面向物联网的传感数据属性语义化标注方法

施昭^{①②*}, 刘阳^①, 曾鹏^①, 于海斌^①^① 中国科学院沈阳自动化研究所工业控制网络与系统研究室, 沈阳 110016^② 中国科学院大学, 北京 100049

* 通信作者. E-mail: shizhao@sia.cn

收稿日期: 2014-11-06; 接受日期: 2015-04-06; 网络出版日期: 2015-05-12

中国科学院战略性先导科技专项 (批准号: XDA06020900) 和国家自然科学基金 (批准号: 61233007) 资助项目

摘要 数据属性在物联网感知数据物理意义表达和智能化处理方面起着非常重要的作用, 可以有效提高数据的使用价值, 是实现异构数据共享和融合、异构资源交互和协同的重要基础. 然而, 随着物联网领域技术的发展, 当前数据属性的描述都是基于具体应用而定制化设计的, 数据属性的结构化较强, 数据很难被跨域理解 and 处理. 针对这一问题, 本文提出一种基于语义网技术的物联网数据属性标注方法. 通过层次本体架构对数据属性进行语义化的描述, 将数据属性从关系型数据库中抽象出来, 使得数据属性独立于具体应用而存在; 在数据访问的过程中, 根据具体应用的需求有针对性地对数据进行属性的语义化标注, 并通过推理技术实现数据属性的扩展. 这种对数据的语义化标注方法有效实现了对数据属性的统一化描述和更为灵活数据模式的构建, 提高了数据物理意义的表达能力和数据的使用价值, 促进了计算机对数据更深入的理解和更智能化的处理, 为物联网领域异构数据的融合共享和异构资源的交互协同提供了有力的技术支撑.

关键词 语义网技术 本体 数据属性 语义化标注 数据处理

1 引言

随着物联网技术的发展和应用, 物联网感知的数据呈现海量性、多源性、多样性等特征^[1], 感知数据的传感器不仅在属性和通信协议上异构^[2], 而且感知的数据在数据格式、属性描述方面也存在着异构的问题^[3,4], 这种异构性不但影响了跨域数据之间的相互理解, 更增加了域间数据处理和应用开发的难度, 很难实现资源的交互协同和数据的共享融合.

在解决设备描述异构问题方面, 文献 [5] 针对智能交通中大量物理设备之间存在的异构性问题, 提出了一种语义化的中间件, 通过语义网技术和智能体技术的结合, 促进设备、服务、人之间的整合和互操作, 提高交通的协同能力和安全性; 文献 [2] 提出了一种针对网络智能设备的语义发现服务 DiscoWoT, 通过对设备属性和提供服务的统一化描述, 实现人和机器对设备功能的语义化发现能力, 促进智能设备的发现、选择和利用; 文献 [6] 针对当前设备存在的异构性, 组织开发了 SSN 本体 (ontology), 通过对传感设备的语义化描述, 提高传感设备间在语义层面的互操作性. 上述文献都是为了解决物理感知设备在描述上异构, 彼此间很难识别和协同操作方面的问题, 都集中在属性或特征的静态描述方面, 而没有考虑反映物理实体状态变化的感知数据的语义异构和实时动态性方面的问题, 而且也没有考虑感知设备和感知数据之间的关联性问题.

引用格式: 施昭, 刘阳, 曾鹏, 等. 面向物联网的传感数据属性语义化标注方法. 中国科学: 信息科学, 2015, 45: 739-751, doi: 10.1360/N112014-00361

感知数据反映物理实体的实时状态和状态的变化, 具有瞬时性和时空关联性等特点, 对其物理意义的理解和智能化的处理都是基于数据属性的描述而实现的. 数据属性是描述数据特征的数据, 是关于数据状况、质量、内容、及其他特征的信息^[7], 为数据提供丰富的特征描述, 是数据定位、发现和处理的关键, 直接决定着数据的物理意义和使用价值. 但是当前对数据属性的描述, 都跟具体应用紧密相关, 数据的物理意义和使用价值都是基于应用而定制化设计的, 不但数据的含义有限, 而且数据属性的结构性强, 很难实现数据属性的扩展和数据的重用. 例如一个室内温度监测系统, 在某个时间点监测到温度值 30°C, 那么在这个系统中 30 这个数值只表示某个时间点上的温度数据, 而不具有其他的物理意义, 例如它表示的位置、监测的对象、影响到的相关人员等等. 随着应用需求的变化, 需要赋予这个数据更多的特征, 这将导致数据结构的更改, 增加系统开发的负担. 另外, 已构建数据结构的变动也会直接影响着上层应用对数据的访问. 这种数据结构和应用竖井式的系统设计模式严重影响了系统开发和数据访问的灵活性.

