

专家视图与本体视图的语义映射方法*

陈德彦^{1,2,4}, 赵宏^{1,2,4}, 张霞^{1,2,3,4}

¹(东北大学 计算机科学与工程学院, 辽宁 沈阳 110819)

²(计算机软件国家工程研究中心(东北大学), 辽宁 沈阳 110179)

³(沈阳东软智能医疗科技研究院有限公司, 辽宁 沈阳 110179)

⁴(东软集团股份有限公司, 辽宁 沈阳 110179)

通讯作者: 陈德彦, E-mail: deyan_chen@126.com



摘要: 由于本体具有强大的知识表示和推理能力, 本体已经在很多领域得到了广泛应用。然而, 本体的深入应用还面临着很多深层次的共性语义映射问题, 已有的本体建模方法仅仅提出了一些简单的指导原则和基本步骤, 使得知识工程师仍然无从下手。在基于领域专家知识构建领域语义知识库时, 针对领域专家知识中的一词多义、多元关系和安全需求等3类共性语义映射问题, 从5个方面展开了深入研究, 提出了相应的语义映射方法, 总结了10条本体建模约定。最后构建了一个完整的应用案例, 并对所提出的5类语义映射方法进行了评价。

关键词: 本体; 领域语义知识库; 专家视图; 本体视图; 语义映射; 问题求解

中图法分类号: TP182

中文引用格式: 陈德彦, 赵宏, 张霞. 专家视图与本体视图的语义映射方法. 软件学报, 2020, 31(9): 2855–2882. <http://www.jos.org.cn/1000-9825/5820.htm>

英文引用格式: Chen DY, Zhao H, Zhang X. Semantic mapping methods between the expert view and the ontology view. Ruan Jian Xue Bao/Journal of Software, 2020, 31(9): 2855–2882 (in Chinese). <http://www.jos.org.cn/1000-9825/5820.htm>

Semantic Mapping Methods Between Expert View and Ontology View

CHEN De-Yan^{1,2,4}, ZHAO Hong^{1,2,4}, ZHANG Xia^{1,2,3,4}

¹(School of Computer Science & Engineering, Northeastern University, Shenyang 110819, China)

²(National Engineering Research Center for Computer Software (Northeastern University), Shenyang 110179, China)

³(Neusoft Research of Intelligent Healthcare Technology, Co. Ltd., Shenyang 110179, China)

⁴(Neusoft Corporation, Shenyang 110179, China)

Abstract: Due to the powerful knowledge representation and reasoning ability, the ontology has been widely used in many domains. Nevertheless, the in-depth application of the ontology still faces many deep common semantic mapping problems. The existing ontology modeling methods only propose some simple guiding principles and basic steps, so that knowledge engineers still have no way to start. When constructing the domain semantic knowledge base based on the domain expert knowledge, the three types of common semantic mapping problems in the domain expert knowledge, such as polysemy, n -ary relationships, and security requirements, are studied in depth from five aspects, the corresponding semantic mapping methods are proposed, and 10 ontology modeling conventions are summarized. Finally, a complete application case is constructed and the five types of semantic mapping methods proposed in this study are evaluated.

Key words: ontology; domain semantic knowledge base; expert view; ontology view; semantic mapping; problem solving

* 基金项目: 国家自然科学基金(61232015); 国家高技术研究发展计划(863)(2015AA020103); 国家重点研发计划(2016YFC1303000); 沈阳东软智能医疗科技研究院有限公司开放课题基金(NRIHTOP1802)

Foundation item: Key Projects of National Natural Science Foundation of China (61232015); National High-Tech Research and Development Plan of China (863) (2015AA020103); National Key Research and Development Program of China (2016YFC1303000); Open Program of Neusoft Research of Intelligent Healthcare Technology, Co. Ltd. (NRIHTOP1802)

收稿时间: 2018-02-10; 修改时间: 2018-08-12; 采用时间: 2018-10-29

本体(ontology)是共享概念模型的明确的形式化的规范说明^[1],其提供了一种结构化地表示领域知识的形式化方法,并提供了推理能力,构造本体可以实现某种程度的知识共享和重用.由于本体具有的强大的知识表示和推理能力,已经在很多领域得到了广泛的应用,例如语义 Web^[2]、知识工程、自然语言处理、信息获取、信息集成、生物医学等领域,用于领域问题求解、异构信息源之间的交互、辅助组织中人与人之间的沟通等.基于本体的信息建模主要借助于一个互补的语言集合,该集合包含 3 种语言^[3]:资源描述框架(resource description framework,简称 RDF)^[4]、RDF 词汇定义(RDF schema,简称 RDFS)语言^[5]、Web 本体语言(Web ontology language,简称 OWL)^[6].由于直接使用这些本体描述语言来进行本体建模很不方便,也难以掌握,于是出现了一些图形化的本体编辑工具,例如 Protégé^[7]、OntoEdit^[8]、KAON^[9]等.这些本体编辑工具很好地简化了领域知识的形式化描述过程,但却并不能很好地帮助指导如何建模领域知识库,而这正是本体工程^[10]所要研究和解决的核心问题.与软件工程、面向对象设计和知识工程类似,本体工程研究本体构建的方法学,涉及本体构建的过程、采用何种工具和语言、以何种顺序使用这些工具和语言、如何应用这些工具和语言、质量控制和资源管理等.针对不同学科领域、工程问题或应用场景,目前有多种本体构建方法,例如面向对象的本体建模方法^[11,12]、基于层次的本体建模方法^[13,14]、面向具体学科领域或工程问题的本体建模方法^[15,16]等.但是目前尚没有一种标准的、适用于所有学科领域或应用场景的本体建模方法,实际上也不可能存在建模本体的唯一的正确的方法,最好的解决方案总是依赖于具体的应用,对本体质量的评价也唯一取决于使用它的应用.因为构建本体本身并不是一个最终目标,而是为了应用的需求而提供一套数据集和它们之间的结构,而且本体开发总是一个迭代和不断精化的过程^[17].无论采用何种方法进行本体开发,一般应遵循 Gruber 提出的 5 条原则^[18]:清晰性(clarity)、一致性(coherence)、可扩展性(extendibility)、编码偏好程度最小(minimal encoding bias)、本体约定最小(minimal ontological commitment).

现有的本体建模方法都只是对建模过程从不同的方法学、学科领域或工程问题的角度提出了一些简单的指导原则和基本步骤,而在进行实际的领域语义知识库建设时,会面临大量更加具体的共性问题或者特定问题,使得知识工程师仍然无从下手.目前公认领域本体的开发需要领域专家的参与,并由知识工程师将领域专家提供的领域知识建模并形式化为可被计算机处理和共享利用的领域本体知识.于是,从专家角度看到的知识视图(简称专家视图,即领域知识模型)和从知识工程师角度看到的本体知识视图(简称本体视图,即本体知识模型)可能是不同的.专家并不理解也不关心基于本体的知识建模和形式化方法,而只关心知识对人的易理解性和易使用性.知识工程师建立本体知识视图时更多考虑的是本体描述语言的描述能力、Gruber 提出的 5 条原则、领域应用需求的可满足性、知识服务实现的合理性和方便性等.由于领域知识本身的复杂性和这两种视图表示方法和目的的不同,专家视图和本体视图几乎不可能是一样的,也即需要建立专家视图和本体视图之间的正确语义映射,才能真正实现领域知识的存储、共享、集成和复用.基于大量实践发现:这个映射过程存在大量需要解决的关键问题,而其中一些关键问题是在建模面向不同学科领域或工程问题的不同类型本体时面临的共性问题,例如:

