文章编号：2095-2163(2019)06-0217-05 中图分类号：TP392 文献标志码：A

知识图谱人物本体模型设计方法

魏玉良，黄 纯，王佰玲
（哈尔滨工业大学(威海) 计算机科学与技术学院，山东 威海 264209）

摘 要： 人物本体在众多领域知识图谱中具有重要的作用，但目前人物本体设计较为简单，难以通用。本文通过人物相关案例构建小型的本体模型，分析其中存在的多元关系问题，结合多元关系的定义介绍多元关系的表示方法。对于在多元关系设计中存在不同设计方案，提出通过将本体模型部分映射为关系数据库，利用关系数据库 N 泛式原则优化多元关系设计。利用现有的本体模型基础上设计领域本体模型，在兼顾领域问题的同时保证扩展性和通用性，并基于 Wikidata 中的 Human 类设计，给出了本文中的人物本体泛式，专门针对地理位置和时间本体进行了优化设计。

关键词： 本体模型；多元关系；领域图谱；知识图谱

Knowledge graph person ontology model design method

WEI Yuliang, HUANG Chun, WANG Bailing
(School of Computer Science and Technology, Harbin Institute of Technology (Weihai) Shandong 264209, China)

【Abstract】 Person ontology plays an important role in many domain knowledge graph, but the current person ontology is relatively simple in design and difficult to be universal. This paper constructs a small ontology model through a case, and analyzes the n-ary relation represention issues. For the different design schemes in the n-ary design, it is proposed to optimize the n-ary design by mapping the ontology model into a relational database, and then N-generic principle is uesd for optimizing. The domain ontology model is designed to ensure the extensibility and versatility. Based on the Human class design in Wikidata, the person ontology pattern is given in this paper. Besides this paper proposes a new design method for the geographic location and time ontology.

【Key words】 ontology model; n-ary relation; domain ontology; knowledge graph

0 引 言

知识是数据中有规律的信息和信息上下文的集合，知识的上下文表示信息之间的关系，知识具有经验性。为了让计算机可以存储和计算知识，上世纪五十年代学者提出的一种可以在计算机硬件中的存储和表示知识形式—语义网络 (Semantic Network)[1]。

语义网使用 W3C 制定的资源描述框架 RDF (Resource Description Framework) 作为知识表示的数据模型[3]，在 RDF 中知识使用 SPO 三元组 (Subject, Predicate, Object) 的形式存储。目前比较知名的开放 RDF 知识数据库有 DBpedia、Freebase 等。RDF 在发布之初定义了常用的 Predicate 关系，通过固定的 IRI 表示，统一的 IRI 定义可以实现不同知识之间的共享，但 RDF 定义中可以表示的知识有限：RDF 预定义的 Predicate 关系中没有区分概念和实体，也无法定义概念的属性和概念之间的关系，RDF 仅能表示 Subject 和 Object 之间的关系，没有泛化和抽象的表达能力。为了提高知识表达范围，在 RDF 的基础上提出了 RDFS (Resource Description Framework Schema)[2]，在知识数据存储之前需要定义知识的概念和关系等，对知识概念和关系的定义成为本体模型 (Ontology Model)。随后在 RDFS 的基础上，根据定义中的实际需求扩展了 OWL (WebOntologyLanguage)[3] 本体语法以及随后的 OWL2[4]，其中 OWL 相比于 RDFS 增加了数值属性和对象属性的不同定义，弥补了 RDFS 的定义中无法区分实体的属性以及实体之间的关系问题。OWL2 在 OWL 基础上增加了角色链，双关等特性定义，规范了表达技巧。目前 OWL2 已经成为本体建模的推荐标准，国际万维网组织 WWW (World Wild Web) 负责本体描述语言的标准制定。

1 相关研究介绍

主流知识图谱大致可以分为通用知识图谱

基金项目： 国家重点研发计划“网络空间安全”重点专项(2016YFB0800802)；山东省重点研发计划(2016ZDJS01A04，2017CXGC0706)。