在数据属性表达方面, 文献 [8,9] 针对当前数据在语义层面上的整合能力问题, 提出了一种实时感知数据的本体化表示方法, 解决语义传感网在语义层面上的互操作问题, 这种方法虽然为数据提供了丰富的含义, 使得感知数据可以被更广泛的应用, 但是该方法只是对数据的含义进行描述, 并没有考虑感知数据动态性和含义描述静态性的结合问题, 同时也没有具体的感知数据标注方法和数据含义描述智能化的扩展. 文献 [10] 提出了一种感知数据含义的标注方法, 在 SWE 的基础上, 通过赋予感知数据时间、空间、主题 3 种元数据, 提高感知数据的描述和含义, 促进感知数据的管理和分析能力, 但这种方法只提及了感知数据时空和主题属性, 缺少更为丰富的其他语义属性, 同时缺少感知数据之间及感知数据与人的关联. 文献 [11~13] 提出了一种面向服务的传感器本体, 通过对数据的本体化描述, 提高了数据含义的理解和重用, 但这个本体是对表达各种物理意义的数据类型进行分类和抽象, 例如刻度类型数据、参数类型数据、物理单位类型数据等, 并不是针对反映物理实体状态及状态变化的动态感知数据而设计的.

因此, 为了有效提高数据属性描述的灵活性和智能性, 增强感知数据的物理意义表达能力和使用价值, 准确反映物理实体的状态及状态的变化, 促进物联网领域跨域异构资源交互协同和数据共享融合的实现, 本文提出了一种基于语义网技术的数据属性标注方法, 通过本体技术对数据属性进行明确化和统一化的语义化描述, 构建数据属性的层次本体架构, 在数据库中将数据属性抽象出来, 使得数据属性独立于具体应用而存在; 通过数据的语义化标注, 在数据访问的过程中根据具体应用的需求, 丰富数据的属性信息, 并通过推理技术实现数据属性的扩展. 这种数据属性描述方法改善了数据模式和数据属性的灵活性, 有效增强了数据物理意义的表达能力和数据的使用价值, 促进了计算机对数据更深入的理解和更智能化的处理, 为物联网领域异构数据的融合共享和异构资源的交互协同提供了有力的技术支撑.

本文第 2 节简要介绍了要用到的预备知识: 语义网技术简介, RDF 和 RDFS, 本体和推理; 第 3 节定义了一种面向数据属性的数据模型和本体的层次架构; 第 4 节基于这种数据模型和本体层次架构, 提出了数据的语义化标注方法和数据属性的扩展; 第 5 节介绍了数据属性存储和访问的方式; 第 6 节对方法性能的实验测试和测试结果进行了分析评估; 第 7 节是结论和进一步的研究展望.

2 预备知识

2.1 语义网技术简介

语义网技术, 由万维网之父 Tim Berners-Lee 在 1998 年提出, 其核心思想是让计算机能够基于人

类的思维和表达方式, 更好地理解 and 处理数据, 自动处理网络上的信息内容, 解决当前网络上基于关键字信息检索准确率低和不完善的问题^[14].

语义网并不是一个单独的网络, 而是当前互联网的扩展, 通过对信息赋予明确的定义, 让计算机程序能够更好地理解信息内容, 基于知识内容和知识结构实现信息的共享和整合, 提高计算机智能化的信息处理能力^[15]. 语义网的层次结构共有 7 层^[16]. 其中, 第 1 层是编码标识层, 包括 Unicode 和 URI, Unicode 是字符集, 是对资源进行编码的国际化标准, URI 表示统一资源标识符, 对资源进行唯一化和统一化的标识; 第 2 层是标记语言层, 是资源描述的标准化语法规则, 包括 XML, XML Schema 和 Namespace, XML 是一种元标记语言, 用来规范化数据描述的语法, XML Schema 用来提供更多的数据类型, 更好地为有效的 XML 文档服务提供数据校验机制, Namespace(命名空间) 用来简化 URI 的书写, 提高数据的可确定性; 第 3 层是资源描述层, 是资源描述的标准化语义规范, 包括 RDF 和 RDFS, RDF 是一种基于图结构的资源描述数据模型, 以“资源 — 属性 — 资源/属性值”三元组的形式描述资源, RDFS 用来定义资源的属性元素, 表示类和属性、值域和定义域在属性上的约束, 保证资源间层次和包蕴关系的准确性; 第 4 层是本体层, 是对资源抽象归类的明确的形式化定义, 用来对资源进行知识化的描述和组织, 在 RDF(S) 基础上描述应用领域的知识、各类资源及资源之间的关系, 揭示更为丰富的语义关系, 是语义网技术的核心; 第 5 层是逻辑 (unifying logic) 层, 用来提供公理和推理规则; 第 6 层是证明 (proof) 层, 是为保证可靠性而提供的一种验证机制; 第 7 层是信任 (trust) 层, 是为保证语义网技术可靠性而建立的信任关系.