- 1) 在专家视图中,一词多义的现象很普遍,资源对象的语义一方面通过采用某种语言描述的资源名称的含义来表达,另一方面,通过资源对象应用的领域上下文来丰富或限定其语义.以健康医疗领域为例,术语“疾病”有时指一个概念,代表所有疾病的集合,有时指一个属性,代表某个疾病诱因关联的疾病;“泌尿道感染”可能代表一种疾病,也可能是另一种疾病的诱因、风险因素或表现症状;“糖尿病”通常指一个疾病实例,但“糖尿病”还可以细分为“1 型糖尿病”、“2 型糖尿病”、“妊娠糖尿病”等,这时“糖尿病”又表示一些实例的集合,即代表一种疾病类型等.而在本体视图中,某个资源对象一旦建立,其 ID 就是唯一的,当然可以明确声明其与另一个资源对象在语义上是等价的或者完全不同的;
- 2) 在任何领域的知识描述中,都会涉及到对资源的某些量化指标的描述,例如人的身高、体重、血糖、血压等.这些指标实际上是一个结构化的值,由数值和对应的单位构成,因为在不同的语境下(例如在不同国家),对相同指标的描述其默认指称单位可能并不一样,即同一个属性存在不同方面的值.另外,在任何领域都存在一些模糊和不确定的知识,例如医生对患者疾病的诊断或风险评估就是一种可能

性的判断,不但要表达患者与疾病的关系,还要描述这种关系的可能性大小或者风险等级.这些结构化值和不确定关系的描述都涉及到多元(n -ary)关系,而本体只能描述二元(binary)关系;

- 3) 访问授权的语义映射.领域本体库提供领域中共同认可的知识,一般不涉及安全或隐私问题^[19];领域实例数据提供关于个体的知识,可能存在显式的关于个体的敏感或隐私信息.基于领域本体库和领域实例数据的语义推理,可能揭示隐含的关于个体的敏感或隐私信息.针对领域中涉及安全、敏感或用户隐私相关的内容,领域安全专家、信息拥有者(例如医院或患者)或信息来源者(例如患者)都可能提出访问授权的需求,需要基于领域语义知识库建立此访问授权的语义映射.

以上列举的部分语义映射的共性问题,在目前基于本体的知识建模方法、建模工具、知识获取、知识复用的研究中并未进行深入探讨,这阻碍了本体在各个领域中的广泛应用.为此,本文就以上列举的 3 类语义映射的共性问题从 5 个方面展开了深入研究,提出了相应的解决方法,总结了 10 条本体建模约定.最后,对这 5 类语义映射方法进行了应用验证和评价.

本文第 1 节给出相关定义.第 2 节针对上述的 3 类语义映射的共性问题从 5 个方面进行深入探讨,提出相应的解决方法.第 3 节采用本文提出的语义映射方法构建一个完整的应用案例.第 4 节对本文的方法进行评价.第 5 节介绍相关研究工作及其存在的不足.最后一节对本文进行总结,并提出进一步的工作.

1 相关定义

定义 1(领域专家知识(domain expert knowledge)或领域专家视图(domain expert view)). 区别于领域大数据中蕴含的碎片化的或者隐含的领域知识,领域专家知识是指领域专家整理的系统化的、显式的领域专业知识以及专家积累的领域经验知识,领域专家知识的形态可以是图、表、文字等多种形式.领域专家知识主要是给人阅读的,而不是计算机可处理的,需要由知识工程师将其映射为领域语义知识库中的知识,以实现领域知识的共享、集成和复用.记领域专家视图为 EKB_{domain} ,其定义如下:

$$EKB_{domain}=\{ek_1,ek_2,\dots,ek_i,\dots,ek_n\},n\geq 1,$$

其中, ek_i 表示其中一位领域专家所贡献的领域知识.

定义 2(领域本体定义(domain ontology schema)). 领域本体定义^[3]通过捕捉领域中共同认可的概念(concept)、概念属性、概念分类关系、概念间存在的语义关联关系及相关约束来描述领域知识.记 O_{domain} 表示领域本体,其定义如下:

$$O_{domain}=(C,A,R,X,I).$$

其中,

- C 表示概念(concept)集,概念又称为类(class),用于表达具有某类相似特征的个体(individual)的集合,个体又称为实例(instance).例如,概念“人(person)”代表所有个体人的集合;
- A 表示所有概念的属性(attribute)集,概念的属性又称为数据类型属性(data type property),它描述概念所包含的实例本身的特征,例如人的姓名、性别、出生日期、身高、体重等;
- R 表示语义关系(semantic relation)集,语义关系又称为对象属性(object property),包含分类关系(taxonomic relation)和非分类关系(non-taxonomic relation),即 $R=R_{vtax}\cup R_{htax}$:
 - R_{vtax} 表示概念、属性或语义关系的纵向分类关系,例如,概念“人”可以进一步划分为“男人”和“女人”两个子概念;
 - R_{htax} 表示横向的非分类关系,例如人的个体间存在的好友关系、父子关系、兄弟关系等;
- X 表示公理集,公理(axiom)用于定义概念、属性和关系之上的约束,比如限定概念某个属性的数量(基数约束)或者在某个属性之上的取值(值约束)、限定属性和关系的定义域(domain)和值域类型(range)等,例如:人的出生日期只有一个,但不同人可以具有相同的出生日期,人在生物学意义上的父亲和母亲都是唯一的;出生日期属性的定义域类型为人,值域类型为日期类型等;
- I 表示实例数据集,用于描述领域中共同认可的常识知识,例如内分泌和代谢类疾病包含糖尿病、甲亢

等.领域本体定义类似于面向对象中的类的定义,领域本体定义中一般不包含实例数据,除非用于表达一般性的领域常识知识,但仅包含实例数据的 RDF 描述片段不能称之为本体定义^[3].

定义 3(领域本体库(domain ontology base)). 领域本体库由一个或者多个领域子本体构成,它们分别用于描述领域中某个方面的知识.记 OB_{domain} 表示领域本体库,其定义如下:

$$OB_{domain}=\{O_1,O_2,\dots,O_i,\dots,O_n\},n\geq 1.$$

通常,一个本体不可能定义关于某个领域的全部知识,它可能只定义该领域中某一个方面的知识. O_i 表示领域本体库中的某个子本体,领域本体库的子本体划分和组织方式依赖于领域知识体系的构成、领域语义知识库服务的应用需求和采用的领域语义知识库建模方法.

定义 4(领域实例数据(domain instance data)). 领域实例数据^[3]为基于领域本体库中的定义来描述的领域中的常识知识或个体的知识.例如,基于医疗健康本体库描述某个人的基本信息和健康状态.领域常识知识为领域中公认的知识,通常位于领域语义知识库中.领域中具体环境或个体的知识,通常作为基于领域语义知识库的知识服务的输入,以获得问题解.记 I_{domain} 表示领域实例数据,其定义如下:

$$I_{domain}=\{(s,p,o)|s\in I,p\in A\cup R,o\in I\cup V\}.$$

其中,

- (s,p,o) 表示描述实例数据的陈述或称为三元组(triple);
- s 为某个实例对象;
- I 表示实例对象集;
- p 表示用于描述实例对象的属性或者语义关系;
- A 和 R 分别表示描述 I 中实例对象所用到的属性集和语义关系集;
- o 表示属性或语义关系的取值,或者为实例对象,或者为字面值(literals)^[4,10];
- V 表示字面值的集合.

