语义传感器Web中的数据管理技术研究

李 琪[1] 吴 刚[2]
(中国铁道科学研究院电子计算技术研究所 北京 100081)[1]
(东北大学信息科学与工程学院 沈阳 110004)[2]

摘 要 语义传感器Web是由传感器网络技术、分布式计算技术、数据库管理技术和语义Web技术整合发展而来的。语义传感器Web能够感知、收集、整合信息,抽取新的知识并为感知器提供增强语义,因此能对环境的变化有更智能的感知,用户可以通过访问Web获取这些信息。作为计算机科学中一个新的研究领域,它有着广阔的应用前景,引起了工业界和学术界浓厚的兴趣。介绍了语义传感器Web的基本概念、特点,并着重讨论了语义传感器Web数据管理中所存在的研究问题、研究现状和研究成果。

关键词 传感器,传感器网络,语义Web,语义传感器Web,数据管理
中图法分类号 TP392,TP182 **文献标识码** A

Research of Data Management on Semantic Sensor Web

LI Qi[1] WU Gang[2]
(Institute of Computing Technology, China Academy of Railway Sciences, Beijing 100081, China)[1]
(College of Information Science and Engineering, Northeastern University, Shenyang 110004, China)[2]

Abstract The semantic sensor Web is an integration of sensor network techniques, distributed computation techniques, database management techniques and semantic Web techniques. They can be used for sensing, collecting, integrating information, deriving additional knowledge and providing enhanced meaning for sensor observations so as to enable situation awareness. Users can conveniently access this information by visiting the Web interface. The semantic sensor Web is a new research area of computer science and technology with a wide application prospects. Both academia and industries are very interested in it. The concepts and characteristics of the semantic sensor Web were introduced, and the issues of the data management of the semantic sensor Web were discussed. The advance of the research on the semantic sensor Web, especially the data management, was also presented.

Keywords Sensor, Sensor network, Semantic Web, Semantic sensor Web, Data management

近年来通信技术、传感器技术和嵌入式计算技术的迅猛发展和日益成熟,促使传感器网络(Sensor Network)技术迅速发展。传感器网络技术在国防军事、国家安全、环境监测、交通管理、医疗卫生、制造业、反恐抗灾等领域得到了广泛深入的应用。但是传感器网络在异构网络间的数据通信、整合、查询,以及异构网络间的控制能力上有明显不足。这种不足使数据流彼此孤立,进而加剧了数据量大而知识不足的矛盾。为了解决这个问题,一种基于传感器网络和语义Web[36]的新技术——语义传感器Web(Semantic Sensor Web)得到了广泛的关注。本文将分别介绍语义传感器Web及其数据管理的概念、特点、需要研究的问题,以及目前的研究进展情况。

1 语义传感器Web

1.1 语义传感器Web的概念

语义传感器Web(Semantic Sensor Web,以下简称SSW)是由传感器网络技术和语义Web技术发展而来的。根据语义Web设计者的想法,Web不仅是人与人交互的信息空间,而且是语义丰富的数据网络,既能够被人浏览,也能够利用计算机程序执行操作。语义Web是以某种方式链接、使全球范围内的计算机均可处理的信息网,可通过标准的标记语言和处理对Web进行扩展[1]。语义Web的基础包括数据表示、查询、规则和应用等标准。核心技术是用于表示的资源描述框架RDF(Resource Description Framework)、用于查询的SPARQL(Simple Protocol and RDF Query Language)和用于构建本体的网络本体语言OWL(Web Ontology Language)。

语义Web是数据的网络[26],从这一角度出发,SSW也可以看作是传感器数据的网络。在SSW中,传感器感知的二进制数据首先形成原始数据流(Raw Data Stream);然后根据某种XML Schema(如Sensor Web Enablement)可以将原始数

到稿日期:2012-06-27 返修日期:2012-12-03 本文受国家自然科学基金(60903010),江苏省自然科学基金(BK2009268),中央高校基本科研业务费专项资金(N110404013)资助。
李 琪(1976—),男,副研究员,E-mail:lq0617@sina.com;吴 刚(1978—),男,副教授,CCF会员,E-mail:wugang@ise.neu.edu.cn(通信作者)。

据(Raw Data)格式转换为便于交换的 XML 数据;接着根据相关的本体模型对 XML 数据进行语义标注,即生成具有自说明能力的 RDF 数据;经过语义标注后,应用软件能够理解并在数据上进行连续的、协作的、准确的推理[12];生成的 RDF 数据将会在 OWL 语言约束下,由推理机从感知数据中推理出隐含的信息和知识。数据在这一语义 Web 传感器处理过程中,经历了从原始数据、特征数据、实体数据到知识的过程,如图 1 所示。

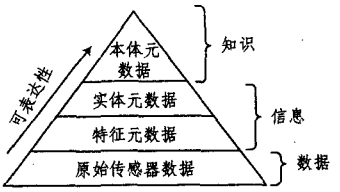

图 1 语义传感器 Web 的数据层次图[1)]

利用 SSW 的特点能够更好地管理异构传感器网络产生的数据,并对来自不同网络、具有不同物理意义的数据进行综合分析,抽取其中的知识,并发布给相关的人员用于制定决策。因此,从更抽象的层次上看,SSW 中的数据处理过程是从产生数据,到将数据转化为信息,再到知识,最后形成决策的过程。