2.2 RDF 和 RDFS

资源描述框架 (resource description framework, RDF) 是一种基于图结构的资源描述数据模型, 以“资源 — 属性 — 资源/属性值”三元组的形式表达资源的属性和关系, 是一种标准的元数据语义描述规范^[17,18], 其中资源、属性、资源/属性值分别对应于自然语言中的主语、谓语和宾语. RDF 是在语义层面上表达资源的属性和关系信息, 是实现异构资源交互和数据共享的基础.

RDF Schema (RDFS) 是描述资源类型及其之间相互关系的词汇集, 通过对资源所属类之间的层次和继承等关系的定义, 对主语和宾语的类型进行约束 (RDF 定义的是 object 层的概念, 而 RDFS 定义的是 Class 层的概念), 以保证陈述表述含义的正确性.

2.3 本体

本体是对资源抽象归类的明确的形式化定义, 用来对资源进行知识化的描述和组织. 本体技术, 是语义网技术的核心, 具有很强的逻辑描述和知识表达能力, 通过某个领域的概念和共识, 构建领域词汇的基本术语和关系, 并结合这些术语和关系定义词汇外延的规则, 增强知识表达的准确性和表达能力, 促进计算机对知识更深入的理解和更智能化的处理, 是知识共享和重用实现的基础. 本体的基元主要包括概念、属性、实例和公理等内容, 以四元组 $O=(C, R, I, A)$ 来表示, 其中, C 表示概念或类的集合, 用来描述资源的抽象和归类; R 表示概念间关系的集合, 用来描述概念间的各种关系, 包括层次关系、逻辑关系、关联关系、操作关系、依赖关系等; I 表示所属概念实例的集合, 用来描述具体的对象和个体; A 表示公理的集合, 用来描述永真命题, 是有效性、一致性检测的基础.

2.4 推理

推理是基于一些设定好的规则和关系, 从已存在的陈述 (或称父陈述) 中, 动态产生新的陈述 (或

图 1 一个 RDF 三元组陈述
Figure 1 An example of a RDF triple

称子陈述), 不断丰富本体库中的知识, 以保证本体库的完备性和访问查询的效率, 扩展本体不支持的表达能力. 另外, 推理技术还被用来对异常现象进行分析判断, 通过预先设定的规则条件, 触发相应的事件或行为, 并生成通知反馈给用户.

当前常见的语义网推理机主要有: Jena, Racer, Pellet 和 Euler. 其中, Jena^[17] 是最常用的推理机, 基于 Java 编程语言设计和实现, 相比于其他推理机更易于使用, 但 Jena 描述逻辑的推理功能 OWL-DL 并不完备, 在使用多条规则推理一个决策时, 会出现不一致的情况, 而且 Jena 的扩展性有限; Racer^[19] 推理机推理功能非常完备, 而且具有三元组回滚功能, 但 Racer 的规则和查询语言是定制化设计的, 不具有通用性, 而且在使用上也非常复杂, 经常给系统带来高负荷的工作负载; Pellet^[20] 推理机一般作为 Jena 的外部推理机来使用, 具有更加完备的推理功能, 与其他推理机之间具有更好的兼容性, 但 Pellet 不像 Racer 那样允许三元组的回滚; Euler¹⁾ 是最轻量级的推理机, 易于使用, 扩展性好, 而且推理速度最快, 但 Euler 同 Pellet 一样, 不具备三元组的回滚功能, 而且 Euler 是一种无记忆的推理机.

3 数据属性的描述方法

3.1 面向数据属性的数据模型

数据属性是描述数据特征的数据. 面向数据属性的数据模型, 是通过 RDF 和本体技术, 对数据属性进行语义化描述, 进而提高数据的物理意义和使用价值. 在数据模型中, 数据属性通过统一资源标识符 (URI) 来表示, 数据属性的描述通过三元组的形式来表示, 如图 1 所示. 为了简便, 我们对 URI 中的前缀 `http://www.owl-ontologies.com/sia.owl#` 用缩略词 SIA 来表示, 则 `SIA:printer_719` 和 `SIA:A719` 分别表示概念 `SIA:Device` 和概念 `SIA:Location` 的实例, `SIA:is_location` 表示两个实例的关系, 图 1 为 RDF 三元组陈述的一个例子.

