜色空间的一种车牌定位和分割方法[J].仪器仪表学报,2005,26(8):71－73.

[2] 郭　捷,施鹏飞.基于颜色和纹理分析的车牌定位方法[J].中国图像图形学报,2002,7(5):472－476.

[3] 李　波,曾致远,付祥胜.基于数学形态学和边缘特征的车牌定位算法[J].电视技术,2005,24(7):94－96.

[4] 熊　军,高敦堂,都思丹,等.应用遗传算法进行车牌定位[J].计算机应用,2004,24(6):163－164.

[5] 陆　锋.基于改进的 BP 神经网络进行车牌定位的研究[J].苏州大学学报,2004,24(6):5－8.

[6] 陈寅鹏,丁晓青.复杂车辆图像中的车牌定位与字符分割方法[J].红外与激光工程,2004,33(1):29－33.

[7] Serra J. Image analysis and Mathematical Morphology[M]. London:Academic Press,1982.

[8] 袁志伟,潘晓露,陈　艾,等.车辆牌照定位的算法研究[J].昆明理工大学学报,2001,26(2):11－14.

[9] 付忠良.图像阈值选取方法——Otsu 方法的推广[J].计算机应用,2000,20(5):53－56.

[10] Song J,Delpe J. The Anaslysis of Morphological Filters with Multiple Structuring Elements[J]. Computer Vision Graphics and Image Processing,1990,50:308－320.

(上接第 49 页)

数据格式转换为对方可理解的格式,实现数据交换的目的。

参考文献:

[1] Bergamaschi S,Castano S,Capitani S. MOMIS:An Intelligent System for the Integration of Semistructured and Structured Data[J]. INTERDATA,1998,45(5):1－14.

[2] Uschold M. Ontologies Principles,Methods and Applications [J]. Knowledge Engineering Review,1996,11(2):42－54.

[3] Doan A,Domingos P,Levy A. Learning source descriptions for data integration[C]//In:Proc. of the Workshop on the Web and Database. Heidelberg:Springer－Verlag,2000:81－86.

[4] Mello Rd S,Heuser C A. A bottom－up approach for integration of XML sources[C]//In:Simon E,Tanaka A K,eds. Proc. of the WIIW. Brazil:[s. n.],2001:118－124.

[5] Volz R,Oberle D,Staab S,et al. OntoLiFT prototype[R]. IST Project 2001－33052 WonderWeb:Ontology Infrastructure for the Semantic Web. Manchester,UK:[s. n.],2003:1－25.

[6] Su X. A text categorization perspective for ontology mapping [R]. Norway:Norwegian University of Science and Technology,2002:1－9.

[7] Ehrig M,Sure Y. Ontology mapping－an integrated approach [EB/OL]. 2005－01－27. http://www. aifb. uni－karlsruhe. de/WBS/ysu/publications/2004-mapping-TR. pdf/.