SSW 作为一种基于 Web 访问传感器的网络,其数据应该能够通过标准协议和应用程序接口进行操作[13]。因此,标准化是 SSW 需要解决的问题。为此 OGC (Open Geospatial Consortium)[2)] 和 W3C (Semantic Web Activity of the World Wide Web Consortium)[3)] 分别制订了相关标准。

OGC 制订了一套关于传感器、传感器数据模型、传感器 Web 服务的标准——Sensor Web Enablement (SWE)。这套标准使我们能够通过 Web 访问和控制各种传感器。这套标准主要包括以下几部分[14]:

1. Observations & Measurements (O&M):观测、度量的标准模型和 XML 模式,用于对来自传感器的观察值和测量值编码,数据可以是存档的和实时的。

2. Sensor Model Language (SensorML):标准模型和 XML 模式,用于描述与传感器观察值相关的系统和过程;提供所需要的信息,包括探测传感器、定位传感器观察值、处理低层传感器观察值,列出可任务化处理的属性,并支持传感器观察值的实时处理。

3. Transducer Model Language (TransducerML or TML):传感器模型语言-概念模型和 XML 纲要,用于描述传感器和支持进出传感器系统的实时数据流。

4. Sensor Observations Service (SOS):传感器观察值服务-标准网络服务接口,用于请求、过滤、提取观察值和传感器系统信息,是位于客户端和传感器观察值的存储系统或实时传感器频道附近的一个中间层。

5. Sensor Planning Service (SPS):传感器规划服务-标准网络服务接口,用于请求用户驱动的探测和观察值,是位于客户端和传感器采集管理环境的一个中间层。

6. Sensor Alert Service (SAS):传感器预警服务-标准网络服务接口,用于发布和订阅从传感器发来的预警。

7. Web Notification Services (WNS):网络通告服务-标准网络服务接口,用于异步传递来自 SAS 和 SPS 网络服务的信息和预警,以及其他服务工作流的因素。

W3C 则将 RDF 和 OWL 引入到 Semantic Sensor Web 中,并且制定了一些和 SSW 相关的本体,例如:基于时间计算的本体 OWL-Time[27]。基于这些本体,使用 OWL 语言,利用规则推理系统可以得到有用的信息。如果一组传感器给出了某地的气温和降雨量,则使用规则可以推理出该地道路的结冰状况,如图 2 所示。

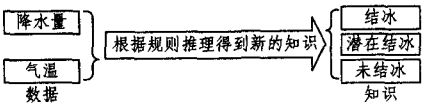

图 2 基于规则的道路结冰情况判别推理过程

下面是基于语义 Web 规则语言 SWRL[37] 描述的推理。若气温低于华氏 32 度,且下雨,则道路会结冰[12]:

Rule:Potentially Icy (with freezing temperature and rain) Observation (? obs) &measured(? obs,? precip) &Rain(? precip) &measured(? obs,? temp) &Temperature(? temp) &temperature_value(? temp,? tval) &lessThanOrEqual(? tval,32) &unit_of_measurement(? temp, Fahrenheit)→ described(? obs,Potentially_Icy)

可见,通过规则的约束和推理,应用程序能够更好地理解传感器的数据进而做出相应的决策。

1.2 语义传感器 Web 的特点

在信息技术日益发展的今天,传感器网络技术的应用也日新月异。传感器网络已应用于很多的领域。但是传统的传感器网络技术在数据融合、信息抽取和知识获取中遇到了一系列的问题。SSW 在解决这些问题上起着十分重要的作用[15],解决这些问题正是 SSW 的特点所在。总结起来主要有以下几个方面:

1. SSW 能够有效地抽取信息

不同的传感器网络产生的数据都是单一种类的。但是在应用的时候,需要将这些来自不同数据源的数据进行整合以获取系统的整体信息。例如需要监测并分析某个生态系统中某类疾病的发病情况的时候,不仅需要传感器网络返回该系统中目前的发病情况,还需要得到该系统中的气候变化情况,例如气温、降水量、湿度,以及地理信息等。这些数据都是由相互独立的传感器网络返回的。因此,将这些信息整合并返回给相应的应用程序是十分重要的。在 SSW 中,将经过处理的 XML 文档进行语义标注,经过标注的信息与相关的本体进行关联形成整体信息,这样就解决了异构的传感器网络之间的信息融合问题。

2. SSW 能够进行不同数据源的数据融合

在数据处理的过程中,可根据相关的标准(例如 SWE)将不同的传感器网络产生的原始数据转换为可以用于信息交换的 XML 数据。这种操作使不同的传感器网络产生的数据可以进行统一的整合。

3. SSW 能够获取丰富的知识

1) 截取自 http://knoesis.wright.edu/library/presentations/Semantic-Sensor-Web-Australia.ppt

2) http://www.open-geospatial.org

3) http://www.w3.org/2001/sw/

引入了语义 Web 中的 OWL 语言。利用语言提供的约束规则，应用程序可以根据已有的信息推理出需要的知识。这些知识将会为决策和分析提供强有力的知识支持，从而解决目前传感器网络中知识不足的问题。

2 语义传感器 Web 中的数据管理

2.1 数据管理是语义传感器 Web 的重要内容

对于传感器网络而言，其上的任何应用系统都离不开感知数据的管理和处理技术。不言而喻，传感器网络数据管理和处理技术是确定传感器网络可用性和有效性的关键技术，关系到传感器网络的成败。对于观察者而言，传感器网络的核心是感知数据，而不是网络硬件。观察者感兴趣的是传感器产生的数据[16]。