在数据属性描述和关系的表达上, RDF 数据模型是一种图模型的数据结构, 与关系型数据库相比, 模型的结构化不强, 不受数据组织结构的限制, 在表达能力上更加灵活, 不但可以表达更加复杂的关系, 而且能根据应用需求的变化对数据属性关系进行延伸, 自动推理得到数据属性间新的关联关系. 如图 2 所示, 基于 `has_own` 本体属性, 可以推导出 `SIA:computer_PC-20130824CKOD` 和 `SIA:LCD.20122244` 属于相同的拥有者 `SIA:David`, 它们之间存在关联关系, 这样在制定对设备的节能控

1) De R J. Euler Yet Another Proof Engine. 2005. <http://eulersharp.sourceforge.net/>.

图 2 数据属性间关系的延伸

Figure 2 The expansion of relationships between data attributes

制策略规则时, 就可以依据两者用电行为是否一致的判断, 识别出能耗中的浪费部分, 及时做出节能的操作, 为控制策略提供更加智能化的依据。

3.2 本体的层次架构

当前本体的构建, 都是针对一个特定的领域, 基于提取的领域特征, 构建数据属性描述的模型。本体也支持其他本体的导入, 通过 owl:import 语句来导入一个或更多的其他本体, 进而实现领域本体的扩展。

本体的层次架构如图 3 所示。Layer 1 包含了特定领域的所有领域术语, 定义了领域本体中用到的所有词汇, 是实现跨域本体对准、信息整合的基础。Layer 2 基于领域术语对数据属性和属性关系进行概念化的抽象和归类, 构建数据属性和属性关系的抽象模型。Layer 3 基于抽象模型, 对具体的数据属性进行实例化的描述。

Layer 1 领域本体词汇库。领域本体词汇库, 在本体层次架构的最上层, 包含了特定领域的所有领域术语, 构成了特定领域的词汇集。这个词汇集是领域本体中概念、属性、实例等类型词汇的集合, 用以表达本体的领域特征。同一事物, 在不同领域的本体里会有不同的词汇描述和词汇表达, 这就是所谓的语义异构^[21,22], 例如我们日常生活中的词汇“水”, 在化学反应中则被表示为“H₂O”。因此, 为了实现跨域的资源交互和信息共享, 就需要在这一层对不同本体间的词汇集进行对准, 将表示同一事物的词汇进行等价标识, 这样就可以实现不同本体对相同资源和信息的识别。

Layer 2 领域抽象模型。领域抽象模型是基于领域特征, 对领域数据属性和属性关系概念化的抽象和归类, 并建立概念的本体属性描述和概念间的关系。数据属性和属性关系, 是对数据物理意义的表达, 对数据使用价值的体现。这一层根据具体领域的特征, 定义了本体的概念、属性和关系, 例如在楼宇节能应用领域, 定义了 Devices, Persons, Instruments, Locations, Surroundings, Data 等与节能有关的数据属性概念, 本体属性定义了 Operation, Parameter, Measurement, Capability, Physical 等对实体和数据特征的描述, 其中, Operation 表示对仪器操作属性的描述, Parameter 表示对仪器参数属性的描述, Measurement 表示对仪器测量属性的描述, Capability 表示对仪器性能属性的描述, Physical 表示对数据及实体物理属性的描述; 数据和实体关系包括 Association, Logic, Operation, Hierarchy, Dependent 等关系。

Layer 3 领域实例模型。领域实例模型是在领域抽象模型的基础上, 根据具体数据和资源的描

图 3 本体层次架构

Figure 3 The ontology layer architecture

图 4 概念 Instruments 实例化的一个例子

Figure 4 An example of an instance of the concept instruments

述和特征, 对抽象的概念进行具体的实例化。例如, 在楼宇节能应用领域中, 我们在 Layer 2 层已经定义了 Device, Person, Instruments, Locations 等概念, 而 SIA:TemSensor_37_270284F0 是概念 SIA:Instruments 的一个实例, 对实例 SIA:TemSensor_37_270284F0 的具体描述如图 4 所示。

图 5 数据的语义化标注过程

Figure 5 The process of the semantic annotation for data

4 数据属性的标注方法和扩展

4.1 数据的语义化标注

数据的语义化标注, 是根据具体的应用需求, 使用面向数据属性的数据模型, 在数据调用的过程中赋予数据丰富的语义属性信息, 提高数据的物理意义和使用价值. 语义化标注的过程如图 5 所示.