定义 5(领域实例库(domain instance base)). 领域实例库由一个或多个领域实例数据片段组成,它们分别用于描述领域中不同方面的常识知识或不同个体的知识,当然也有可能描述某个个体不同方面的知识,例如描述某个人的健康状况、研究兴趣、个人爱好等.记 IB_{domain} 表示领域实例库,其定义如下:

$$IB_{domain}=\{I_1,I_2,\dots,I_i,\dots,I_n\},n\geq 1.$$

定义 6(领域语义规则集(domain semantic rule set)). 领域语义规则同时服务于领域本体的定义和领域实例数据的描述,一方面用于描述领域中领域专家所获得的启发式经验知识,另一方面用于补充本体描述语言的语义描述能力.记 F_{domain} 表示领域语义规则集,其定义如下:

$$F_{domain}=\{r_1,r_2,\dots,r_i,\dots,r_n\},n\geq 1.$$

其中, r_i 表示其中一条语义规则.语义规则是典型的条件语句:if-then 子句,只有当特定陈述(statement)集合为真时,才会添加新的知识.语义 Web 层次结构^[2,3]提供了多种知识表示形式,包括从 RDF 到最新版本的 OWL 等多种格式,每一层都对表达能力进行了进一步的扩展,并且允许用户根据语义程序具体所需的语义量来采用相应的表示方式.但本体描述语言在表达能力和灵活性方面仍然存在一些不足,语义规则用于扩展本体描述语言的描述能力以及灵活性.W3C 建议的语义规则描述语言为语义 Web 规则语言(semantic Web rule language,简称 SWRL)^[20].

定义 7(领域语义知识库(domain semantic knowledge base)或本体视图(domain ontology view)). 领域本体库、领域实例数据和领域语义规则集一起构成了领域语义知识库.

记 SKB_{domain} 表示领域语义知识库,其定义如下:

$$SKB_{domain}=(OB_{domain},IB_{domain},F_{domain}).$$

本体定义的结束,便是知识库构建的开始^[17].基于领域本体库中的本体定义和语义规则来描述具体的实例,形成实例数据和语义规则集,以领域本体库作为领域背景知识,以实例数据和语义规则集作为具体的知识,它们一起形成了面向特定应用需求的语义知识库.例如,基于医疗健康领域本体库构建面向慢性病患者的健康风险

评估、疾病自我诊断、疾病辅助诊断、疾病干预方案(药物、运动、饮食、心理、睡眠等)、远程监护服务、健康知识问答服务、健康教育/咨询服务等需求的语义知识库。

W3C 推荐的针对语义知识库的语义层(即 RDF 层)查询标准语言为 SPARQL^[21],这种查询语言不仅理解 RDF 的语法,而且理解 RDF 的数据模型和 RDF 词汇的语义,几乎所有的 RDF 查询工具都提供了对 SPARQL 查询语义的支持^[10,19]。

2 语义映射方法

以本体视图来建模领域知识,目的是为了实现领域知识的存储、共享、集成和复用.为此,需要基于共同认可的标准或规范来对本体视图进行形式化描述,以实现计算机可准确理解其语义并可做出正确的推理.本文对本体视图的描述完全遵循 W3C 推荐的描述语言规范(RDF,RDFS,OWL 等),使用这些标准本体描述语言提供的语义组件来描述领域专家知识。

本节针对专家视图与本体视图的语义映射中的 5 个通用问题进行了深入探讨,给出了相应的解决方案,总结了 10 条约定,并分别通过实例对映射方法进行了详细说明。

2.1 资源名的语义映射

针对引言中提到的一词多义的语义映射问题有很多类型,本节探讨专家视图中的资源在本体视图中的命名规则,以及在本体视图中属性的定义位置。

第 2.2 节和第 2.3 节分别探讨了另外两类问题及其语义映射解决方案。

如图 1(a)所示的专家视图,如果进行直接的语义映射,将存在如下 3 个问题。

- 1) 在图 1(a)中,专家同时定义了“疾病”概念和“疾病”属性、“症状”概念和“症状”属性,这对于人类理解并没有问题,但在本体视图中,概念和属性对应的语义类型分别为 owl:Class 和 rdfs:Property,其定义和语义约束是完全不同的,无法将同一个 ID 表示的资源同时声明为类和属性;
- 2) 在图 1(a)中,“疾病”属性对应的定义域类型分别有“人”“风险因素”“医学实验”和“疾病诱因”;“症状”属性对应的定义域类型有“人”和“疾病”,即分别通过属性“疾病”“症状”建立了多个概念之间的语义关系.同样,这对于人类理解并不会产生歧义;但如果在本体视图中分别建立对应的对象属性“med:diseases”和“med:symptoms”来描述这些概念之间的语义关系,或者不声明属性的定义域,或者将属性的定义域同样声明为多种类型,这都将引起语义推理结果的歧义.例如,基于推理结果认为概念“人”“风险因素”“医学实验”和“疾病诱因”在语义上是等价的,因为它们的实例都可以作为属性“med:diseases”的主语.对于症状属性,存在同样的问题;
- 3) 从本体工程的角度,同一个专家视图可能被映射为一个或者多个本体视图.在本体视图中,由于属性具有方向,属性定义域和值域声明中对应的概念的定义可能位于不同的 O_{domain} 中,属性连接的两个实例的定义也可能位于不同的 I_{domain} 中.这种情况下,属性定义的位置约定就很重要,否则可能导致属性的冗余定义(有可能不同位置定义的属性其语义约束并不完全一致),进而导致知识的冗余定义,同时可能会引起语义理解和语义推理上的歧义.例如在图 1(b)中,疾病、症状等概念的定义位于 $O_{medicine}$ 本体中,概念“人”的定义位于 O_{people} 本体中,那么在定义人与疾病或者人与症状之间的语义关系时,需要决定该语义关系应该定义在哪个本体中.如果分别定义,可能会导致属性冗余定义,除非明确声明它们是等价的(通过 owl:equivalentProperty)或者互逆的(通过 owl:inverseOf),否则可能会引起语义理解和语义推理上的歧义。

在 SKB_{domain} 中,除了字面值(数值、布尔值、字符串等)和规则以外的所有对象都被称为资源(resource).资源是我们想要讨论或记录的对象或事物^[10],这些资源可以是物理存在并可通过网络访问的资源,例如网页、网络图片、Web 服务等;或是物理存在但不可通过网络访问的资源,例如人、公司、图书馆中的书籍等;或是物理不存在的抽象概念,例如本体、语义 Web 等.RDF 将资源定义为任何可被 URI(uniform resource identifier)引用(URI references,简称 URIs)标识的事物^[4].URI 引用是一个在尾部附加了可选的片段识别符(fragment identifier)的

URI.例如,在这个 URI 引用“http://www.example.org/index.html#section2”中,符号#前面的部分为 URI,#后面的部分为片段标识符.URI 引用可以包含 Unicode 字符,这就允许在 URI 引用中使用多种语言.因此,使用 URIs,RDF 实际上可以描述任何事物,并陈述这些事物之间的关系.在领域语义知识库中,除了不使用 URI 引用格式表示的字面值以外,其余对象都为资源,包括概念、属性、语义关系和实例对象.