同理，对于 SSW 而言，数据管理也是具有重要意义的。在 SSW 中，为用户和应用程序提供的是信息和知识。这些信息和知识都是原始的数据经过标准化和语义标注后，再经过基于一些规则的推理而得出的。这个过程中，首先是原始数据的消化、异构数据的融合，之后是标注和推理。与此同时，还要处理相关的用户查询等要求。整个过程就是一个数据管理的过程，如图 3 所示。因此，数据管理是 SSW 的重要内容之一。

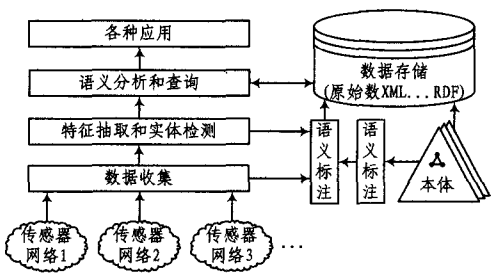

图 3 SSW 整体框架[4)]

2.2 语义传感器 Web 数据管理的难点

采用 SSW 技术解决传感器网络的异构数据融合、信息抽取以及知识不足等问题时，不仅要解决原始数据失真等传感器网络本身面临的问题；同时还要解决语义 Web 所引入的新问题；另外，还要解决这两种技术融合所带来的一些问题：

1)由于传感器网络所处的工作环境以及传感器自身设计上的一些问题，获取的数据往往会出现一些错误和缺失。虽然已经有个别的数据库系统提供抽取、转换和载入感知数据的工具(ETL)，但这些系统都缺少数据校正和填补的能力。尽管一些设备生产商会提供数据调校软件，但此类软件大都是基于特定设备元件的。因此，研究通用的感知数据获取技术对于 SSW 将是十分重要和有意义的。

2)在 SSW 中，数据的存储系统在整体架构和具体存储形式上都存在一定的问题。一方面，系统设计者们希望将数据分散存储到传感器中，但由于传感器运算能力、存储空间、功耗以及通信易受环境影响，一般要设置专用的数据存储设备来存储这些数据，这种分布式的存储结构面临一些问题，另一方面，SSW 中数据的存储主要有两个解决途径：其一是存储传感器网络的流数据，然后在上层进行相关的融合和处理；其二是将数据转换成 RDF 再进行存储[28]。对于第一种途径，由于流数据是多重、连续、随时间快速变化的，因此传统的关系数据库难以对其进行处理。对于第二种途径，需要解决 RDF 的数据存储问题。RDF 是一种高度规范化的数据模型[17]。RDF 自身的特性使之不便于用传统的关系数据库存储。如何构建适合 RDF 的数据库系统至关重要。

3)传感器网络中的数据以数据流形式存在，因此 SSW 中处理的也是各种形式的数据流，包括：原始数据流、XML 数据流、RDF 数据流等。为了满足应用系统的实时性等需求，需要设计一系列针对流数据的高效查询处理方法。

4)由于传感器网络的地理分布和数据量大等原因，分布式数据查询是较为可行的解决途径。采用分布式查询在带来优点的同时也导致了一些问题，其中最为突出的是联邦数据(Federal data)的处理问题。然而这正是语义 Web 的优势。因此，将语义 Web 技术与分布式查询处理有效结合具有重要意义。

5)本体匹配问题。在 SSW 中需要将 RDF 数据与本体进行关联，以便为推理做准备。然而由于本体并不是全局统一的，不同的团体和组织在构建 SSW 的过程中可能采用不同的本体表达相同的概念。这是信息融合和知识抽取面临的主要问题。本体匹配相应变得十分重要。本体引擎需要做两方面的工作：识别相关的本体；匹配和记录属于相关本体的实体之间的关系[24]。

6)SSW 的最大优点在于能够整合异构传感器网络的数据信息，通过推理从已知的数据信息中抽取出内在的隐含信息——知识。这将有助于在线决策制定等应用的开发。SSW 上的推理研究刚刚起步，尚不成熟。

2.3 语义传感器 Web 数据管理的研究现状

在 SSW 的数据管理这一研究领域，学术界和工业界做了很多探索工作，研究现状如下。

1. 原始数据的获取

对于某些传感器节点的数据缺失，目前通行的做法是首先对原始的数据进行网格化处理。所谓的网格化处理，是指将传感器感知的原始数据进行时间和空间上的定位，即每个传感器节点产生的感知数据都有时间和空间两个坐标[2]。对于缺失的数据可以通过对比该节点的历史数据或者相邻节点的数据进行填补和调校。虽然一些 ETL(extract, transform, load)工具，如 Apatar[5)]、Altova MapForce[6)] 和 HiT Software-Allora[7)] 等都有抽取、转换和载入感知数据的功能，但是它们均缺少数据校正、填补和网格化的软件。尽管一些设备生产商会提供数据调校的软件，但此类软件大都是基于特定设备元件的。因此建立一个通用的数据获取工具对于语义传感器 Web 将是十分重要和有意义的。

2. 链接传感器数据的发布

语义传感器 Web 中的数据集成与融合主要是指异构数据流上的集成与融合。在这一阶段可以引入一定的语义技术。当前的研究趋势是从传感器网络数据流中产生链接数

4) 截取自 http://knoesis.wright.edu/library/presentations/Semantic-Sensor-Web-Australia.ppt