作者简介： 魏玉良(1989-)，男，博士研究生，主要研究方向：知识图谱；黄 纯(1996-)，男，硕士研究生，主要研究方向：自然语言；王佰玲(1978-)，男，博士，教授，主要研究方向：网络与信息安全。

通讯作者： 王佰玲 Email：wbl@ hit.edu.cn

收稿日期： 2019-03-29

UKG 和领域知识图谱 DKG。UKG 是面向全领域信息构建知识表示和关联关系,强调的是广度,而 DKG 是面向特定的垂直领域构建知识关系,对于数据有更严格的前置数据模式和准确度要求,强调的是深度。DKG 在金融量化交易[5]、学者信息搜索[6]、智能教育[7]、历史研究[8]、生物医学[9]等垂直领域有广泛的应用。构建 UKG 和 DKG 时,两者之间的主要区别在于 UKG 一般使用“自底向上”的方法构建知识库,而 DKG 使用“自顶向下”的方法。UKG 的“自底向上”方法体现在利用开放式关系抽取(Open Information Extraction,OIE),通过语法结构分析文本中的实体和关系构建三元组,构建 DKG 的“自顶向下”方法需要在设计之初首先确定待解决的领域问题,通过本体建模的方法明确问题的范围、包含的实体以及实体的属性和关系,并且根据领域内的规律构建推理规则。DKG 与 UKG 之间相辅相成,DKG 可以从 UKG 中获取通用性的知识,而 DKG 本身就是 UKG 在具体领域的丰富和延展,为了通用性,DKG 在设计时需要考虑与 UKG 的兼容性。

近年来为了实现知识计算和共享,DKG 的研究逐渐增多[10]。文献[11]中介绍了目前自动构建本体模型的主要方法,通过自动识别实体,语法分析获取实体之间概念上的层级关系,文中指出目前自动构建方法主要针对层次关系(is-a 关系)的构建,而对于应用中的领域本体模型,大量非层次关系更为重要,因此自动构建的方法只能在领域实体和概念的发现过程中有所帮助。

从目前本体模型的研究可以发现:

(1)自动化构建本体模型的方法主要应用于 UKG 中的层级关系,在 DKG 中大量的非层级关系仍无有效地自动化构建方法,以领域专家人工构建为主。

(2)DKG 在各行各业中逐渐产生重要的作用,相比 DKG 指导工业应用和生产的价值更高。

(3)目前没有健全的 DKG 本体模型的构建思路和方法,ODP 的设计理念可以提高领域模型的设计规范,但仍处于工业探索阶段,仍需要大量的领域专家构建不同的 DKG 积累量变,逐步到质变的过程。

(4)公开的 ODP 中关于人物、机构、事件的 ODP 研究较少,目前定义最完善的人物本体是 Wikidata 的 Human 定义,多元关系定义依赖于传统的百科词典的词条转化,为了保证兼容性,定义冗余程度高,表意区分度不明显。

本文主要研究人物本体应该如何在满足本体要求的情况下,精简概括地建模,并提出包含多元关系的人物、机构、事件相关的 ODP,供构建领域知识图谱中与人物相关的本体模型参考。

2 人物本体建模案例分析

本体模型案例:“HA 在 2010 年 7 月从 OB 学校计算机专业研究生毕业,HA 的本科就读于 OB 校信息安全专业,2010 年 8 月 HA 进入 OD 公司工作,刚入职就非常有热情,工作积极主动,在 2012 年 12 月的“年度公司综合竞赛”中获得第一名的成绩,很快在 2013 年 4 月升职为项目负责人,独立带领团队。2015 年 5 月 OD 公司改组,HA 离开 OD 公司进入 OE 公司担任大数据分析组项目负责人,并工作至今。HA 的感情生活并不像事业那样如意,2011 年 12 月 HA 与 HF 结束三年的爱情长跑步入婚姻,但是因为种种原因,在 2014 年 4 月协议离婚,在 2015 年进入新公司后,遇到 HG 让 HA 再一次激起了爱情的火花,2016 年 3 月,HG 与 HA 组成新的家庭,并在 2017 年 10 月喜得千金。”