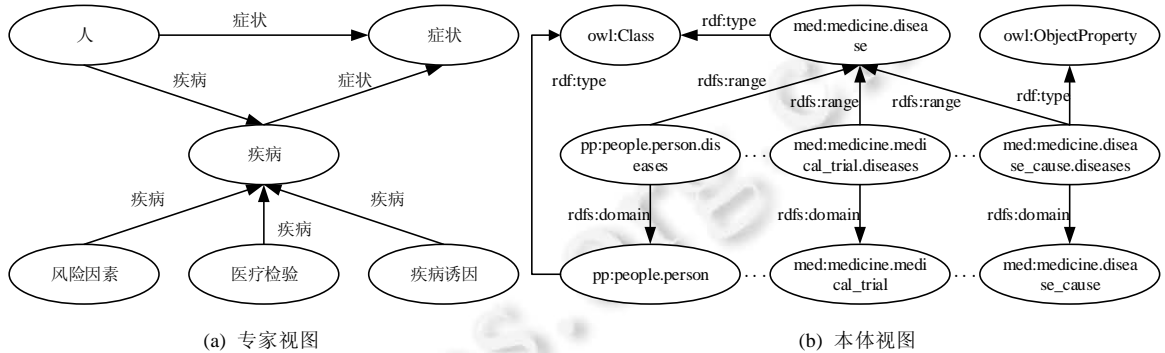


Fig.1 Semantic mapping of resource names

图 1 资源名的语义映射

在本体视图中,资源 ID 采用 URIs 进行标识,只需要保证其唯一性即可,这个 ID 对于语义推理来说没有任何语义贡献,因为推理是基于本体描述语言中定义的语义组件的语义(即本体推理)和定义的语义规则的语义(即规则推理)来进行的.由于在 SKB_{domain} 中, OB_{domain} 中的对象规模相比 IB_{domain} 中的对象规模少很多,几乎不是一个数量级的,比如一个疾病概念下的疾病实例可能有几万种.所以一般情况下,可以对 OB_{domain} 中的资源对象取一个不重复的、有意义的人类可识别 ID.由于实例的规模非常大,对于 IB_{domain} 中实例的 ID,要取一个不重复的、有意义的名称比较困难,所以可以根据领域特定上下文或应用需求采用某种约定的编码规则,例如,健康医疗领域中的疾病诊断的 ICD-10 编码^[22]、临床操作的 ICD-9-CM 编码^[22]等;或者取一个不重复的随机数,例如 Freebase^[23]知识库中的对象 ID 采用 MID(a machine identifier)进行标识,例如 ns:m.0c58k 表示“糖尿病”实例的 ID,ns 表示 Freebase 的默认名称空间.对于实例的人类可识别名称和描述,可以通过本体描述语言提供的标注属性(比如 rdfs:label,rdfs:comment)去体现.

约定 1. 在 OB_{domain} 中,概念和属性 ID 的命名规则如下.

- 概念:namespace:domain.concept;
- 属性:namespace:domain.concept.property.

其中,

- namespace 表示领域的名称空间(例如 med 表示 medicine 领域的名称空间);
- domain 表示领域 ID,名称空间内唯一;
- domain.concept 表示以领域 ID 为前缀的概念 ID,名称空间内唯一;
- domain.concept.property 表示以概念 ID 为前缀的属性 ID,名称空间内唯一.

这种命名规则约定,对于机器处理和人类理解都是有价值的.语义描述的原则是,尽可能清晰明确地描述所有对象的语义(包括语义约束)及与其他对象的语义关系(包括语义约束).图 1(b)为基于约定 1 对这些概念间的语义关系的定义,明确给出了属性的定义域和值域限定,无论对于计算机推理还是人类理解都不会存在歧义.rdf,rdfs,owl 分别为本体描述语言 RDF,RDFS 和 OWL 提供的语义组件定义的名称空间前缀,med 和 pp 分别为本体 $O_{medicine}$ 和本体 O_{people} 的名称空间前缀,在本体模型的形式化描述中进行声明.

针对以上的问题 3),这里做出如下的本体建模约定.

约定 2. 如果属性的定义域所对应的概念定义在某个 O_{domain} 中,那么也将该属性定义在此 O_{domain} 中.

约定 2 的含义是:无论是数据类型属性还是对象属性,根据其定义域所对应的概念所在的本体来确定该属性定义所在的本体.这样做的理由是:属性用于描述概念所代表的实例集中的实例的特征或者这些实例指向其他概念代表的实例集中的实例的语义关联,所以属性的定义理应与其服务的概念的绑定在一起.如图 1(b)所示:将描述概念“人”的疾病和症状的属性定义在 O_{people} 本体中,将描述 $O_{medicine}$ 本体中概念的属性定义也定义在 $O_{medicine}$ 本体中.

某些情况下,可能需要通过导入其他本体中的概念,并对其属性进行扩展,使得属性定义和其定义域声明中的概念的定义位于不同的本体中.但通常情况下,在定义某个新的概念时,往往都会同时定义此概念的一些属性.此时,应尽可能将概念和描述此概念的属性(以此概念为定义域的属性)定义在一起,以避免属性的冗余定义.

2.2 同一实例具有多种类型或角色

健康医疗领域的专家可能构建如图 2(a)所示的专家视图,在多个概念的实例定义中同时包含了糖尿病、泌尿道感染等实例.而在本体视图中,每个实例对象的 ID 必须是唯一的,但多个 ID 不同的实例对象,在语义上可以是等价的.在本体视图中,某个实例对象也可以具有多种类型,或者属于由多种类型通过复杂类^[6]构造方法构造的复杂类.

约定 3. 在专家视图中,基于不同的语义类型定义了多个相同的实例时,在对应的本体视图中可以定义一个唯一的实例,并声明其具有多种语义类型.在进行语义理解或语义推理时,实例对象在具体语境下的语义类型由连接该实例的属性的定义域或值域的类型来确定.

例如,图 2(a)所示的专家视图对应的本体视图如图 2(b)所示,其中,白色椭圆表示类和属性,灰色椭圆表示实例.下同.

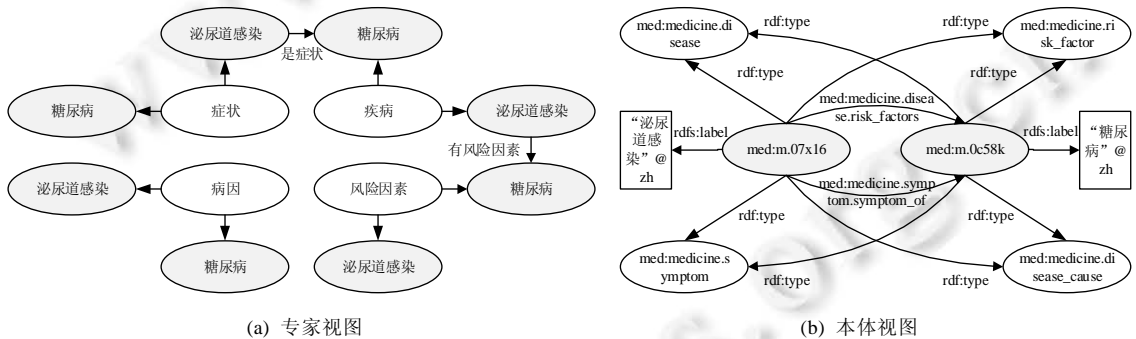


Fig.2 Instance objects have multiple types

图 2 实例对象具有多种类型

下面为采用 Turtle^[3]语法描述的 $O_{medicine}$ 本体的一个片段.