5) http://www.apatar.com/

6) http://www.altova.com/

7) http://www.hitsw.com/

据[40]，主要做法是通过将基于传感器的数据转换成 RDF 数据并通过使用传感器相关 URI 使其可通过 HTTP 协议访问[41,42]。这一做法保证了传感器数据(及其它类型数据)间的无缝导航。目前，较为成熟的链接传感器数据发布平台是由 Payam Barnaghi 和 Mirko Presser 开发的 Sense2Web[43]，它通过 SPARQL endpoint 来发布链接数据并使其它 Web 应用可访问该数据。Sense2Web 支持对传感器描述数据、观测数据和度量数据的发布。Harshal Patni 等人在其研究中[44]构建了两个实际的 RDF 数据集，即 LinkedSensorData 和 LinkedObservationData，分别用于描述链接传感器数据和链接观测数据，总的三元组数量已经超过 10 亿。

众所周知，RDF 数据比较适合表示专题元数据(thematic metadata)，而不能直接表示传感器获取的具有时间性和空间性的数据等。因此，传感器数据发布过程中，将传感器数据转换成 RDF 数据格式是一个关键问题。根据 W3C 语义传感器网络孵化器小组(W3C Semantic Sensor Network Incubator Group)的总结，目前有如下 3 种与 OGC 标准相兼容的语义标注机制[8]。

1)基于 XLink 的方法：将 xlink:href 映射成 rdf:resource (OGC 所使用的地理标识语言 GML 具有类似于 RDF 的结构)，这种方式存在的问题是在 OGC 内部有多种对 XLink 的解释方法；

2)基于 RDFa 的方法：RDFa 标准[9]提供了针对语义标注的形式化语法和解释，但目前主要在 XHTML 领域中使用；

3)从 XML 元素值和属性中抽取语义：SWE 标准中的某些元素中可能含有 RDF 资源的 URI，但它们未使用 XLink，例如 srsName 属性。

下面展示的是基于 RDFa 的方法将时间和地理信息数据标注成 RDF 数据的脚本片段。

```
〈swe:component rdfa:about="time_1" rdfa:instanceof="time:Instant"〉
  〈swe:Time rdfa:property="xs:date-time"〉2008-03-08T05:00:00
  〈/swe:Time〉
〈/swe:component〉
〈swe:value name="satellite-data" rdfa:about="Dayton" rdfa:instanceof="geo:City"〉
  0011000111001111
〈/swe:value〉
```

3. 语义传感器 Web 的数据存储

在 SSW 中，数据的特点是分散而且数据量大，并且传感器自身的存储和处理能力偏弱。因此，一般的方法就是构建一些分布式的数据库，专门用于数据的存储和处理。P2P 技术是一个较好的选择。在 P2P 应用中，需要有效定位所存储的数据条目的节点。Chord 是麻省理工大学计算机科学实验室提出的一个用于分布式查询的协议。该协议解决了 P2P 中定位的问题，其中为每一数据分配一个关键字，进而将“数据-关键字”对存储到以关键字做映射操作得到的节点上。这样，即使在连续更新数据的情况下也可以快速定位到查询节点，提供性能稳定的查询服务[18]。Ryan Huebsch 等人构建了首个通用的关系查询处理器 PIER(Peer-to-peer Information Exchange and Retrieval)。PIER 是以 P2P 为技术架构，使用互联网上的数千个节点构建而成的，用于对大规模的分布式数据库风格的数据流进行快照和连续查询[19]。Timothy Roscoe, Scott Shenker, Ion Stoica 和 Aydan R. Yumerefendi 分别在 IrisNet, Aurora/Medusa, Borealis, TelegraphCQ 等基于分布式哈希表(DHT)的系统中使用了相关的技术，提供了解决大规模、连续、分布式数据的存储和处理等问题的方案[2]。

在 Danh Le-Phuoc 和 Manfred Hauswirth 构建的 SSW 系统中，采用了先将传感器产生的流数据进行存储，再在其数据管理系统之上由 SensorMasher 层来完成数据融合和语义标注的方法[29]。该系统中提出使用数据流管理系统(Data Stream Management System, DSMS)[30]作为黑盒数据库系统来管理和维护数据。DSMS 是在斯坦福流数据管理 (Stanford Stream Data Management, STREAM)[32]的基础上构建完成的。DSMS 解决了流数据存储时存在的多重、连续、快速、时间可变等难题。在这个系统基础上，还出现了对于大容量的、多变量的大规模流数据的连续查询工具 TelegraphCQ[31]。Rajeev Motwani 等人研究了将 DSMS 用于连续、多重、资源有限条件下的数据流近似查询[30]。

在处理 SSW 中的 RDF 数据存储的问题上主要有以下几种思路：利用传统的关系数据库存储数据，然后将其转换成 RDF 数据，例如使用 D2R 处理器将数据从关系数据库映射到 RDF 数据[20]；同时也可考虑将 RDF 数据用关系数据库进行存储，例如 3Store[21]。但使用传统的关系数据库存储 RDF 数据将会在数据库中造成大表(Large Table)。大表的存储和查询处理都是十分困难和低效的。目前常见的做法是开发专用的 RDF 数据库系统，其中影响力比较大的有：Sesame, Jena, Kowari, 3Store 和 RDFStore 等。前三者支持 OWL 特性[22]。这些系统在性能上各有所侧重，测试发现 3Store 和 Sesame 总体上通过配置数据可以得到不错的性能，而 Kowari 和 RDFStore 在这方面稍有逊色[23]。

