昆明学院学报　2018，40（3）：70～74　ISSN 1674－5639　CN 53－1211/G4
Journal of Kunming University

物联网多源异构数据的自动语义标注方法研究

韩黎晶，袁凌云*
（云南师范大学 信息学院，云南 昆明 650500）

摘要：针对各类物联网设备因原始感知数据缺乏明确含义，以及表现形式异构而导致的物联网信息资源之间难以实现交互协同和数据融合共享等问题，提出一种面向物联网多源异构数据的自动语义标注方法. 首先建立物联网领域应用本体模型，然后在此基础上给出测量对象的语义描述框架及自动语义标注架构，同时研究实现标注的核心技术，以此增强异构数据资源之间的互操作性，并为上层的语义服务及物联网多设备联动提供支持.

关键词：物联网；多源异构数据；语义标注；本体
中图分类号：TP391.1　**文献标识码**：A　**文章编号**：1674－5639（2018）03－0070－05
DOI：10.14091/j.cnki.kmxyxb.2018.03.013

Study on Automatic Semantic Annotation of Multi-source Heterogeneous Data in Internet of Things

HAN Lijing，YUAN Lingyun*
（College of Information，Yunnan Normal University，Kunming，Yunnan，China　650500）

Abstract：It's because the raw perception data coming from different IoT devices is lack of clear meaning and with different manifestations so that increases the difficulty of interaction，integration and sharing between the information resources. So an automatic semantic annotation method for multi-source heterogeneous data of Internet of Things has been put forward. The construction process of IoT domain application ontology model is expounded，then based on which，a semantic description framework of monitoring object and an automatic semantic annotation architecture are designed. At the same time，this research has conducted some thorough discussion about the core technology to achieve the automatic semantic annotation，thereby to enhance the interoperability between heterogeneous data resources and to provide support for the upper semantic service and the interconnection between multiple devices.

Key words：Internet of Things（IoT）；multi-source heterogeneous data；semantic annotation；ontology

感知设备为物联网应用提供信息来源，是物联网系统进行信息交互和协同的基础. 由于物联网技术的多样化应用及飞速发展，数以万计不同类型的物联网设备在多个领域的各种应用和服务中产生并交互大量的数据. 而以传感器和 RFID 为核心代表的前端感知设备所获取到的原始感知数据信息呈现出明显的多源性、异构性以及高度冗余性等特征[1]，为物联网信息资源之间的交互协同和数据的融合共享处理等操作带来了极大的困难和挑战. 除此之外，物联网用户对各类物联网信息服务的要求也在不断提高，诸如如何屏蔽数据异构性，解决数据孤岛问题，从而实现用户对多设备联动以及多数据融合的需求是物联网研究中亟待解决的问题[2]. 针对此类问题，将日益成熟的语义标注技术作为突破口，能有效地为物联网发展过程中所面临的上述瓶颈提供一种新的解决思路.

语义标注是解决物联网异构行之有效的技术，它为设备所获取到的数据资源提供结构一致且明确的语义描述，有利于让物联网实体设备之间更好地理解彼此所产生的数据信息含义，从而提高数据的

收稿日期：2017－12－01
基金项目：国家自然科学基金资助项目（61561055）；教育部人文社会科学研究青年基金资助项目（13YJCZH233）.
作者简介：韩黎晶（1995—），女，云南曲靖人，硕士研究生，主要从事物联网技术及其应用研究.
*通讯作者：袁凌云（1980—），女，云南昭通人，副教授，博士，主要从事物联网、传感器网络技术及其应用研究，E-mail：blues520@sina.com.