1. med:medicine.disease.risk_factors rdf:type owl:ObjectProperty;
2. rdfs:domain med:medicine.disease;
3. rdfs:range med:medicine.risk_factor;
4. owl:inverseOf med:medicine.risk_factor.diseases.
5. med:medicine.symptom.symptom_of rdf:type owl:ObjectProperty;
6. rdfs:domain med:medicine.symptom;
7. rdfs:range med:medicine.disease;
8. owl:inverseOf med:medicine.symptom.symptom_of.
9. med:m.07x16 rdfs:label “泌尿道感染”@zh;
10. rdfs:label “Urinary tract infection”@en;

11. `rdf:type med:medicine.disease,`
12. `med:medicine.disease_cause,`
13. `med:medicine.risk_factor,`
14. `med:medicine.symptom.`
15. `med:m.0c58k rdfs:label “糖尿病”@zh;`
16. `rdfs:label “Diabetes mellitus”@en;`
17. `rdf:type med:medicine.disease,`
18. `med:medicine.disease_cause,`
19. `med:medicine.risk_factor,`
20. `med:medicine.symptom.`
21. `med:m.07x16med:medicine.disease.risk_factors med:m.0c58k.`
22. `med:m.07x16 med:medicine.symptom.symptom_of med:m.0c58k.`

第 1 行~第 4 行为对属性 `med:medicine.disease.risk_factors` 的定义的一个片段;第 5 行~第 8 行为对属性 `med:medicine.symptom.symptom_of` 的定义的一个片段;第 9 行~第 14 行是对实例 `med:m.07x16` 的声明,该实例被声明为了多种类型;第 15 行~第 20 行是对实例 `med:m.0c58k` 的声明,该实例也被声明为了多种类型;第 21 行、第 22 行是对实例 `med:m.07x16` 和 `med:m.0c58k` 之间语义关系的声明.基于属性 `med:medicine.disease.risk_factors` 的定义和第 21 行中的实例语义关系声明,可以推理出在第 21 行中的实例 `med:m.07x16` 的类型为 `med:medicine.disease`,实例 `med:m.0c58k` 的类型为 `med:medicine.risk_factor`;而基于属性 `med:medicine.symptom.symptom_of` 的定义和第 22 行中的实例语义关系声明,可以推理出在第 22 行中的实例 `med:m.07x16` 的类型为 `med:medicine.symptom`,实例 `med:m.0c58k` 的类型为 `med:medicine.disease`.

在执行 SPARQL 检索时,语义也是很清晰的.例如,在以上 $O_{medicine}$ 的本体片段中,实例 `med:m.0c58k` 被同时声明为了 4 种类型:`med:medicine.disease`,`med:medicine.disease_cause`,`med:medicine.risk_factor` 和 `med:medicine.symptom`.当分别检索这 4 种类型的实例时都会包含 `med:m.0c58k`,这也符合医学上的解释:在医学上,糖尿病是一种疾病,同时也是其他疾病的诱因、风险因素和表现症状.在基于图模式执行复杂语义关联的检索时,基于实例的类型声明和属性的定义域和价值域限定,可以检索出正确的结果.如果语义关系中的实例类型不符合属性的定义域和价值域限定,在本体构建过程中可以通过语义一致性检查进行正确识别,从而进行修正.例如 Protégé 就提供了本体创建过程中的语法和基于推理的语义一致性检查机制.

在图 2(a)所示的情况中,类似“泌尿道感染”、“糖尿病”这样属于多种类型的实例可能并不是很多,大部分病因可能只是一种病因,而不是一种疾病;同样,大部分症状可能只是代表一种症状,而不属于疾病.而有些情况下,几乎所有实例都可以映射到多种类型下的某种子类型,如图 3(a)所示的专家视图.在图 3(a)中,运动项目从不同角度进行分类,而羽毛球(单打)、羽毛球(双打)实例可以映射到每种分类方法中的某个子类中.在此专家视图中,虽然将运动项目按不同科目类型、不同强度、不同能量代谢类型和不同运动益处进行分类,但它们都属于运动项目的子类,对于描述运动项目的属性和语义关系,也并不需要按子类的不同进行区分.而在图 2(a)所示的专家视图中,疾病、症状、病因和风险因素是完全独立的一级概念,它们在语义上也是完全不同的.

约定 4. 在专家视图中,所有实例都归属于同一祖先类下的不同子类时,在对应的本体视图中,可以将这些子类建模为类型实例,实例对象通过分类属性连接这些类型实例.

基于约定 4,对于图 3(a)所示的专家视图,可以建立如图 3(b)所示的本体视图.在如图 3(b)所示的本体视图中,针对运动项目实例,只建立了一个运动项目一级分类,并将所有运动项目建模为其实例;而将专家视图中的不同特化(specialization)分类建模为类型实例,同时增加多个对应的分类属性来连接运动项目实例和这些类型实例.

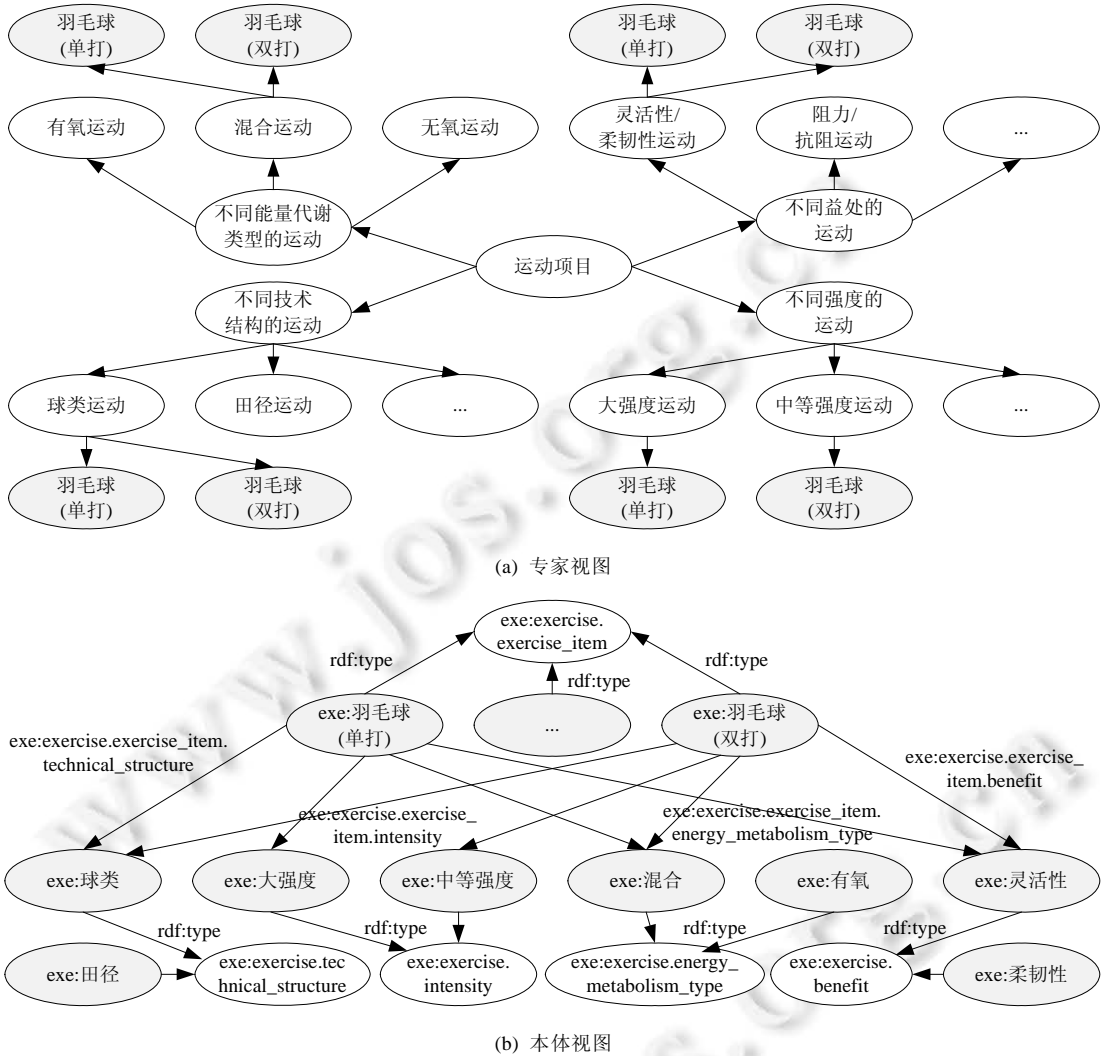


Fig.3 Instance objects have multiple classification properties
图3 实例对象具有多种分类属性

2.3 同一资源既是类型又是实例

在所有领域的知识中都会存在这样的情况,即某个资源自身既表示实例又可以作为分类代表更多实例对象的集合.例如在健康医疗领域,糖尿病和高血压都属于典型、多发的慢性疾病,大多数时候都表示疾病实例,但糖尿病还可以进一步细分为 1 型糖尿病、2 型糖尿病、妊娠糖尿病、糖尿病前期、糖尿病性神经病等多种疾病,如图 4(a)所示的专家视图.

