• 软件技术与数据库 •

文章编号: 1000-3428(2011)14-0044-03

文献标识码: A

中图分类号: TP311

带格式的:字体颜色:自动设置

基于本体的异构数据集成框架

钟 将,宋 娟

(重庆大学计算机学院, 重庆 400044)

摘 要: 针对电力系统数据集成中存在的语义异构问题,提出一种基于本体的数据集成框架。依据电力参数估计系统的数据需求模型,分析数据集成存在的语义冲突类型,在传统数据集成框架的数据集成中间件模块中加入本体语义描述模块。采用本体描述信息资源域中的概念,通过实现语义冲突主动发现并构建语义映射关系。实验结果表明,该框架能有效解决数据集成过程中的语义异构问题。

关键词: 本体; 中间件; 语义异构; 数据集成; 语义映射

Heterogeneous Data Integration Frame Based on Ontology

ZHONG Jiang, SONG Juan

(College of Computer Science, Chongqing University, Chongqing 400044, China)

[Abstract] According to the data request model of power parameter estimate system, and semantic conflict type existing in the process of data integration, a heterogeneous data integration frame based on ontology is proposed to solve semantic heterogeneity problem. It improves traditional data integration structure by adding ontology semantic describing structure into data integration mediator. Based on describing the concept of domain by ontology, this structure solves the semantic heterogeneity problem existing in heterogeneous data integration, by finding semantic conflict initiative and constructing semantic mapping relations. Experimental result shows the feasibility and effectively of integration framework based on ontology proposed in this paper.

[Key words] ontology; mediator; semantic heterogeneity; data integration; semantic mapping **DOI:** 10.3969/j.issn.1000-3428.2011.14.013

1 概述

目前电力系统中电网结构越来越复杂,为保证电网的安全稳定运行,对各种辅助监视、分析、决策、控制系统的电力仿真分析应用需求愈加迫切。输电网参数的准确性是各种电网分析计算软件的基础。为正确识别电网参数错误,参数估计系统实现了基于单一设备多时段 PMU 数据或 PMU 数居或 PMU 数据或 PMU 数据或 PMU 数据对分别全网参数进行辨识。系统需要集成由广东省电力调度中心提供的电网模型数据和各子区域的实时数据,由于这些数据种类繁多、表示和存储形式各异,集成时系统间会出现大量的语义冲突,给数据的集成带来很大的困难。传统解决语义冲突的办法是由领域专家手工定义语义匹配表来解决集成时出现的语义冲突。随着数据源数量的不断增加和各种新技术的出现与应用,主动发现语义冲突,以化解和消除数据集成中的语义冲突问题已经成为目前研究的热点和难点。

随着语义网技术的发展,本体被引入到数据集成中[1],而本体本身具有很强的表达概念语义和获取知识的能力,使用它可以准确地描述概念含义以及概念之间的内在关系,通过逻辑推理能够获取概念之间蕴涵的关系^[2]。本文归纳了参数估计系统集成时可能存在的各种类型的冲突,提出一种基于本体并且采用语义技术的完整解决方案,通过实现主动识别语义冲突以消解集成时的语义冲突。

2 数据集成冲突及解决方案

2.1 参数估计数据需求模型

参数估计系统主要的数据源有数据采集与监视控制系统 (Supervisory Control And Data Acquisition, SCADA)提供的量 测信息和相量测量单元(Phasor Measurement Unit, PMU)提供 的高精度相量信息和电网模型数据。根据需求集成上述类型 的数据,生成统一的参数估计计算需求模型。分析上述电网数据可知,其中,SCADA数据是TXT格式的非结构化数据; PMU数据是EXCEL格式的非结构化数据;电网模型数据是XML格式的半结构化数据,以及存在于MySQL数据库中的电网实时数据。数据集成需求模型如图1所示。

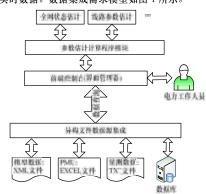


图 1 数据集成需求模型

2.2 语义冲突类型

为实现上述需求的数据集成,传统电力数据集成框架大均采用 Wrapper/Mediator 的方式构建数据集成系统。它可以

基金项目: 重庆市自然科学基金资助项目(CSTC, 2010BB2046); 国家科技支撑计划基金资助重大项目(2008BAH37B04); "211 工程" 三期建设基金资助项目(S-10218)