在例子中首先可以明确确定 4 个主要类别:Human 人物类别、Organization 机构类别、Event 事件类别、Position 职位类别,在 Human 与 Organization 关系中,还存在 Position 的职位属性,为了在知识库中进一步表示职位属性,需要综合考虑三个类别之间的关系。Postion 属于 Organization 的组成属性,公司中一定会包含各种不同的职位从 CEO、CTO 到普通雇员、HR 等。因此使用“hasPosition”属性关联 Organization 和 Position。Human 和 Postion 之间也可以通过类似“hasPosition”的方式关联,但是这样会产生歧义,如图 1 所示。由于 RDF 表示的知识中是没有时序性的,因此“ed.com/human/1”通过“hasPosition”只能表示“ed.com/human/1”曾经担任过“ed.com/pos/1”和“ed.com/pos/2”,但无法知道是在“ed.com/org/1”和“ed.com/org/2”公司中分别担任哪些职务,Human 和 Postion 之间的关联需要第 3 个实体 Organization 才能确定,这种涉及到多个不同实体之间的关联的关系属性称为多元关系(N-ary)。

OWL 通过 SPO 三元组表示的知识只能表达二元关系,但在真实数据中存在大量的多元关系(N-ary Relation),一个具体的多元关系 C_{R_n} 被定义为一种特殊的本体类,通过定义 C_{R_n} 的类关系确定多元关系中共现的不同本体类。对应前文中确定的 Human、Organization、Position 之间的关系,可以抽象为同一个 Employee 类表示多元关系,如图 2 所示。Employee 继承自 N-ary Relation 表明是一个关系

类，而不是对应的实体类。

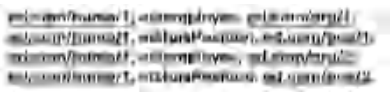

图 1　RDF 知识三元组示例

Fig. 1　RDF knowledge triple example

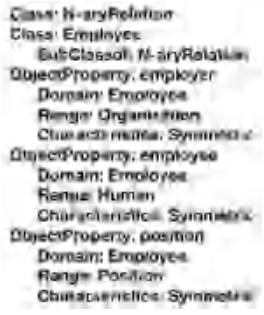

图 2　多元关系类 OWL 定义示例

Fig. 2　OWL N-ary relation class definition example

图 2 的定义中虽然实现了 3 个之间多元关系，但是进一步详细分析会发现对于一个 Employee 关系，Human 和 Organization 是固定的，而 Position 并不是唯一的，因一个人在一个公司可以担任多个职位。当增加时间属性时，问题会更加明显。一个 Employee 关系包含入职时间和离职时间，而对 Position 也需要描述具体职位的当选时间和离开时间，如果按照图 2 的定义，则需要在此基础上增加 4 个时间属性，如图 3 所示。从例中得知，HA 在 OB 公司从员工升职为项目负责人，因此需要创建 2 个 Employee 关系的实体，分别描述当员工时的信息和担任项目负责人时的信息，这 2 个实体中“workStartTime”和“workEndTime”重复出现，属于冗余的知识信息，在本体建模中需要避免冗余性的出现。

图 3　多元关系类属性定义示例

Fig. 3　N-ary relation class property definition example

为了更好的解释图 2 和图 3 中本体建模的问题，本文提出将本体模型部分映射到关系型数据库表的方法，利用数据库设计的 3NF 原则指出设计的不规范性，并将数据库设计的泛式原则转化为多元关系的定义准则。本体模型映射到关系型数据库的步骤如下。