参考 ICD-10 可以知道,1 型糖尿病还可以进一步细分为脆弱型糖尿病、幼年型糖尿病、趋酮症性糖尿病等疾病.同样,高血压等疾病也可以做进一步的细分.

针对以上情形,尽管 OWL 提供了这样的表达能力,但我们需要在表达能力和推理能力上做出平衡.OWL 提供了 3 种表达能力递增的子语言 OWL Lite、OWL DL 和 OWL Full,以分别用于特定的实现者和用户团体^[3,6]. OWL Lite 仅提供用于描述分类关系和简单约束的语义组件.OWL DL(description logic)包括了 OWL 语言的所有成分,但有一定的限制,如类型的分离,即一个类不能同时是一个个体或属性,一个属性不能同时是一个个体

或类.采用 OWL DL 语义组件描述的本体能够确保语义推理计算的完全性(computational completeness)和可判定性(decidability).OWL Full 具有最强的表达能力和完全自由的 RDF 语法,例如,在 OWL Full 中,一个类本身可以作为一个个体,并允许一个本体增加预定义的(RDF,OWL)词汇的含义,但由此带来的问题是,没有推理系统能够支持对 OWL Full 所有成分的完全推理.

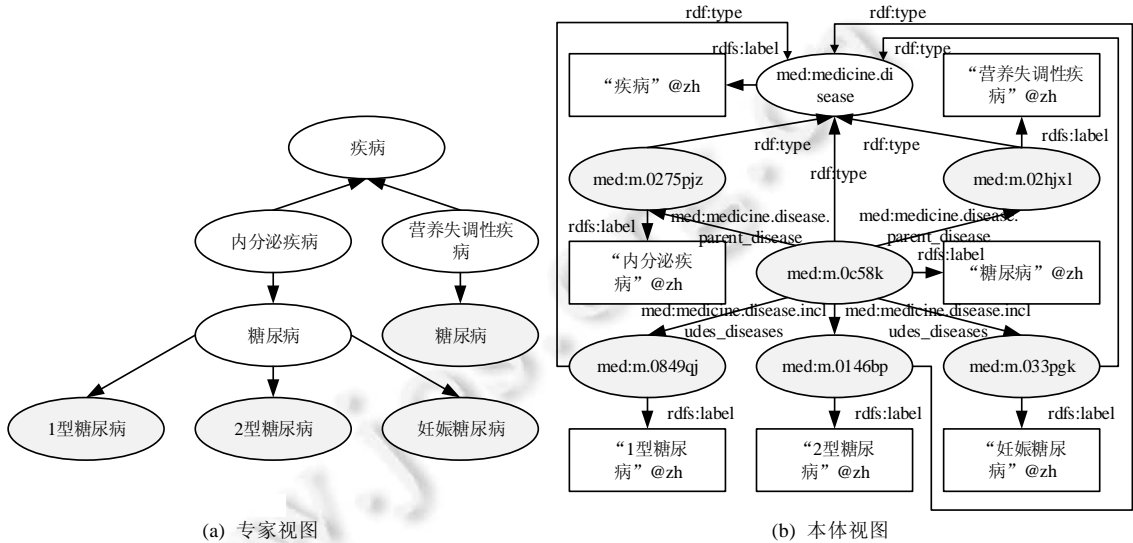


Fig.4 The same resource is both a class and an instance

图 4 同一资源既是类型又是实例

为了平衡表达能力和推理能力,这里选择 OWL DL 作为本体视图的描述语言.同时,为了支撑这里的领域知识描述需求,即某个资源自身既表示实例又可以作为分类代表更多实例对象的集合,这里提出了约定 5 的领域知识建模方法.

约定 5. 在专家视图中,对于自身既表示实例又可以作为分类代表更多实例对象的集合的资源对象,在对应的本体视图中,将其所在分支的父类、子类及实例对象都建模为最近公共父类的实例,并通过对象属性表达实例间的分类关系.

基于约定 5,对于图 4(a)所示的专家视图,可以建立如图 4(b)所示的本体视图.本体视图将专家视图中疾病的所有子类及实例都建模为疾病类的实例,并通过对象属性 `med:medicine.disease.parent_disease` 表达实例的父类(实例),通过对对象属性 `med:medicine.disease.includes_diseases` 表达实例作为类型包含的实例集合.这里,对象属性 `med:medicine.disease.parent_disease` 和 `med:medicine.disease.includes_diseases` 都是传递(transitive)属性,具有类型 `owl:TransitiveProperty`^[3].如果属性 p 为传递属性,那么如果存在陈述 $(A p B)$ 和 $(B p C)$,则意味着存在陈述 $(A p C)$.

除了这里的情况以外,在本体视图中,将一个资源表示为一个概念(即类)还是一个实例,或者将资源建模为实例还是某个数据类型属性的值,这依赖于具体的领域范围和该本体潜在的应用需求.

如果具有超类(superclass)没有的附加属性,或者具有不同于超类的约束,或者参与了与超类不同的语义关联,这时可以建模新的子类而不是超类的实例^[17];如果资源具有很多属性特征,或者与其他资源存在语义关联,那么应该将资源建模为实例而不是作为某个数据类型属性的值.

2.4 多元关系的语义映射

在任何领域的知识描述中,都会涉及到对资源的某些量化指标的描述,例如人的身高、体重、体温、心跳次数等.在不同的国家和地区,对相同指标的默认指称单位并不完全一样,例如在中国,衡量人的身高、体重一般

使用米(m)和公斤(kg),而在英国,计量人的身高、体重的单位分别为英寸(inch)和磅(pounds).一些领域专家在建模领域知识时,会假定领域知识的使用区域或使用者,确信使用此领域知识的使用者会正确推断出默认的计量单位,从而在表示领域知识时常常省略计量单位,如图 5(a)所示.但在将专家视图映射为本体视图并应用到语义 Web 中时,这种假定是不成立的,所以必须对计量单位或类似的信息进行明确表述.

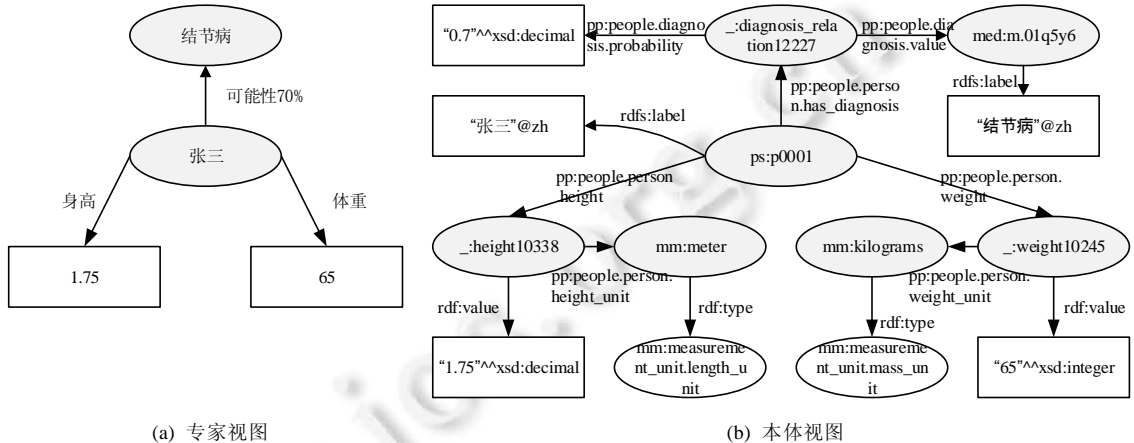


Fig.5 Semantic mapping of structured values and uncertain relations

图 5 结构化值和不确定关系的语义映射

另外,在任何领域的知识描述中都存在一些模糊和不确定的知识,例如医生对患者疾病的诊断就是一种可能性的判断,这种可能性的判断依赖患者主诉的症状、体征、检查、化验、医生经验等.医生在给出诊断结果时,并不能给出这种可能性的大小.健康医疗大数据的出现,一些基于数据挖掘、机器学习等方法构造的针对单病种或者专病的疾病辅助诊断模型,不但可以提供比单个医生(基于大数据的疾病辅助诊断模型综合了更多医生的经验知识)更高的疾病检出率,而且还能提供鉴别诊断,给出数值化的患病可能性参考值.本文不探讨各种模糊性和不确定知识的建模方法,而是给出这种数值化关系的语义描述方法.