利用率. Terziyan 等[3]设计了一款将智能体技术与语义技术有效结合的语义中间件（Semantic Middleware），通过它来整合各项应用，解决智能交通中各类异构物联网设备之间的互操作问题，以此提高交通的协同监管效力及服务质量. 时念云等[4]将待标注文档中的特征词汇提取出来，通过解析得出这些词汇与本体中概念的对应情况，再据此进一步建立出两者之间的映射关系. 但由于中文语义的复杂性，相同词汇在不同语境中含义可能不同，因此该方法在对于中文资源的语义标注上准确率略低. 王浩然等[5]提出了一种将本体构建与元数据标注联系在一起的方法，该方法首先利用 XML Schema 构建知识本体，并建立 XML 结构与本体概念的映射关系，最后利用映射关系来实现对 XML 文档内容的自动语义标注. 虽然该方法应用到实际项目中表现出不错的效果，但因其依据 XML Schema 所自动构建的本体中并没有包含一个领域完备的知识与概念，因此就标注效果的完善度和准确度而言还差强人意. Wang 等[6]为物联网领域中的知识表示设计了一个轻量级的语义描述模型，在构建本体时考虑了轻量级和完整度之间的权衡，通过将构建好的室内定位本体和信息描述本体结合，形成链接数据，使用者可以通过在其基础上链接其他本体和关联数据来对此语义描述模型进行扩充，使其能够提供更完善的语义描述信息. Liu 等[7]使用语义技术来解决分散、分级和异构的物联网设备信息之间的互操作问题，提出了一种面向设备的自动标注方法，通过对信息抽取、文本分类、属性信息划分、语义标签选择以及信息融合等算法的研究与应用，达到对设备信息进行自动标注的效果. 施昭等[8]利用语义技术实现了对物联网数据属性的标注. 其首先搭建出本体层次模型，在此基础上再对物联网的数据属性添加标准的语义描述，之后将数据属性从关系型数据库中抽象出来，使数据属性独立于具体应用而存在. 该方法实现了对数据属性的统一化描述和更为灵活的数据模式构建，有效提高了数据物理意义的表达能力和数据使用价值.

鉴于目前相关研究大多集中于对单一设备的属性描述或同类感知信息的语义标注，并没有考虑多个设备的协同作业，以至于弱化了“物物相连”的理念. 因此，本文在现有研究的基础上提出一种自动语义标注方法，旨在解决物联网领域中不同设备来源的物联网原始数据自动语义标注及统一化描述问题. 本研究将以一整个测量对象（如智慧教室）为单位，把处于该应用场景内对此测量对象各参数进行监测的所有设备都看作为其中的信息感知实体，通过系统的形式将来自于不同信息感知实体设备的各类数据进行自动化语义描述与标注，最后将这些不同来源的标注信息整合为一体，以此作为对整个测量对象各方面信息的完整语义描述.

1　面向物联网领域的应用本体构建

作为自动语义标注的核心技术，一个功能友好的物联网领域本体是实现面向物联网数据自动语义标注的基础和关键步骤. 考虑到 W3C 传感网工作组（SSN-XG）所研发的 SSN 本体[9]在传感器功能、属性以及传感器观测值等方面具有比较完备的描述，本文参照该本体的构建形式与内容，在遵循正确性、一致性、可扩展性和有效性等原则的前提下，进行物联网领域应用本体的构建.

使用 W3C 推出的 OWL 作为本体描述语言来构建本体模型，选用 Protégé 5.0 作为本体建模工具. 在本体的建立过程中，首先需收集物联网领域的相关知识及概念来确定本体的类和属性，然后在本体中建立这些类及属性之间的关系，最后根据相关领域专家的意见和已经比较完善的本体模型来形成该本体的逻辑结构及具体模型. 遵循以上步骤，在构建好一个物联网领域本体的基础上，增加某一特定应用场景下的特殊实例，构建出一个以“感知设备 - 数据资源 - 时空环境”为核心的应用本体，其层次结构如下图 1 所示.

2　物联网数据自动语义标注方法

自动语义标注是一个为数据添加概念实例、数据属性和对象属性的过程[10]，这个过程将包含物联网设备原始感知数据的 XML 文档作为输入内容，把经过标注之后结构化的、富有明确含义的信息作为输出结果. 通过系统的形式将这个过程具象化地表现出来，下面就系统实现的相关思路及主要技术进行说明.

2.1　系统功能设计

使用 Java 作为编程语言，采用 MyEclipse 10 作为集成开发平台. 从宏观的角度看待整个系统，其数据流向和处理过程如下图 2 所示.