作者简介: 钟 将(1974—), 男, 副教授、博士, 主研方向: 数据挖掘; 宋 娟, 硕士研究生

收稿日期: 2011-01-21 E-mail: songjuanlinda@foxmail.com

很好地解决系统异构、数据结构异构和语法异构的问题,但对于集成数据时如何主动识别语义冲突问题,主要是字段冲突、表冲突以及记录冲突问题^[3]仍无效。语义冲突是指当描述同一现实世界事物时,2个对象在描述方式、结构和内容上的不同造成的语义不一致性。从采用语义技术解决问题的角度看,按层次的不同,本文对集成上述电力数据中可能遇到的语义冲突进行如下归纳:

(1)物理冲突: 指数据源的存储格式不同引起的冲突。比如: SCADA 数据是 TXT 格式的非结构化数据, 而电网模型数据是 XML 格式的半结构化数据。

(2)表冲突:包括命名冲突、结构冲突、关系冲突。命名冲突包括表名重复,比如开关对象可以用 Breaker 或 Switch 表示。结构冲突指表达相同概念的表使用的是不同的结构描述。关系冲突指各系统内部的表间关系集成到一起时所呈现出的不一致,比如在 A 系统中,2 张表之间表达的是父子关系,而在 B 系统中,类似的2 张表却表达的是等价关系。

(3)字段冲突:包括命名冲突、类型冲突、长度冲突、精度冲突、计量单位冲突、表达方式冲突。命名冲突指字段同名异义、异名同义,电压基准值和电压等级描述的是电压的同一个属性;类型冲突指表达相同的特征在不同表中采用不同的数据类型;长度冲突指表达相同特征的不同字段的数据长度不一致;精度冲突指表达相同特征在不同表中采用不同数据精度的字段;计量单位冲突指表达相同特征的数据在不同表中具有不同的计量单位;表达方式冲突指数据表示不一致,如数据格式差异、缩写差异等。

(4)记录冲突: 指描述同一数据的数据记录不同,因计量单位冲突引起的数值不同,例如变比和档位的换算: 1.23(变比)=0.5(档位)

2.3 基于本体的异构数据源集成框架

面对集成电力数据时存在的上述冲突类型,冲突的主动发现是解决语义冲突的关键。为实现数据集成时语义冲突的主动发现,本文扩展了基于 Wrapper/Mediator 的方式构建数据集成系统,将本体作为解决语义异构的工具引入到系统中。该系统利用本体在描述语义上的优势设计了语义冲突检测机制,并且通过构建语义映射关系可以很好地解决数据集成中的语义异构问题。系统框架如图 2 所示。

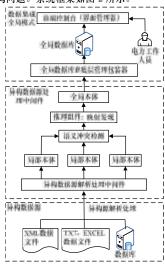


图 2 基于本体的数据集成框架

该系统对异构数据源处理中间件模块进行了扩展,加入了本体的语义描述模块。该模块实现异构数据源到局部本体的构建,并且利用语义冲突检测机制和构建语义映射关系实现局部本体到全局本体的合成。最后利用生成的全局本体,将异构的数据源集成到统一的模式中,并且将基于中介模式的查询转换为基于各局部数据源的模式查询,提供给用户统一的数据查询界面,完成异构数据源的集成。

基于本体的数据集成框架具有以下特点:

(1)局部本体的自动生成:面对来源丰富的异构数据源,为了更有效率地将其解析预处理建立局部本体,系统设计了动态适应统一异构数据源的处理解析接口,采用的是策略模式来实现该接口,实现了对异构数据源分析、预处理和数据提取的统一性,并且利用此接口实现了局部本体的自动建立。该接口的实现关键部分在于从结构化、半结构化以及非结构化的数据文件中提取出构建局部本体的语义信息。由于结构化关系数据库的模式跟局部本体模式最相似,本文在预处理3种类型的数据时,都会先将半结构化数据以及非结构化数据转换成结构化数据,再统一使用结构化数据构建局部本体。通过实现自动构建局部本体,可以解决数据集成时存在的物理冲突。