4. 数据流查询

由于 SSW 中数据存储的解决思路不同：一种途径是存储传感器数据，然后经过上层的模块转换成为 RDF 数据；另一途径是直接存储 RDF 数据。因此，针对 SSW 数据流查询的解决思路也有两大方向。

较为常见的对传感器产生的数据流进行查询的方法大都基于前述提到的 DSMS 而构建。例如 TelegraphCQ 采用窗口查询的方式进行流查询[31]。Rajeev Motwani 等人改进了 DSMS，使其支持描述式查询语言，并且采用近似给出结果的方法进一步缓解了高数据率和查询工作的负担[30]。

另外由布兰迪斯大学、布朗大学和麻省理工共同开发的第二代分布式查询处理引擎 Borealis 也是一个非常出色的流数据处理工具。Borealis 继承了 Auroa[33]的流查询内核和 Medusa[34]的分布式处理功能，满足了新出现的流处理应用的常见需求[35]。

目前针对 RDF 数据进行查询所使用的语言主要是 SPARQL[3]。SPARQL 是为 RDF 数据模型定义的查询语言和数据获取协议。SPARQL 可以用于对多种数据源进行查

8) http://www.w3.org/2005/Incubator/ssn/wiki/Semantic_Mark_up

9) http://www.w3.org/TR/xhtml-rdfa-primer/

询而不用考虑数据是直接存储的 RDF 数据，还是通过中间件转换而来的 RDF 数据[1]。目前，具有 SPARQL 查询能力的 RDF 数据管理系统较多，如 Jena[4] 是一个为语义 Web 应用而建立的编程框架，它为 RDF、RDFS 和 OWL 提供了编程环境。另外，开源项目 Sesame[5] 也具备 SPARQL 查询处理能力。

由于 SPARQL 并非是为数据流形式的 RDF 查询而设计的，因此一些研究人员对 SPARQL 进行了扩展和优化，以便支持 RDF 数据流查询。DF Barbieri 等人提出了 SPARQL 的一个扩展 C-SPARQL[6]，其主要侧重于对连续的 RDF 数据流进行查询。Sven Groppe 等人构建了一个对 RDF 数据流进行查询的引擎[7]，它对 SPARQL 语言在查询过程中的图匹配和合取过程的顺序进行了优化，并将哈希方法引入到模式匹配中以提高查询效率。

5. 分布式查询

如前所述，SSW 查询处理中的另一个突出问题是联邦数据处理。联邦数据系统(Federal Information System)的概念最早是由 Busse S,Kutsche 等人提出的[8]。解决方案主要是利用多数据库查询语言(Multi-Database Query Languages，MDBQL)，如 SchemaSQ[9]。除此之外，还有联邦数据库和基于调解器的信息系统(Mediator Based Information Systems，MBIS)。多数据库查询语言要求用户提供用于查询的数据源的详细说明，与之相对的 MBIS 则隐藏了这些要求，取而代之的是提供了单独统一的模式。值得一提的是，SPARQL 可以被认为是用于 RDF 查询的多数据库查询语言[10]。但就目前来看，SPARQL 查询主要是本地查询。虽然在一些系统中使用 SPARQL 来完成联邦数据查询，但此类查询都是将分布的数据整合到本地数据库中，然后再采用 SPARQL 查询。载入的数据可以来自基于 RDF 存储的数据源，也可以来自传统的关系数据库。如果是后者，则需要经过中间件的转化[10]。由于将远程数据载入本地后再进行合并查询的解决方法可能存在技术和法律上的问题，Bastian Quilitz 和 UlfLeser 提出了 DARQ[11] 系统。DARQ 是一个联合的 SPARQL 查询引擎，提供了对于多数据源的透明式 SPARQL 查询。这些查询的数据在实际中是分布式存储的，而用户在感觉上却认为是在一个单一的 RDF 图上进行的。DARQ 还使用了重写查询和基于代价的查询优化来加速查询的执行。