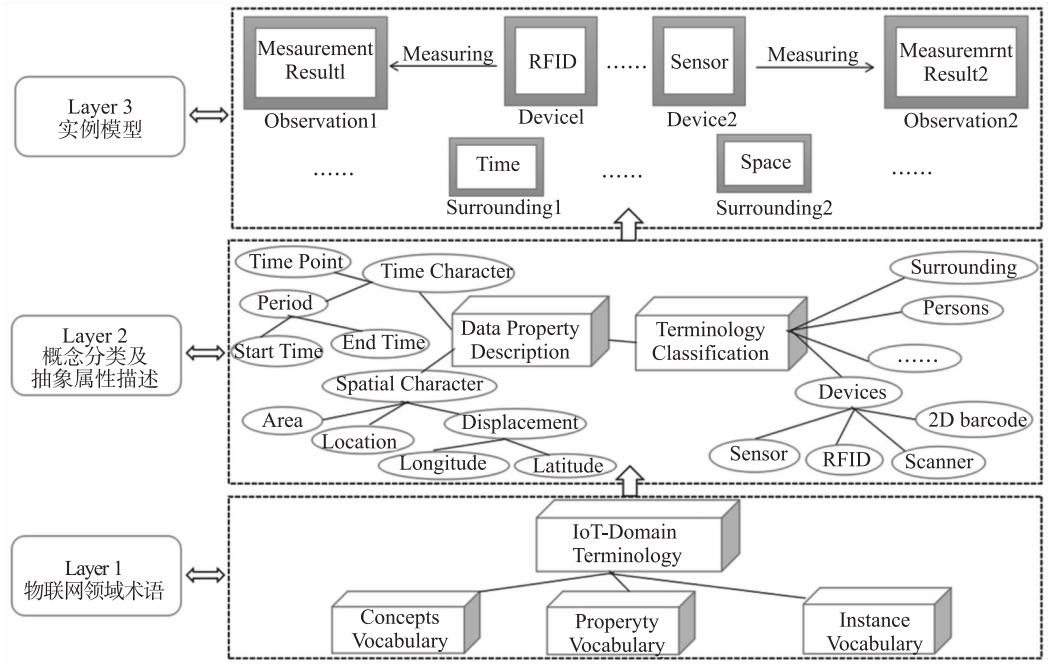

图1 物联网领域应用本体模型

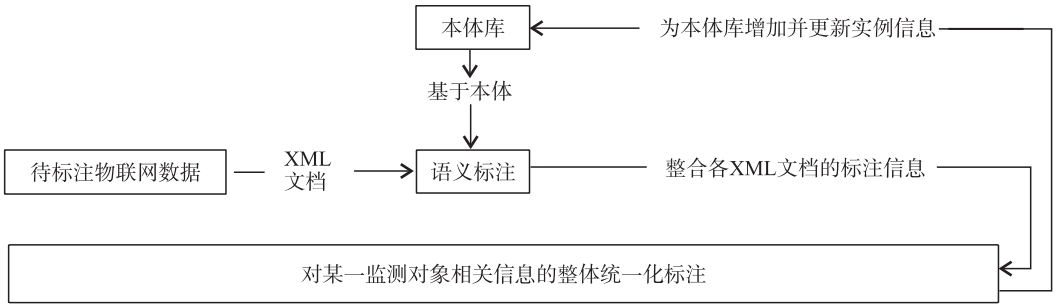

图2 物联网多源异构数据标注处理系统数据流向示意图

2.2 测量对象语义描述框架

本文面向 XML 文档，对其中的物联网原始数据进行语义标注. XML 文档本身只能表达数据的语法和结构，而不能表示形式化的语义，隐藏在 XML 文档中的语义信息以及 XML 的标签内容仅对人们来说有较大意义，而计算机却难以理解. 因此，通过特定的方法将 XML 文档中数据的显式及隐式语义信息抽取出来，并进行形式化标注，对于数据资源的最大化利用具有重大意义. 本文以某一测量对象为单位，对其所包含的各类感知数据属性信息进行标注，语义描述框架如下图 3 所示.

语义描述框架从“设备属性”“时空属性”“测量值属性”3 个方面对测量对象的相关数据信息进行描述. 其中设备属性包括对测量对象进行感知的物联网设备的标识 ID、名称及设备类型等信息. 时空属性包含测量值的时间与空间属性：时间属性包括时间戳（设备获取数据的时间点）和有效时间范围两部分；空间属性描述的是设备实体的地理位置、经度、纬度 3 个属性. 对于时空属性的详细描述有助于之后基于时空相关性的数据融合处理. 测量值属性则包含监测数据的类别（如湿度）、数值及单位.

2.3 自动语义标注架构

总的来说，可将语义标注视为一个语义信息抽取及概念映射的过程，基于上述语义描述框架，提出如图 4 所示的物联网数据资源自动语义标注架构.