(2)语义冲突检测机制:语义冲突检测机制实现了数据集成过程中语义冲突的主动发现,并且利用发现的语义冲突完成语义映射关系的构建,完成本体的映射过程。本体映射是指2个本体存在语义级的概念关联,通过语义关联,实现将源本体映射到目标本体,映射最重要的过程就是发现语义冲突。通过本体映射和化解语义冲突,从而消除语义异构。为了实现语义冲突类型的主动发现,本文设计了基于语义树的映射发现策略。映射发现策略分别定制了属性对映射发现策略、概念对映射发现策略以及实例对映射发现策略,从而从不同层次分别解决了表冲突、字段冲突以及记录冲突的主动发现问题,同时定义了基于语义树的映射发现规则来提高映射发现的效率以及准确率。

(3) 语义映射类型: 语义映射类型定义了本体之间的概念、属性以及实例的语义关系,为解决异构数据源间的语义冲突提供依据。为了解决冲突检测机制发现的语义冲突问题,本文定义了应用于电力数据集成的 4 种语义映射类型。表 1 列出了本系统中所支持的语义映射类型。

表1 语义映射类别

			M 74-11
	映射类型	对应异构问题	举例
	Equal	命名冲突(表冲突、字段冲突)	开关: breaker=Equal(switch)
	Contact	表冲突的结构冲突(对象/属性)	节点=Contact(厂站,电压,连接点)
	Union	表冲突的关系冲突(父集/子集)	交流线=Union(模型,量测)
	Algorithm	记录冲突(数值冲突、字段冲突)	电压值=Algorithm(变比)

3 系统关键技术

3.1 基于语义树的多策略映射发现

面对电力系统丰富的数据类型以及海量的数据,实现高效率的映射发现很重要。为全面考虑本体中概念-概念、属性-属性、实例-实例之间的映射关系,本文提出了基于语义树的多策略映射算法,并且在此基础上定义了映射发现过程的规则来提高映射发现的效率。下面详细介绍映射发现的各个子策略,以及基于语义树的映射发现规则。

3.1.1 名称相似度算法

名称相似度算法^[4]用于计算本体中属性对的相似度。本 文中的名称相似度算法基于 Wordnet 语义字典, Wordnet 是 一个义类词典,其中每个节点 s 表示一个词义,节点中保存 了多个同义词或者短语,每个单词或短语又可以存于多个语 义节点中。定义词 w1 和词 w2 的名字相似度为:

$$sim(w_1, w_2) = 1 - t \sqrt{\frac{\alpha - 1}{\alpha} \times \beta \times Dist(w_1, w_2)}$$
 (1)

其中, $\beta = \frac{Dep(w_2)}{Dep(w_1) + Dep(w_2)}$; t 和 α 是可变因子,其中, $\alpha > 2$ 。

Dep(c) 表示任一非根节点概念 C 在语义树中的深度,定义为: Dep(c) = Dep(parent(c)) + 1 (2)

基于义类词典的概念相似度的基本思想是: 2 个单词通过上位关系(hypemym)连接的距离越近,相似度越大; 反之,相似度越小。如果它们在一个节点上,即 s1=s2,则 $sim(w_1, w_2)=1$,如果他们在有限上位层次中没有共同的父节点,则 $sim(w_1, w_2)=0$ 。

3.1.2 概念相似度算法

概念相似度算法^[5]用于计算本体中概念对的相似度。概念定义描述信息包括 2 个方面:表示概念的同义词集和概念的特征集。其中,特征集可以分为功用(function)、组成(part)和属性(attribute)3 个部分。同义词集表示同一个概念的名词的词集。定义概念相似度计算方法:

$$sim(A,B) = \frac{|a \cap b|}{|a \cap b| + \alpha(A,B)|a/b| + (1-\alpha(A,B))|b/a|}$$
(3)

其中,a n b分别表示概念 A n B 的描述集合,包括同义词集和特征集; $a \cap b$ 表示集合 a n b 交集的元素个数;a/b 表示属于集合 a 但是不属于集合 b 的元素个数。比例 a 满足:

$$\alpha(A,B) = \begin{cases} \frac{depth(A)}{depth(A) + depth(B)} & depth(A) \leq depth(B) \\ 1 - \frac{depth(A)}{depth(A) + depth(B)} & depth(A) > depth(B) \end{cases}$$
(4)