杨梦东和吴刚最近提出了一种基于频繁 RDF 图模式划分的并行 RDF 数据处理方法[45]。传统的基于单机架构的集中式 RDF 数据管理工具已经越来越难以应付不断增长的 RDF 数据。在这样的背景下，发展出了很多分布式 RDF 数据管理方案。现有分布式 RDF 数据库主要关注两个问题：1) 对 RDF 数据的索引；2) SPARQL 查询优化。现有分布式 RDF 数据管理工作在 RDF 数据划分方面给予的关注相对较少，往往仅采用较为简单直观的划分方法与划分粒度(如 RDF 文档、RDF 命名图、RDF 句子/分子/三元组)，缺乏针对 RDF 图数据和 SPARQL 查询特性设计的 RDF 数据划分方法。仅有的根据图数据特性划分 RDF 数据的工作也因算法复杂度过高而难以胜任大规模 RDF 数据的划分。该研究主要考察分布式 RDF 数据库中的数据划分及相应的查询处理问题。根据已有工作中存在的上述不足之处，提出了一种基于模式图划分的分布式 RDF 数据处理方法，包括如下两个主要技术点：1)一种基于模式图的 RDF 数据划分方法。这种划分方法同时考虑了数据(data)和工作负载(workload)两方面的因素。首先，通过对 SPARQL 查询进行模式化处理，得到模式化查询图(Patternized Query Graph，PQG)，PQG 代表了一个比具体 SPARQL 查询定义的基本图模式(Basic Graph Pattern，BGP)更一般的图模式。PQG 是发现模式图的依据。在三元组加载前，对于每个 PQG 默认有一个空的模式图与其对应，通过对这些模式图建立倒排索引，可以在三元组加载时发现模式图。通过对这些发现的模式图建立倒排索引，可以进一步发现更多的模式图。发现的模式图被索引，索引键是那些原始 SPARQL 查询中出现了具体值的节点位置。索引后的模式图按照其首三元组的主语值被哈希划分至各个节点进行存储，首三元组为将模式图中的三元组按照谓语值排序后的首个三元组。2)在所提出的模式图的 RDF 数据划分方法基础上，设计了有针对性的 SPARQL 查询处理算法。该研究在开源 RDF 数据处理框架 Sesame 上进行了实现，并在 3 个主流测试基准(benchmark)上进行了实验测试。实验数据表明了本文所提方法的正确性和有效性。由于 RDF 的数据存储粒度为模式图，因此避免了以三元组作为存储粒度时需要进行大量连接操作的问题，查询处理性能有较为明显的提升。该项研究为分布式环境下的 RDF 数据管理提供了一种新的数据划分与查询处理方法，对分布式 RDF 数据管理研究具有一定的贡献与参考价值。

6. 本体匹配

由于本体不是全局统一的，在 SSW 中不同系统可能使用不同本体来构建，导致进行信息的整合过程遇到了一些问题。本体匹配(ontology matching)因此成为了重要的研究课题。相关的本体匹配技术主要有 4 项[25]：基于词法的方法(基于对标签和词表的词法比较)；基于结构的方法(利用本体的结构信息)；基于背景知识的方法(使用额外的外部知识)；基于实例的方法(使用经过分类的实例数据)。这些技术主要用于：

(1)问答(question answering)：在这种情况下，本体通常非常巨大和复杂，并且有大量的文字重叠，不能直接使用基于词法的技术进行处理，必须先分析本体的域名和使用的结构，然后才使用词汇匹配来解决。

(2)数据集上的统一视图(unified view over collections)：在这里命名规格和决策模型可能不同，但是基于词法的技术解决了绝大多数的问题。基于结构的技术在这里很少使用。

(3)浏览中的主题偶然组配(serendipity in browsing)：词汇匹配在这里不适合。这主要是由于命名概念模糊，以及缺少标准化的分类实例数据。前者可以通过基于背景知识的方法来解决。后者只能通过基于实例的方法来解决。实际上，匹配若建立在实例数据上，则必须通过分析给定类子类的上下文来确定。

(4)数据迁移(data migration)：这种情况下主要通过基于词法的技术和基于背景知识的技术来解决。

7. 推理的研究

目前大多数的推理都是基于语义 Web 规则语言 SWRL(Semantic Web Rule Language)，从现有的事实中获取信息，得到新的知识[12]。比如根据传感器网络返回的某地气温、降雨量等数据，就可以利用定义的一些规则来推测出该地的道

路湿滑程度和结冰状况等。目前有很多系统都是基于 SWRL 技术来构建的。例如 ES3N[4] 系统根据湿度传感器和温度传感器返回的数据以及传感器分布的位置信息，便可推测出应该在何时对粮仓进行通风处理。Amit Sheth 等构建了一个用于在时间范围内进行推理的语义传感器 Web 系统[12]，对来自俄亥俄州巡逻系统的摄像机产生的视频进行语义和地理信息标注。这些数据被上传至 YouTube 站点中，使用者可以在谷歌地图中对某地在某时间段内的情况进行查询，系统则会根据标注的信息进行判断和推理，从而提供符合用户要求的视频信息。

吴刚等研究了对推理结果提供合理解释的技术[46]。通常，可以通过找出导致该推理结果的所有推理过程（或称为验证——justification）的方法来实现。现有的本体数据验证方法的可伸缩性普遍较差，无法处理规模达到百万（million）数量级的三元组（本体的基本组成单位）的大规模本体数据。即使采用一些优化技术，例如基于模块抽取的优化技术，也无法处理十亿（billion）数量级三元组，其广泛存在于真实的链接开放数据 LOD 中的大规模 OWL 本体中。研究中考虑到传统人工智能领域问题求解（Problem Solving）中广泛采用的真值维护系统（Truth Maintenance Systems，TMS）的问题求解过程与本体推理结果验证中的白盒求解过程（glass-box）之间所存在的相似之处，提出了将 JTMS 系统（一种最简单的 TMS 系统）部分用于解决本体推理结果的解释问题。研究发现由无环 JTMS 依赖图中某一节点出发递归回溯，可以查找到该结点所表示的三元组的所有解释。由于 TMS 系统能够较好地处理在较大搜索空间中的问题求解，因此基于 TMS 系统的本体推理结果验证能够在一定程度上改善性能。为了处理十亿数量级三元组本体，在前述算法基础上，提出了相应的 MapReduce 算法。查找推理结果的所有推理过程可以分为两个阶段，即构造阶段和查找阶段。构造阶段，在基于开源 Hadoop MapReduce 框架的分布式 OWLpD^* 推理引擎基础上构建 TMS 系统。查找阶段，在前一阶段构造好的 TMS 系统之上，设计 MapReduce 算法用于查找相应结果。在拟合数据集 LUBM 和真实数据集 Dbpedia[10] 上的实验结果显示他们提出的方法具有较好的可伸缩性，并能处理含有 10 亿数量级三元组的本体。