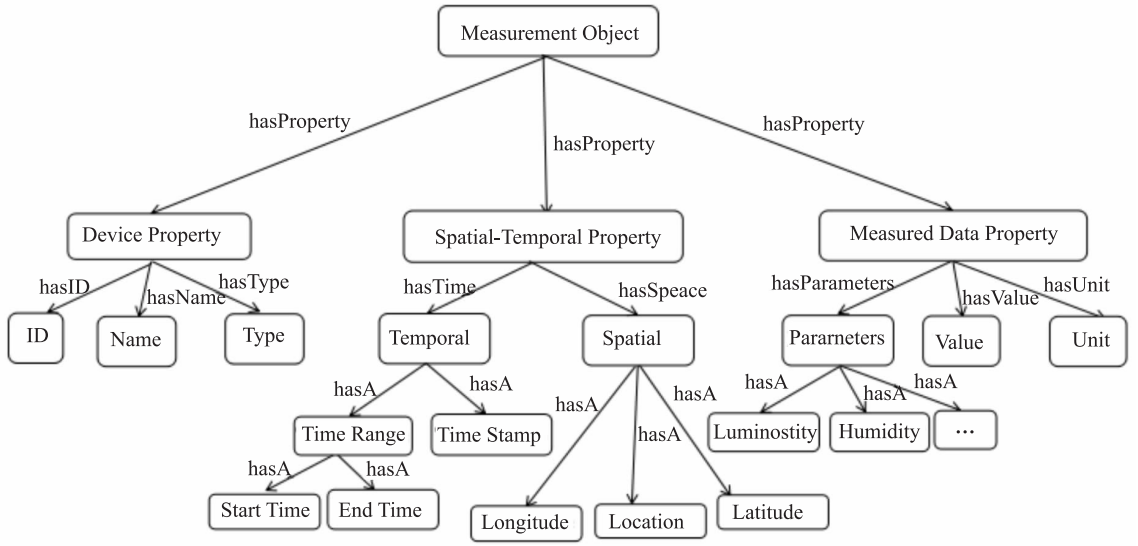

图3　测量对象的语义描述框架

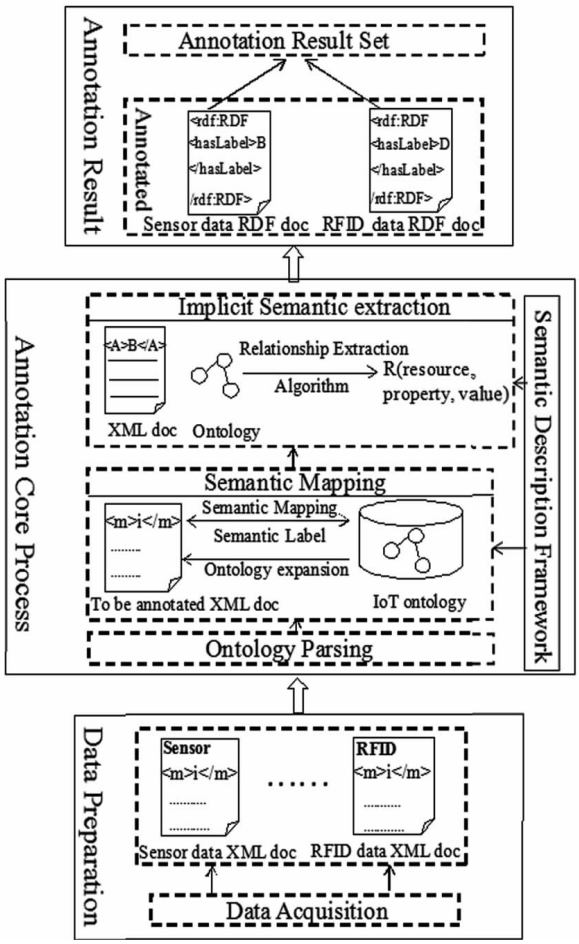

图4　物联网数据资源自动语义标注架构

2.4　语义标注实现的关键方法及技术

2.4.1　本体解析

进行标注前需要将 XML 数据文档和本体文件导入系统，并对本体库进行解析，得到其所包含的概念信息、实例信息以及对象关系信息，为下一步基于本体的语义映射奠定基础. 本文使用 Jena 的 Ontology API 接口的 OntModel，OntClass 和 OntProperty 这 3 个大类对 OWL 本体文件进行解析.

2.4.2　XML 数据属性标注

XML 文档包含两部分信息：< metadata > data instance </metadata >. 元数据是描述数据的数据，即数据属性. XML 数据资源语义标注的第一步就是对于数据属性的标注，这相当于是一个为数据实例（值）添加明确语义标签的过程. 本文通过计算语义相似度的方法来找出 XML 元素与本体中各概念间的对应关系，以此建立 XML 文档中元数据与本体元素之间的语义映射关系，如下所示：

< metadata >

< xml_element > a </xml_element >

< ontology _element > A </ontology _element >

</metadata >

语义映射关系生成之后需要将其存储至本体中，考虑使用 OWL 的 < rdf：comment > 标签来为本体中的概念声明其对应的 XML 元数据. 之后再基于包含语义映射关系的物联网应用本体 I，对于 XML 文档中的每一个元数据 a，找出在 I 中与之匹配的概念 A，并将 A 作为 a 的语义标签进行标注. 经上述处理之后，即完成对于 XML 文档内容数据属性的标注.