数据属性是数据语义化标注的内容, 体现着数据的物理意义. 数据的属性越丰富, 则数据物理意义的表达能力就越强, 可使用的价值就越大. 根据物联网传感数据的特点, 我们将数据属性分为动态属性和静态属性. 动态属性是随着数据的不断采集而实时变化的属性, 例如数据的时间属性 when, 仪器的状态属性 Status; 静态属性几乎是不随着数据的不断采集而实时变化的属性, 例如数据所在的物理位置属性 where, 数据关联的人员属性 who 等. 为了更好地描述数据的含义, 我们定义了用于描述数据含义的几个基本维度属性, 如下: who, where, what, when, do what (object), how and others. 其中, who 表示与数据关联的人员信息, where 表示与数据关联的物理位置信息, what 表示数据的类型信息, when 表示数据的时间信息, do what (object) 表示数据对象的信息, how 表示数据采集方式的信息, others 表示数据其他方面的信息. 例如, 一个房间的温度值为 29°C, 则 29 的 who 表示使用这个房间的具体的人, where 表示这个房间, what 表示温度值, when 表示温度采集的具体时间, do what

图 6 基于具体应用的语义化标注

Figure 6 An example of application-specific semantic annotating for data

(object) 表示是这个房间的温度, how 表示是非接触式的感知方式. 在具体的应用中, 如果房间使用者设定的适宜温度为 25°C , 则通过温度值的比较和这些特征的描述, 可以实现房间温度的自动调节, 以满足房间使用者的舒适度.

另外, 基于具体的应用需求, 可以有针对性地对数据进行语义化标注, 例如一个室内温度传感器, 它的属性描述如图 6 左图所示. 在设备是否正常的自动监测应用中, 并不需要关心传感器的所有属性, 只需要判断设备的采集值是否在有效测量值范围内, 据此判断设备是否运行正常, 并自动通知系统管理者进行实时的维护. 因此, 对温度值进行语义标注的属性如图 6 右图所示. 这种标注的方式将数据属性和数据分离, 实现了数据属性更为方便和灵活的表达能力.

4.2 数据属性的扩展

数据属性的扩展, 是通过资源关系, 在数据已有属性的基础上, 推导出新的数据属性, 实现对数据属性的延伸, 为数据提供更丰富的属性描述, 以满足特定应用数据处理方面的需求.

资源关系是资源基于某个特征而形成的一种关联, 通过关系的表示和处理, 提高数据处理的智能性. 资源关系主要包括简单关系、关联关系、逻辑关系、依赖关系、时序关系, 其中, 简单关系表示资源间紧密程度的相互关系, 关联关系表示资源的取值之间存在某种规律性的相互关系, 逻辑关系表示资源间因果推导先后顺序的相互关系, 依赖关系表示资源间依附存在的相互关系, 时序关系表示资源在某个时间段内是否一起发生的相互关系.

例如, 室内温度传感器 SIA:TemSensor_37_270284F0 感知到当前房间 A719 的温度值为 30°C , 如图 7 所示, 30 这个数值通过温度感知器 SIA:TemSensor_37_270284F0 获得所在位置 where、监测对象 Object 等属性特征, 但是不能直接获得关联的人员属性 who. 而在领域本体中, SIA:TemSensor_37_270284F0 实例和 SIA:John 实例都有相同的本体属性 is_location 和相同的属性值 SIA:Room_A719, 由此可以推理出 SIA:John 和 SIA:Room_A719 存在着简单关系, 进一步推导出温度值 30 的 who 属性值为 SIA:John. 接下来再根据 John 的喜好温度, 在数据处理模块中进行判断分析, 可是实现温度控制操作的触发和自动调节.

图 7 数据属性的扩展

Figure 7 The expansion of data attributes

5 数据属性的存储和访问

5.1 数据属性的存储

为了更好地实现对数据属性的访问和处理, 需要一个更加灵活和高效的存储方式. 本体的语义化存储, 是对 RDF 类型数据模型的存储和管理, 当前主要以基于关系型数据库的 RDF 存储和基于图模型的 RDF 存储两种存储方式为主, 本论文中提到的数据属性存储方式采用的是基于关系型数据库的存储.

基于关系型数据库的 RDF 存储方式, 使用数据表来存储主语、谓语、宾语构成的三元组. 这种存储方式的特点是所有的 RDF 三元组都存储在同一个表中, 在访问时并不需要了解表结构, 易于查询, 查询的适应性强, 更适用于情景环境, 有更好的扩展性和灵活性. 但是, 其缺点在于, 与同数据量的关系型数据库相比, 语义三元组是关系型行数的 20 倍, RDF 格式的数据量大小是 CSV 数据量的 18 倍^[23]. 语义索引的代价很大.