在本体视图中,为了描述张三的身高及度量单位,可以在描述身高的属性的 ID 中附上度量单位,比如 pp:people.person.height(m),或者在属性的标注属性(rdfs:label,rdfs:comment 等)中注明使用的计量单位,或者将身高属性的值表示为字符串,在字符串中同时表达身高的数值和单位,例如“1.75m”.对于描述张三与疾病诊断关系的概率值也可以采用类似的方法.这些描述方式对人是可理解的,但却不是机器可处理的;而且对于不同的度量单位和关系权重(例如描述张三患某种疾病的风险等级),需要分别处理,可扩展性差.

描述身高的数值和单位组成了一个结构(structure),要描述张三和这个结构的关系,就要涉及到处理一个 N 元(n-ary)关系^[24].在这里 N 为 3,分别是张三、身高数值和身高单位.要描述张三与诊断关系的概率值,需要一个附加的属性用于描述这个关系的概率,同样构成了一个 N 元关系.而 RDF 只能表示二元(binary)关系,为此作出如下的建模约定:

约定 6. 对于专家视图中明确描述或隐含的结构化值,在本体视图中引入 N 元关系来进行明确表示;使用空节点(blank nodes)^[4,10]或有名资源(具有 URIs)表示 N 元关系:对于每一个 N 元关系,将其切分为一元(作为这个 N 元关系的主体,比如张三)和 N-1 元(比如身高值和计量单位),并创建一个空节点或有名资源及其相关属性来进行连接(即描述).如果 N-1 元中仍然存在多元关系,可采用相同的方法进行切分和描述.

如图 5(b)所示,分别建立了描述人和描述度量单位的本体 O_{people} 和 $O_{measurement_unit}$,这两个本体定义中的资源 ID 对应的名称空间前缀分别为 pp 和 mm.然后定义了一个 $I_{zhangsan}$ (名称空间为 ps),基于 O_{people} , $O_{measurement_unit}$ 和 O_{domain} 来描述个体张三.图中创建了 3 个空节点,分别用于描述张三的身高、体重和诊断结果.这里并没有为空节点分配 URIs,但表达了它应该表达的含义即提供了图中各个部分之间必需的连通作用.这个空节点本身可

能从来不会被从这个 RDF 图的外部引用,因此可能不需要通用的标识符.如果需要被从外部引用,则可以创建有名资源来代替这里的空节点.但在把 RDF 图序列化三元组形式的时候,需要空节点标识符来表示和区分图中的空节点.因此,在三元组表示法中,使用“_:name”形式的空节点标识符来表示空节点^[4],例如图 5(b)中的 `_:height10338`, `_:weight10245` 和 `_:disease12227`.在使用工具程序(例如 Apache Jena^[25])创建空节点时,可以指定空节点标识符,也可以由程序自动生成.

与 URIref 和字面值不同,空节点标识符并不被认为是 RDF 图的一个实际组成部分(因为图中的空节点可以没有空节点标识符).另外,因为空节点标识符表示的是空节点而非弧,所以在一个图的三元组表达式中:空节点标识符只能出现在三元组主体和客体的位置上,而不能出现在谓词的位置上^[4].

结构化值中包含两个或者多个元素,有时领域专家可能想强调某个元素的重要性,例如前面讨论领域专家在描述某些量化指标时通常省略其单位,因为领域专家认为指标的值更重要,单位在某个语境中很容易达成共识.这种重要性更多的是一种主观认识,在领域专家的知识中可能不会特意强调,更多地是一种习惯认识.例如,在描述张三的身高和体重的结构化值中,身高和体重的数值对结构化值的贡献最大,而身高和体重的记录单位用于修饰身高和体重的数值.在其他类型的结构化值中,可能也存在这样的情况.

约定 7. 识别专家视图中明确描述或隐含的结构化值的主值(main value)^[5],在本体视图中使用 RDF 提供的预定义的属性 `rdf:value` 来描述结构化值中的主值.

尽管 `rdf:value` 本身没有含义,但 RDF 鼓励这样使用.如图 5(b)所示,图中属性 `pp:people.diagnosis.value` 为属性 `rdf:value` 的子属性,以表明诊断结果的重要性.

2.5 访问授权的语义映射

针对领域中涉及安全、敏感或用户隐私相关的内容,领域安全专家、信息拥有者(例如医院或患者)或信息来源者(例如患者)都可能提出访问授权的需求.如何基于领域语义知识库建立此访问授权的语义映射,是本节要讨论的主题.

所有领域(尤其是医疗健康、金融等领域)都会涉及到与安全、敏感或用户隐私相关的内容^[19,26].信息共享与隐私保护的目标是共享数据,同时保护个人身份信息,并确保信息的使用与信息收集的目的相一致.隐私保护的核心是隐私控制,即信息的拥有者是否有权完全控制个人信息(全部或部分)的收集、访问和共享,并可以要求停止对个人信息的处理.随着信息化程度的不断提高,越来越多的组织和个人开始担心安全和隐私问题,安全和隐私也成为信息化时代需要首先解决的关键问题^[19].

领域专家在建模领域知识的时候,关心的是尽可能全面、详尽、准确、一致地表达领域中的知识;但在提供知识服务时,领域安全专家必须要考虑领域知识中显式存在的涉及安全、敏感或用户隐私相关内容的管理和控制问题,或者可能由推理、授权传播、数据集成、数据挖掘等过程导致的隐含的安全、敏感或隐私信息泄露问题^[19].领域本体定义提供领域中共同认可的知识,一般不涉及安全和隐私问题;而领域实例数据描述领域中个体(比如人、组织等)的知识,会涉及到与个体相关的安全、敏感或隐私问题;但领域本体定义为个体的描述提供语义支持,并为个体知识的推理提供背景知识.本体描述语言提供的语义描述组件可能为基于推理和数据集成的安全和隐私泄露提供支持.例如,电子邮件地址通常是唯一的,可以唯一标识一个人,所以描述人的电子邮件属性一般定义为 `owl:InverseFunctionalProperty` 类型的,即一个人可以有多个电子邮件,但一个电子邮件地址可以唯一地确定一个人;一本书的 ISBN 号是唯一的,所以书籍的 ISBN 属性通常定义为 `owl:FunctionalProperty` 类型的,即一本书可以唯一的确定一个 ISBN 号.在语义 Web 上,可以通过邮件地址聚合关于某个人的很多信息,通过 ISBN 号可以聚合关于一本书籍及其作者的很多信息,从而可能引发安全和隐私泄露问题.