结束语 以上讨论了语义传感器 Web 的概念及特性。该技术能够解决传统传感器网络在数据融合、信息获取和知识融合方面的不足，为用户提供友好的访问方式，避免用户必须掌握足够多的专业背景知识才能使用相关数据的缺陷，能够有效地融合异构数据，整合信息，通过基于规则的推理获得知识。语义传感器 Web 技术融合了传感器网络技术和语义 Web 技术，在将语义 Web 技术引入的同时也带来了数据管理上的一些难题，包括系统整体的结构问题、RDF 数据存储问题、流查询的处理问题、分布式查询的问题，以及本体匹配和推理方面的问题。

参 考 文 献

[1] 高志强，潘越，马力. 语义 Web 原理及应用[M]. 北京：机械工业出版社，2009

[2] Balazinska M, Deshpande A, Franklin M J, et al. Data Management in the Worldwide Sensor Web[J]. IEEE Pervasive Computing, 2007, 6(2): 30-40

[3] Prud'hommeaux E, Seaborne A. A SPARQL Query Language for RDF, W3C Recommendation[EB/OL]. http://www.w3.org/TR/rdf-sparql-query/, 2011-11-27

[4] Feng Tao, Campbell J, Pagnani M, et al. Collaborative Ocean Resource Interoperability: Multi-use of Ocean Data on the Semantic Web[C]// The Semantic Web: Research and Applications. The 6th Annual European Semantic Web Conference (ESWC2009). Berlin: Springer-Verlag, 2009: 753-767

[5] openRDF.org. openRDF.org: Home[EB/OL]. http://www.openrdf.org/, 2011-11-27

[6] Barbieri D F, Braga D, Ceri S, et al. C-SPARQL: SPARQL for continuous querying[C]// Proceedings of the 18th international conference on World Wide Web (WWW'09). New York: ACM, 2009: 1061-1062

[7] Groppe S, Groppe J, Kukulenz D, et al. A SPARQL Engine for Streaming RDF Data[C]// Proceedings of the Third International IEEE Conference on Signal-Image Technologies and Internet-Based System (SITIS'07). Washington: IEEE Computer Society, 2007: 157-168

[8] Busse S, Kutsche R-D, Leser U, et al. Federated information systems: Concepts, terminology and architectures[R]. Technical Report Forschungsberichtedes Fachbereichs Informatik 99-9. Berlin: Technische Universität Berlin, 1999

[9] Lakshmanan L V S, Sadri F, Subramanian I N. SchemaSQL-A language for interoperability in relational multi-database systems[C]// Proceedings of the 22th International Conference on Very Large Databases (VLDB'1996). San Francisco: Morgan Kaufmann Publishers Inc., 1996: 239-250

[10] Langegger A, Wöβ W, Blöchl M. A Semantic Web Middleware for Virtual Data Integration on the Web[C]// Proceedings of the 5th european semantic web conference on the semantic web: research and applications (ESWC'08). Berlin: Springer-Verlag, 2008: 493-507

[11] Quilitz B, Leser U. Querying Distributed RDF Data Sources with SPARQL[C]// Proceedings of the 5th European Semantic Web Conference on The Semantic Web: Research and Applications (ESWC'08). Berlin: Springer-Verlag, 2008: 524-538

[12] Sheth A, Henson C, Sahoo S S. Semantic Sensor Web[J]. IEEE Internet Computing, 2008, 12(4): 78-83

[13] Botts M, Percivall G, Reed C, et al. OGC® Sensor Web Enablement: Overview and High Level Architecture[C]// GeoSensor Networks, Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2008: 175-190

[14] OGC. Sensor Web Enablement DWG | OGC(R)[EB/OL]. http://www.opengeospatial.org/projects/groups/sensorweb, 2011-11-27

[15] Compton M, Henson C, Neuhaus H, et al. A Survey of the Semantic Specification of Sensors[C]// Proceedings of the 2nd International Workshop on Semantic Sensor Networks, at 8th International Semantic Web Conference. CEUR-WS.org, 2009: 17-32

[16] 李建中，李金宝，石胜飞. 传感器网络及其数据管理的概念、问题

10) http://downloads.dbpedia.org/3.6/en/

与进展[J]. 软件学报,2003,14(10):1717-1727

[17] Manola F. A Database Perspective on the Semantic Web: A Brief Commentary[J]. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering,2003,26(4):5-11

[18] Stoica I, Morris R, Karger D, et al. Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications[C]// Proceedings of ACM SIGCOMM 2001. New York: ACM Press, 2001: 149-160

[19] Huebsch R, Chun B N, Hellerstein J M, et al. The Architecture of PIER: An Internet-Scale Query Processor[C]// Proceedings of the 2nd Biennial Conference on Innovative Data Systems Research (CIDR'05). VLDB Endowment, 2005: 28-43

[20] Bizer C. D2R MAP-Database to RDF Mapping Language and Processor[EB/OL]. http://www4. wiwiss. fu-berlin. de/bizer/d2rmap/D2Rmap. htm, 2011-11-28