基于图模型的 RDF 存储方式, 是一种能够更加直接地对 RDF 数据的结构进行建模的数据结构, 能够提供一种高效的方式来定位与之共享相同资源的陈述. 常见的基于图的 RDF 存储的实现使用了相互链接的陈述列表, 这样每条共享相同资源 (这些资源可能作为陈述主语、谓语或者宾语) 的陈述就被安置到了一个连续的链接列表中, 或者使用特殊的索引数据结构来链接在 RDF 图中相邻的陈述. 这样就提供了一种机制, 可以快速遍历包含某一特定资源的所有陈述. 这种设计的特性与 RDB 方法中的列索引和附加表十分类似, 然而, 基于图的实现方法有更为一般的用途, 并且并不需要用到与数据一同存储的高级知识.

5.2 数据属性的访问

在本体库建立好以后, 需要从本体库中访问或检索需要的数据属性信息. 其中, SPARQL 是一种专门用来访问 RDF 三元组的本体查询语言. SPARQL 的查询类似于 SQL 对关系型数据库的查询, 通过条件的设置访问相关的信息, 但不需要了解表结构. 基于设定的条件, SPARQL 可以用来进行

图 8 语义查询性能测试结果

Figure 8 The results of performance test by semantic querying

简单的查询和复杂的查询, 甚至可以进行具备同属性值实例集合的查询, 例如, 一个简单查询, 查询传感器 “Sensors_zhl_lcd” 的所属者, 通过查询语句 ‘select DISTINCT?name where {sia:Sensor_zhl_lcd sia:is_owned ?name }’, 就可以从 RDF 三元组中直接得到当前传感器监测设备的所属者 “ZhangHl”; 一个多条件的复杂查询, 查询监测打印机 “printer_719” 的传感器所在的位置, 通过查询语句 ‘select DISTINCT?name where { ?sub sia:is_monitoring sia:printer_719. ?sub sia:is_location ?name }’, 就可以先从 RDF 三元组中确定打印机 “printer_719” 对应的传感器, 再得到该传感器所在的位置是 “A719”, 实现多条件的复杂查询。

6 实验与评估

数据的语义化标注, 在数据的处理方面起着非常重要的作用, 直接影响着数据物理意义的表达能力和数据的使用价值。而领域本体的查询访问时间, 是语义化标注性能的关键影响因素, 因此在这一章节, 我们对不同三元组行数量的查询访问时间做了测试和比较, 如图 8 所示, 其中横坐标表示三元组的行数量, 实线变化曲线表示查询访问时间的变化情况, 虚线变化曲线表示数据集大小的变化情况。

性能测试实验编程环境采用 Java 语言, myeclipse 10 编程环境, 数据的存储采用 Oracle 11g r2 作为动态实时数据的存储载体, 采用 MySQL 5.0.18 作为本体库的存储载体; 实验系统硬件采用 Intel(R) Core(TM) i7-2600 处理器, 3.4 GHz 主频 CPU, 4 G 内存, 1 TB 硬盘和 Window 7 操作系统。

在性能测试实验中, 我们将部分的楼宇设备能耗数据转化成 RDF 三元组的形式, 以最简单的形式, 构建 RDF 三元组数据集, 对于不同 RDF 三元组行数数据集的数据量大小和查询访问时间进行测试和比较, 为语义化标注方法的可行性的验证和数据存储方式的设计提供有力的依据。本实验采用 SPARQL 本体查询语言进行了简单属性值的查询。所有的测试都执行了 30 次, 测试结果为 30 次的平均值。

从测试结果中可以看出, 本体查询访问时间, 随着三元组数据量的变化呈现线性的增长, 当三元组行数达到 10^7 量级时, 数据集大小为 1.1 GB, 查询访问时间将近 3 s, 而在我们的楼宇设备能耗监测实验平台中, 能耗量采集频率为 1 s, 采集节点为 100 个, 而基于这样的实验参数, 两周采集的感知数据就达到了 3×10^7 行, 如果按照文献 [23] 所述将关系型数据库中的数据转成三元组的描述方式, 则 RDF 格式的总行数将达到 6×10^8 , 按照本实验测试结果的变化规律, 则查询时间将达到 150 s, 这不但

严重影响系统对于异常现象反应的即时性, 而且感知数据的数据量会随着时间的不断的增大, 系统性能也会越来越慢. 因此, 在数据含义的表达能力和系统数据处理性能之间的平衡方面, 数据存储方式的设计是一个关键性的问题, 本文所设计的数据存储方式, 是用关系型数据库存储数据的动态属性数值, 例如表示数据时间属性的 when 数值, 表示数据具体数值的 Value; 用 MySQL 存储数据的静态属性数值, 例如 where, who, what 等属性的数值. 这种设计方式, 将传统数据存储中的一些静态属性描述提取到本体库中存储, 不但减少了传统数据库存储的复杂性和冗余性, 而且增强了数据属性的扩展性和灵活性, 实验中对 100 个节点的数据属性和设备属性描述, 以 RDF 三元组形式构建后生成的行数只达到 10^5 量级, 而感知数据还是以传统的关系型数据库形式进行存储, 在对感知数据进行访问和处理时, 通过本体库的属性描述赋予感知数据丰富的物理意义, 就可以提高数据应用的广泛性和数据处理的智能性. 这种关系型数据库和本体库相结合的数据存储方式, 验证了通过语义化标注方法, 提高数据智能性的可行性.