2.4.3 XML 对象属性标注

此步骤是对 XML 文档中数据内容之间关系的抽取阶段，在上一步的基础上，还需要进一步挖掘文档中各数据元素在嵌套、并列排列之下隐藏的相互关系，标注出 XML 文档中数据的对象属性.

XML 中数据内容的嵌套及并列关系对应着本体中概念间的关系. 借鉴文献［11］的研究思路，可将 XML 元素的层次结构看作是一种树结构：把最外层的 XML 元素看作根节点，内部的嵌套元素看作下面的父节点，嵌套元素的数据实例（值）看作叶子节点（图 5）. 通过抽取算法将 XML 树结构中的层次关系抽取出来，并将其映射为 RDF 三元组，每个三元组代表一个（实体资源、资源属性，属性值）陈述.

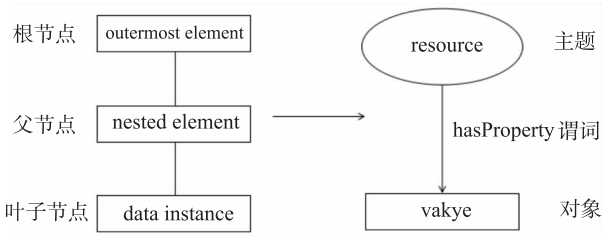

图5 XML层次关系到RDF三元组的映射

上述关系抽取算法的主要思想是对 XML 文档元素的层次结构进行遍历. 首先遍历到 XML 文档最外层元素，即根节点，再从根节点开始对内嵌元素即父节点进行深度遍历，如果内嵌元素节点带有属性，则将该内嵌节点、属性和属性值组成一个 RDF 三元组描述；如果遍历到的节点已无子节点，说明已经遍历到了最下层的叶子节点，即数据实例（值），那么就抽取该元素节点的上层父节点、父节点和节点值组成 RDF 三元组描述，这样依次遍历直至最后一个元素.

为方便计算机对于数据资源语义信息的处理，采用 Jena RDF API 来实现本体实例三元组到 RDF 文件格式的编码，以 RDF 文件格式来存诸 XML 物联网数据文档的标注结果. 最后将关于测量对象的各个 RDF 标注结果整合在一起，形成一个描述该测量对象各类数据信息的整体标注结果集. 至此，即可完成对于物联网多源异构数据的自动语义标注.

3 结语

本文使用语义标注技术对传感器及 RFID 收集到的原始感知数据进行处理，提出物联网多源异构数据的自动语义标注方法. 本研究旨在获取更优的物联网数据，提高物联网不同种类设备之间在语义层面上的互操作性，从而为物联网提供更好的应用服务. 在下一步工作中，需将特定的应用背景（一个整体的测量对象）作为试点，对物联网异构数据资源的自动语义标注系统进行可行性实践与验证，同时考虑基于标注后冗余数据的时空融合处理问题，为物联网的大规模应用及扩展奠定理论与实践基础.

［参考文献］

［1］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［J］. 信息通信技术，2014（5）：26－31.

［2］冯建周，宋沙沙，孔令富. 物联网语义关联和决策方法的研究［J］. 自动化学报，2016，42（11）：1691－1701.

［3］TERZIYAN V，KAYKOVA O，ZHOVTOBRYUKH D. Ubi road：semantic middleware for context-aware smart road environments［J］. International Journal on Advances in Intelligent Systems，2010（3）：295－302.

［4］时念云，杨晨. 基于领域本体的标注方法研究［J］. 计算机工程与设计，2007，28（24）：5985－5987.

［5］王浩然，徐建良，张巍. 一种 XML 元数据的自动语义标注方法［J］. 计算机科学，2008，35（4）：266－268.

［6］WANG W，DE S，CASSAR G，et al. Knowledge representation in the internet of things：semantic modelling and its applications［J］. Automatika，2013，54（4）：388－400.

［7］LIU F，P LI P，D DENG D. Device-oriented automatic semantic annotation in IoT［J］. Journal of Sensors，2017（5）：1－14.

［8］施昭，刘阳，曾鹏，等. 语义网关键技术概述［J］. 中国科学，2015，45（6）：739－751.

［9］SCHLENOFF C，HONG T，LIU C，et al. A literature review of sensor ontologies for manufacturing applications［J］. IEEE International Symposium，2013（10）：96－101.

［10］崔愉. 面向文本的自动语义标注技术研究与实现［D］. 陕西：西安电子科技大学，2014.

［11］乔卫. 基于领域本体的 XML 语义信息抽取的研究与实现［D］. 湖北：武汉理工大学，2009.