[21] Reggiori A, van Gulik D-W, Bjelogrlic Z. Indexing and Retrieving Semantic Web Resources: the RDF Store model[C]// Proceedings of SWAD-Europe Workshop on Semantic Web Storage and Retrieval. 2001

[22] Owens A. Semantic Storage: Overview and Assessment [EB/OL]. http://eprints. ecs. soton. ac. uk/11985/1/owens-semanticStorageOverview-ecs. pdf, 2011-11-28

[23] Lee R. Scalability Report on Triple Store Applications [R]. Technical report. Massachusetts Institute of Technology. USA: Massachusetts Institute of Technology, 2004

[24] Euzenat J, Shvaiko P. Ontology Matching[M]. Berlin: Springer, 2007: 10

[25] Aleksovski Z, van Hage W R, Isaac A. A survey and categorization of ontology-matching cases[C]// Proceedings of the Workshop on Ontology Matching (OM2007) Collocated with the 6th International Semantic Web Conference (ISWC-2007) and the 2nd Asian Semantic Web Conference (ASWC-2007). CEUR-WS. org, 2007

[26] Bernes-Lee T. Semantic Web road map[EB/OL]. http://www.w3. org/DesignIssues/Semantic. html, 2011-11-28

[27] Hobbs J R, Pan Feng. Time Ontology in OWL, W3C Working Draft[EB/OL]. http://www. w3. org/TR/owl-time/, 2011-11-28

[28] Babcock B, Babu S, Datar M, et al. Models and issues in data stream systems[C]// Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS'02). New York: ACM Press, 2002: 1-16

[29] Le-phuoc D, Hauswirth M. Linked opendata in sensor data mashups[C]// Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09). CEUR-WS. org, 2009: 1-16

[30] Motwani R, et al. Query Processing, Resource Management, and Approximation in a Data Stream Management System[C]// Proceedings of the Conference on Innovative Data Systems Research (CIDR'2003). VLDB Endowment, 2003

[31] Chandrasekaran S, et al. TelegraphCQ: Continuous Data Flow Processing for an Uncertain World[C]// Proceedings of the Conference on Innovative Data Systems Research (CIDR'2003). VLDB Endowment, 2003

[32] Stanford University InfoLab. Stanford Stream Data Management (STREAM) Project[EB/OL]. http://www-db. stanford. edu/stream. 2011-11-28

[33] Carney D, et al. Monitoring streams: a new class of data management applications[C]// Proceedings of the 28th International Conference on Very Large Data Bases (VLDB'02). VLDB Endowment, 2002: 215-226

[34] Zdonik S B, Stonebraker M, Cherniack M, et al. The Aurora and Medusa Projects[J]. IEEE Data Engineering Bulletin, 2003, 26(1): 3-10

[35] Abadi D J, et al. The Design of the Borealis Stream Processing Engine[C]// 2nd Biennial Conference on Innovative Data Systems Research (CIDR'05). VLDB Endowment, 2005: 277-289

[36] Berners-Lee T, Handler J, Lassila O. The semantic web[J]. Scientific American Magazine, 2001, 5

[37] Horrocks I, et al. SWRL: A Semantic Web Rule Language Combining OWL and RuleML, W3C Member Submission[EB/OL]. http://www. w3. org/Submission/SWRL, 2011-11-28/

[38] Yang Meng-dong, Wu Gang. Semantic Caching for Semantic Web Applications[C]// Proceedings of Joint International Semantic Technology Conference (JIST2011). Berlin: Springer-Verlag, 2011: 192-209

[39] 刘翔宇,吴刚. 基于 Prüfer 序列的 RDF 数据索引与查询[J]. 计算机学报,2011,34(10):1977-2008

[40] Corcho O, Garcia-Castro R, et al. Five Challenges for the Semantic Sensor Web[J]. Semantic Web Journal, 2010, 1(1/2): 121-125

[41] Sequeda J, Corcho O. Linked Stream Data: A Position Paper[C]// Proceedings of the 2nd International Workshop on Semantic Sensor Networks (SSN09). CEUR-WS. org, 2009

[42] Page K, de Roure D, Martinez K, et al. Linked Sensor Data: RESTfully serving RDF and GML[C]// Proceedings of the 2nd Int. Workshop on Semantic Sensor Networks (SSN09). CEUR-WS. org, 2009

[43] Barnaghi P, Presser M. Publishing Linked Sensor Data [C]// Proceedings of the 3rd International Workshop on Semantic Sensor Networks 2010 (SSN10) in Conjunction with the 9th International Semantic Web Confe-rence (ISWC 2010). CEUR-WS. org, 2010

[44] Patni H, Sahoo S S, Henson C, et al. Provenance Aware Linked Sensor Data[C]// Proceedings of the 2nd Workshop on Trust and Privacy on the Social and Semantic Web Co-located with ESWC2010, the 7th European / Extended Semantic Web Conference. 2010

[45] Yang Meng-dong, Wu Gang. A Workload-based Partitioning Scheme for Parallel RDF Data Processing[C]// Proceedings of the Sixth Chinese Semantic Web Symposium and the First Chinese Web Science Conference (SWWS'12). 2012

[46] Wu Gang, Qi Gui-lin, Du Jian-feng. Finding All Justifications of OWL Entailments Using TMS and MapReduce[C]// Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM 2011). Berlin: Springer-Verlag, 2011: 1425-1434