7 结论

这篇论文介绍了一种基于语义网技术的物联网数据属性的标注方法, 通过 RDF 和本体技术构建面向数据属性的数据模型, 该模型对数据属性进行了明确化、统一化的语义化描述, 使得数据属性从特定的应用中独立出来, 提高了数据模式的灵活性. 在数据调用的过程中, 通过对数据的语义化标注, 赋予数据更丰富的属性含义, 提高数据物理意义的表达能力和使用价值, 促进了计算机对数据更深入的理解和更智能化的处理, 为物联网领域异构数据的融合共享和异构资源的交互协同提供了有力的技术支撑.

在物联网领域, 语义网技术多用在资源描述方面, 用以提高资源设备发现、设计和调试的自动化能力, 而用在数据处理方面的研究还比较少. 在数据处理的智能化方面, 数据挖掘技术是通过从海量的数据中发现数据间的关系而提高数据的使用价值, 而语义网技术是通过表示和处理关系来提高数据的使用价值, 两者虽然在研究方向上不一致, 但是否能够结合而产生出新的技术, 还有待于进一步的研究. 另外, 在本体库的访问过程中, 不同复杂度的语义查询, 例如上述内容中提到的简单查询和多条件复杂查询, 查询响应时间有多大差别, 查询响应时间的差别是否跟 RDF 三元组的行数有关, 这些都有待于进行进一步的研究.

参考文献

- 1 Ganz F, Barnaghi P, Carrez F. Information abstraction for heterogeneous real world internet data. *IEEE Sens J*, 2013, 10: 3793–3805
- 2 Mayer S, Guinard D. An extensible discovery service for smart things. In: *Proceedings of the 2nd International Workshop on Web of Things*, New York, 2011. 1–6
- 3 Ma Z G, Yang Y, Sebe N, et al. Knowledge adaptation with partially shared features for event detection using few exemplars. *IEEE Trans Pattern Anal Mach Intell*, 2014, 36: 1789–1802
- 4 Mao J L, Jia X Q, Liu H Q. Research on semantic architecture and semantic technology of IoT. *J Inf Commun Technol*, 2014, 5: 26–31 [毛峻岭, 贾雪琴, 刘红旗. 物联网语义架构和语义关键技术研究. *信息通信技术*, 2014, 5: 26–31]
- 5 Terziyan V, Kaykova O, Zhovtobryukh D. UbiRoad: semantic middleware for context-aware smart road environments. In: *Proceedings of 5th International Conference on Internet & Web Applications & Services*, Barcelona, 2010. 295–302
- 6 Compton M, Barnaghi P, Bermudez L, et al. The SSN ontology of the W3C semantic sensor network incubator group. *Web Semant Sci Serv Agent World Wide Web*, 2012, 17: 25–32

- 7 Mize J, Habermann R T. Automating metadata for dynamic datasets. In: Proceedings of IEEE Conference on the OCEANS, 2010. 1–6
- 8 Henson C A, Neuhaus H, Sheth A P, et al. An ontological representation of time series observations on the semantic sensor web. In: Proceedings of the 1st International Workshop on the Semantic Sensor Web, Heraklion, 2009. 79–93
- 9 Henson C A, Pschorr J K, Sheth A P, et al. SemSOS: semantic sensor observation service. In: Proceedings of the International Symposium on Collaborative Technologies & Systems, Baltimore, 2009. 44–53
- 10 Sheth A, Henson C, Sahoo S S. Semantic sensor web. IEEE Internet Comput, 2008, 12: 78–83
- 11 Kim J, Kwon H, Kim D, et al. Building a service-oriented ontology for wireless sensor networks. In: Proceedings of the 7th IEEE/ACIS International Conference on Computer & Information Science, Portland, 2008. 649–654
- 12 Eid M, Liscano R, Saddik A E. A novel ontology for sensor networks data. In: Proceedings of IEEE International Conference on Computational Intelligence for Measurement Systems & Applications, LaCoruna, 2006. 75–79
- 13 Eid M, Liscano R, Saddik A E. A universal ontology for sensor networks data. In: Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems & Applications, Ostuni, 2007. 59–62
- 14 Lee T B, Hendler J, Lassila O. The semantic web. Sci Am, 2001, 5: 35–43
- 15 Janev V, Vranes S. Applicability assessment of semantic web technologies. Inform Process Manag, 2011, 47: 507–517
- 16 McGuinness D L, Van Harmelen F. OWL web ontology language overview. W3C Recommendation, 2004, 10: 1–19
- 17 Carroll J J, Dickinson I, Dollin C, et al. Jena: implementing the semantic web recommendations. In: Proceedings of the 13th International World Wide Web Conference on Alternate Track Papers & Posters, New York, 2004. 74–83
- 18 Lassila O, Swick R R. Resource description framework (RDF) model and syntax specification. W3C Recommendation, 1999
- 19 Haarslev V, Möller R. Description of the RACER system and its applications. Descrip Logic, 2001, 49: 132–141
- 20 Sirin E, Parsia B, Grau B C, et al. Pellet: a practical OWL-DL reasoner. Web Semant Sci Serv Agent World Wide Web, 2007, 5: 51–53
- 21 Huang Y, Luo X G, Guo M Q, et al. Semantic-oriented metadata management model. In: Proceedings of Conference on Environmental Science and Information Application Technology, Wuhan, 2009. 740–743
- 22 Rueda C, Bermudez L, Fredericks J. The MMI ontology registry and repository: a portal for marine metadata interoperability. In: Proceedings of the Oceans, Mts/IEEE Biloxi-Marine Technology for Our Future: Global & Local Challenges, Biloxi, 2009. 1–6
- 23 Lopez V, Kotoulas S, Sbodio M L, et al. QuerioCity: a linked data platform for urban information management. In: Proceedings of ISWC, Boston, 2012. 148–163