其中, depth(A) 表示从概念 A 到根(root)的最短路径距离。 3.1.3 基于语义树的映射发现规则

本体语义模型是一棵概念树^[6],概念树的叶子节点代表 本体中的属性或者实例,其他节点代表本体的概念(类)。因 此,定义如下基于概念语义树的规则来提高映射发现效率。

规则 1 在语义树中,如果 A 节点和 B 节点的父节点 parent(A)和 parent(B)以及子节点 son(A)和 son(B)分别都存在 映射关系,则 A 节点和 B 节点也可能存在映射关系。

规则 2 在语义树中,如果 A 节点和 B 节点的兄弟节点 brother(A)和 brother(B)存在映射关系,则 A 和 B 节点也可能存在映射关系。

規则 3 在语义树中,如果 A 节点和 B 节点是相似的,则 A 节点和 B 节点拥有的实例节点是相似的。

规则 4 在语义树中,如果 A 节点和 B 节点拥有相同的属性节点,则 A 和 B 节点是相似的。

3.2 本体映射算法

本文设计的的本体映射算法综合使用上述各个相似度策略来计算本体之间元素的相似度,主动发现2个本体中的概念、属性以及实例之间的语义冲突,最后输出本体映射类型完成映射发现过程。映射发现的具体算法如下:

输入 局部本体 O₁和 O₂

输出 概念映射表(ClassMap)、属性概念映射表(ProMap)、实例映射表(InstanceMap)

SearchMap(O₁, O₂){

//属性映射算法

 $SearchProMap(O_1,O_2)$ {//匹配本体树上所有的属性节点对,生//成属性映射表

//树深度搜索本体 O1 树上所有的属性节点

PropertyNodeDeepSearch(O1);

//树深度搜索本体 O2 树上所有的属性节点

PropertyNodeDeepSearch (O2);

//使用式(1)名称相似度算法计算 2 个属性节点的相似度

Int sim=NameMatchAlgorithm(proNode1,proNode2);

//如果相似度大于阈值,则将属性节点对记录在属性映射表中

If (sim>w) putProMap(proNode1,proNode2);

//根据规则 1, 更新属性映射表

UpdateProMap();}

//概念映射算法

SearchClassMap(O_1 , O_2) {//匹配本体树上所有概念节点对,生//成概念映射表

//树深度搜索本体 O₁树上所有非叶子节点

ClassNodeDeepSearch(O1);

//树深度搜索本体 O₂树上所有非叶子节点

ClassNodeDeepSearch (O2):

//使用式(3)概念相似度算法计算2个概念节点的相似度

Int sim = ObjectMatchAlgorithm(objcetNode1,objectNode2);

//如果相似度大于阈值,则将属性节点对记录在属性映射表中

 $If (sim > f) \ putClassMap(objcetNode1, objectNode2); \\$

//根据规则 1、2、4 更新属性映射表

UpdateClassMap();}

//实例映射算法

//根据规则 3 生成实例映射表

CreateInstanceMap ();}

测试数据使用电力参数估计数据需求模型里的 PMU 和SCADA 实例数据各一套。PMU 数据中包含概念 16、属性 54、实例 129, SCADA 数据中包含概念 8、属性 32、实例 156。执行本体映射发现算法,统计输出的概念映射表、属性概念映射表、实例映射表所发现的映射对,与实际存在的映射对个数进行比对,结果如图 3 所示。



图 3 实际存在映射与实验发现映射个数比较

采用信息检索领域的查准率作为评价映射算法的主要指标,定义查准率(precision)为:

发现的正确映射对

根据查准率公式,计算得到概念对查准率 pobject=0.75; 概念对查准率 pproperty=0.85; 实例查准率 pinstance=0.78。查准率说明本文定义的映射发现算法可以主动发现数据集成中大部分的语义冲突,但仍有少量冲突没有办法识别。这些冲突主要是因为电力内部自定义了一些简写方式,如记录数据"北花甲线"和"北花J线"表示的同一记录数据,还有名称的缩写:"期望"对应于"Exp",这些冲突的解决还需要领域专家制定语义映射规则。

4 结束语

通过实验分析以及实际应用结果可知,本文设计的映射 发现算法可以高效并准确地发现集成数据的语义冲突,同时 利用定义好的语义映射类型,可以解决数据集成中的大部分 (下转第49页)