(1)包含多元关系类的所有类分别转化为一张表，以类名作为表名。

(2)所有类的数值属性转化为表的字段，表的键值对应本体中属于该类别的实体 IRI。

(3)非多元关系类的关系属性独立生成一张关系表，表包含双键值，分别对应关系属性的 Domain 和 Range 类的实体 IRI；

(4)多元关系类表的键值是多元关系的关系属性中所有 Range 对应类的 IRI。

通过转化可以得到多元关系转化的关系型数据库，ER 图如图 4(a)所示，多元关系表中的“workStartTime”和“workEndTime”属性只依赖于 Human 和 Organization 键值，而不依赖于 Position 键值，违反数据库定义中第二范式原则“非主属性完全依赖于主关键字”，本例 Employee 表是多键值表，存在属性依赖于部分主键，而不是整体键值，因此需要进行修改。根据关系型数据库的修改规范，将只依赖于部分主键的属性独立成表，创建新的键值，原表中使用新表的键值代替原来的部分主键，如图 4(b)所示。得到转化的 ER 图后，根据从关系型数据库转化到本体模型的算法，可以转化为本体模型，再经过修改增加相应的属性描述。

3　复用现有本体模型

人物摘要本体模式，是为了给具体的领域知识图谱设计者提供基本概念的设计思路和复用泛式。本节介绍基于 Wikidata 的基础概念，结合前文中介绍的多元关系设计，给出本文中设计的人物摘要本体 ODP，方便其它领域知识图谱参考。图 5 是本文中涉及的人物摘要本体模型，主要的实体类和关系类。涉及的对象包括表示人物的 Human 类，表示机构的 Organization 类以及表示事件的 Event 类。为了领域知识图谱可以直接兼容 Wikidata 中现有数据，顶层继承关系承袭自 Wikidata 的 schema，Object、Subject、Agent、Individual、TemporalEntity 借鉴自 Wikidata 中的抽象概念，Subject 表示具有独特意识或独特个人经历的人，或与其它实体存在关系的实体；Object 描述与 Subject 相反的概念，表示物体不具有独立意识；Agent 表示能够执行行动的个人和可识别实体，可以在事件中担任行为的发起方；Individual 指人或特定物体；TemporalEntity 表示可以在一段时间内包含的内容，或者状态的变化。Wikidata 中构建了大量较为完善的抽象层概念，可以在此基础上通过多继承的方式丰富领域内实体的概念，便于实现逻辑上的推理和定理的描述。核心实体类包括 Organization、Human、Award、Event，分别

表示机构、人物、荣誉和事件,可以根据具体的领域问题方便增加社交网络属性信息,通过集成 Relation 类扩展网络中账号之间的关系,图中省略了地理位置和时间的定义,在存在基于时间和地理位置查询索引时,可以增加相关的实体设计。在通过扩展实体时,如果实体不具有主动意识,可以通过继承 Object 类进行定义,如增加作品实体的定义,可以适用于学者论文、明星作品等不同领域的知识表示。对于可以作为事件主动者的实体,可以继承自 Agent 类。新增多元关系时,可以参考已有的多元关系。