Semantic annotating method of sensing data attributes for data processing in IoT

SHI Zhao^{1,2*}, LIU Yang¹, ZENG Peng¹ & YU HaiBin¹

1 Department of Industrial Control Network and System, Shenyang Institute of Automation Chinese Academy of Sciences, Shenyang 110016, China;

2 University of Chinese Academy of Sciences, Beijing 100049, China

*E-mail: shizhao@sia.cn

Abstract Data attribute description has a vital role in the physical significance and intelligent processing of sensing data and improves the usefulness of data effectively. It is an important basis of heterogeneous data sharing and integrating and heterogeneous resource interaction. However, with the development of IoT technology, established descriptions for sensing data attributes are specialized for specific purposes. The physical significance of sensing data is designed for specific applications. The expression of data significance is limited and the structure of data attribute descriptions is strong, which is difficult to achieve the extension of data attributes and reuse of data. In cross-domain applications, this hinders further data processing, which is strongly required in the IoT. To overcome these problems, this paper presents an ontology-based data attribute description method for

sensing data processing. In this method, the approach of semantic description for data attributes with a layered ontology architecture, which extracts the data attribute information from the database storing the sensing data, allows the data attributes to be independent of specific applications. In the process of data access, the sensing data can be semantically annotated with data attributes for the requirements of specific applications. Further, extension of the data attributes can be achieved by reasoning technologies. This manner of semantic annotation for data unifies the semantic description of the data attributes and creates flexible data schemata. This method can effectively improve the physical significance and useful value of sensing data and can provide computers a more in-depth understanding and facilitate more intelligent data processing of the sensing data. The sharing and integrating of heterogeneous data and interacting of heterogeneous resources can be implemented based on this method.

Keywords semantic web technology, ontology, data attribute description, semantic annotation, data processing

SHI Zhao was born in 1983. He is currently working towards a Ph.D. degree in mechatronics at the Laboratory of Industrial Control Networks and Systems, Shenyang Institute of Automation, Chinese Academy of Sciences. His research includes semantic sensor data processing for IoT and building energy management.

LIU Yang was born in 1981. She received a Ph.D. degree from Northeastern University, Shenyang, in 2011. Currently, she is an associate professor in the Shenyang Institute of Automation, Chinese Academy of Sciences. Her research interests include industrial IOT and semantic sensor data processing.

ZENG Peng was born in 1976. He received his Ph.D. degree in mechatronics from Shenyang Institute of Automation, Chinese Academy of Sciences in 2005. He is currently a professor and the Director of the Laboratory of Industrial Control Networks and Systems, Shenyang Institute of Automation, Chinese Academy of Sciences. He also holds the position of Member, Expert

Group of IEC TC65 WG16. His research interests include industrial automation and wireless networks for such automation.

YU HaiBin was born in 1964. He received a Ph.D. degree in control theory and control engineering in industrial automation from Northeastern University in 1997. He is currently a professor and the President of Shenyang Institute of Automation, Chinese Academy of Sciences. He also holds the positions of Member, Expert Group of Advanced Manufacturing in 863 Plan and Expert,

National Science & Technology Award Evaluating Committee. His research field includes industrial automation.