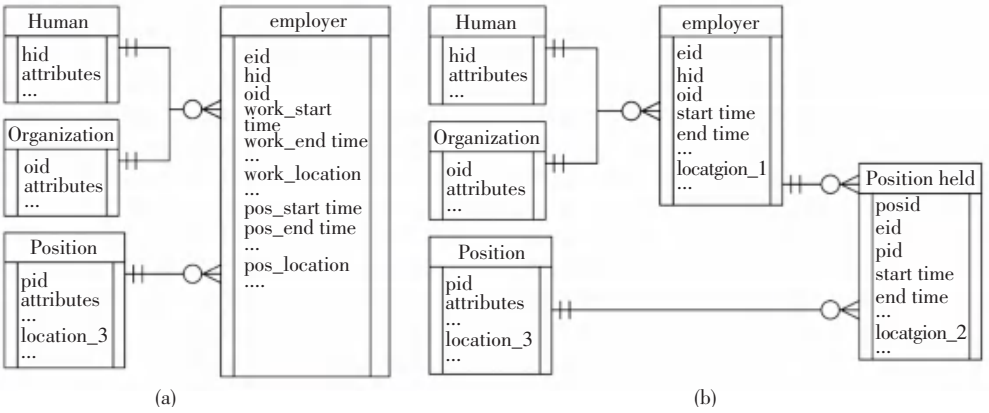

图 4 多元关系 ER 图表示示例

Fig. 4 Example of ER diagram for N-ary relation

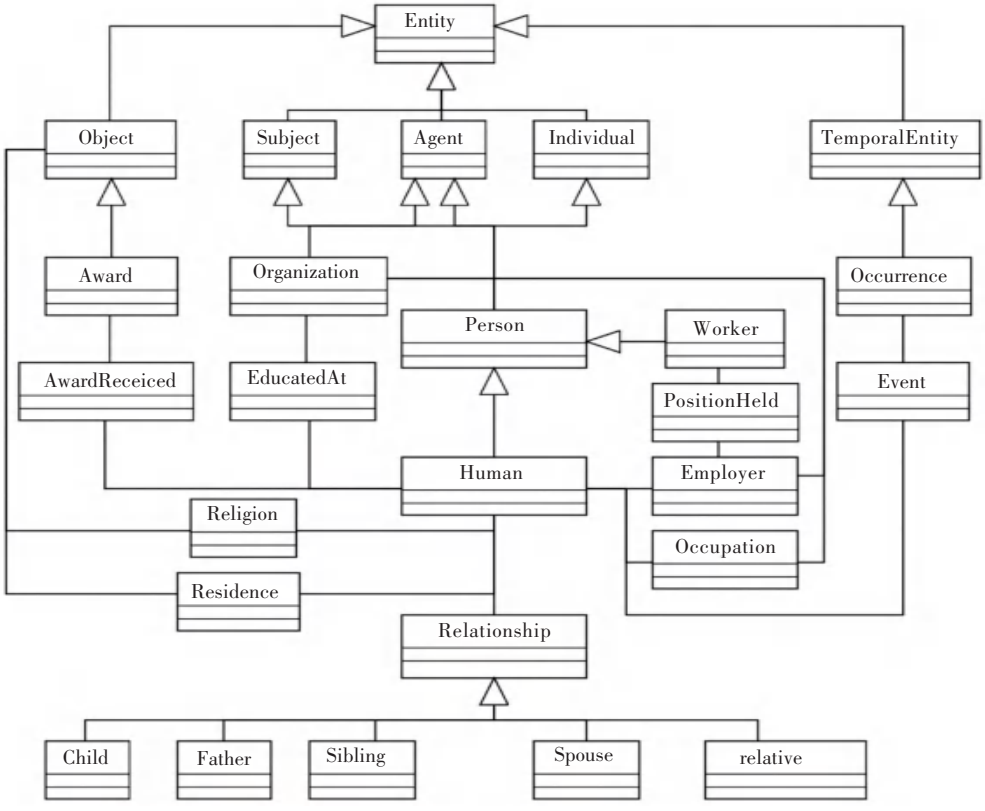

图 5 人物摘要本体 ODP

Fig. 5 Ontology design pattern of person abstract

4 结束语

本文中介绍了目前本体模型设计的基本语法结构和设计思路,并给出了通过二元关系表示多元关系的方法,通过例子分析了不同情况中多元关系的设计思路。其次针对多元关系设计中可能存在的冗余问题,本文提出本体模型到 ER 图的映射算法,通过数据库设计 N 泛式的规则又换多元关系设计。最后以 Wikidata 为主要模板,给出了人物摘要本体 ODP,便于在具体应用中知识图谱的设计参考。

(下转第 224 页)