语义 Web 技术提供的知识表示和推理能力可以增强传统的数据安全和隐私保护能力,并为语义 Web 本身提供保驾护航.本体模型可以很好地建模领域知识,同样可以细粒度地捕捉领域安全和隐私保护需求.基于本体建模的领域知识,为发现显式的和隐含的安全和隐私保护需求提供了很好的支持.传统的安全和隐私保护模型采用增量式的实现方法,没有详细的、一致的形式化模型,不适用于包含大量涉及安全和用户隐私敏感信息的系统,而本体正好提供了一种可扩展的、灵活的、健壮的形式化模型.使用 SWRL 作为安全策略规范,并基于领

域语义知识库中定义的资源 and 陈述,可以构建基于语义的安全和隐私保护约束规则,一方面可以发现由推理引发的安全和隐私泄露问题,另一方面可以检查在策略执行上的一致性,避免授权冲突^[19].

图 6 为一个本体库定义的示例,包含 O_{people} 和 $O_{medicine}$ 的部分片段,其名称空间前缀分别为 pp 和 med .

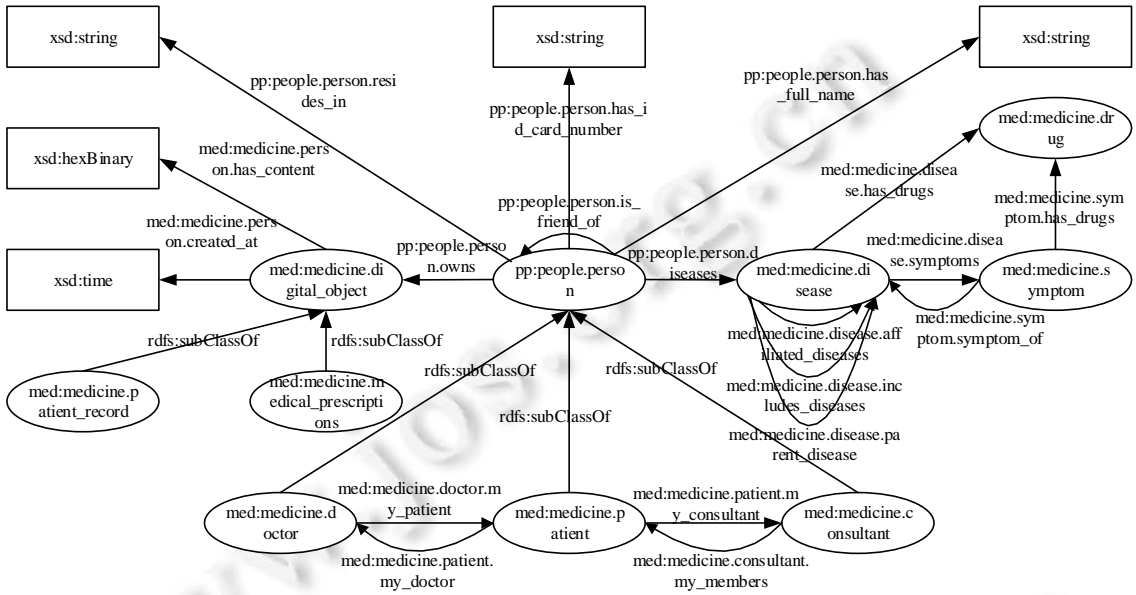


Fig.6 An example of an ontology library definition

图 6 一个本体库定义示例

图中连接概念的属性除了表达分类关系($rdfs:subClassOf$)以外,仅用于声明属性的定义域和值域类型,例如属性 $pp:people.person.diseases$ 的定义域为 $pp:people.person$,值域为 $med:medicine.disease$.本体定义中包含的一级概念有数字对象(电子病历、医疗处方等)、人、疾病、症状;数字对象包含患者记录和医疗处方两个二级概念,人包括医生、患者、顾问这 3 个二级概念;人的主要属性有身份 ID、拥有的数字对象、患有疾病等.

图 7 为基于图 6 的本体定义构建的实例数据示例,分别构建了一个顾问、患者和疾病实例及它们之间的语义关系,描述了患者患有的疾病、拥有的患者记录和身份 ID.

为了实现患者和顾问之间的服务功能,领域安全专家或者患者需要授权顾问(访问主体)对患者相关健康状况(访问客体)的访问权限,例如允许顾问查看患者的诊断结论、电子病历或某次影像诊断结果.从传统的安全视角,访问主体和访问客体都是孤立的,它们通过访问行为建立联系.而在领域语义知识库中,资源之间以及资源和字面值之间都可能存在某种语义关系.与患者相关的信息都和患者存在语义关联,患者和他/她的顾问之间也可能建立了显式的顾问/会员关系.构成领域语义知识库的最小单位是三元组,对语义知识库的访问实际就是对某个或某些三元组的访问,比如张三的诊断、张三的患者记录等.为此,作出如下的约定.

约定 8. 基于 SKB_{domain} 提供语义知识服务时,将三元组作为最小的访问控制保护单元,即访问客体.

领域安全专家或者患者可能为顾问做出如下的授权.

- 1) 允许顾问李四访问患者张三的诊断记录;
- 2) 允许顾问李四访问患者张三的患者记录.

为基于领域语义知识库建立此访问授权的语义映射,作出如下的建模约定.

约定 9. 为基于 SKB_{domain} 的三元组建立访问控制策略规则,采用 RDF 陈述具体化(reification)^[4,10]的方法建模和引用三元组.

一条陈述是由客体、谓词和主体构成的三元组,陈述的 RDF 具体化用于表达关于该三元组的附加信息,例

如三元组的创建时间、创建人等.RDF 具体化词汇包括类型 `rdf:Statement`、属性 `rdf:subject`、`rdf:predicate` 和 `rdf:object`.RDF 具体化词汇被设计用于讨论陈述本身,即 `rdf:Statement` 的实例.图 8 所示建立了两条陈述的具体化,`med:triple0001a` 和 `med:triple0001b` 为陈述实例,通过具体化词汇分别指向了两条陈述.陈述实例的 ID 可以是有名 ID,也可以是无名的空节点.但为了在访问控制策略规则中引用此陈述实例,需要采用有名 ID.

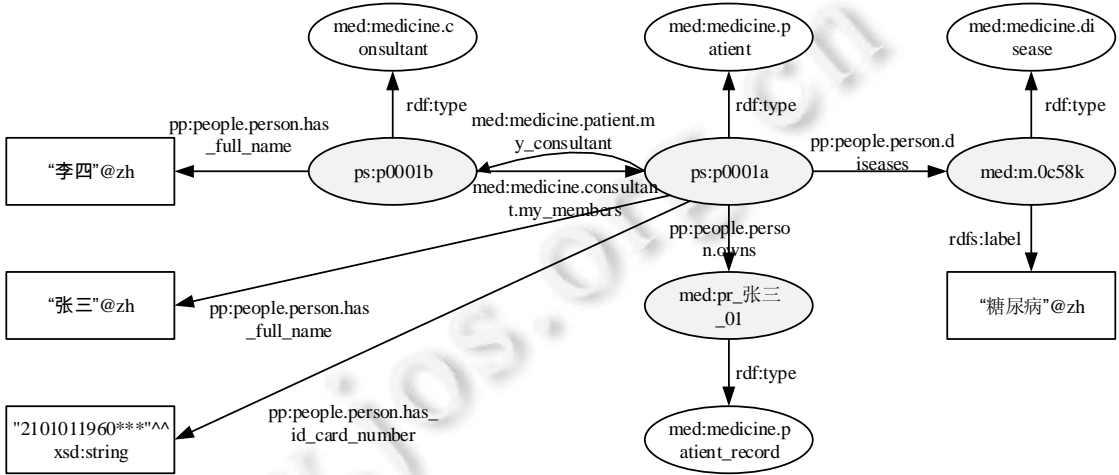


Fig.7 An example of instance data

图 7 一个实例数据示例

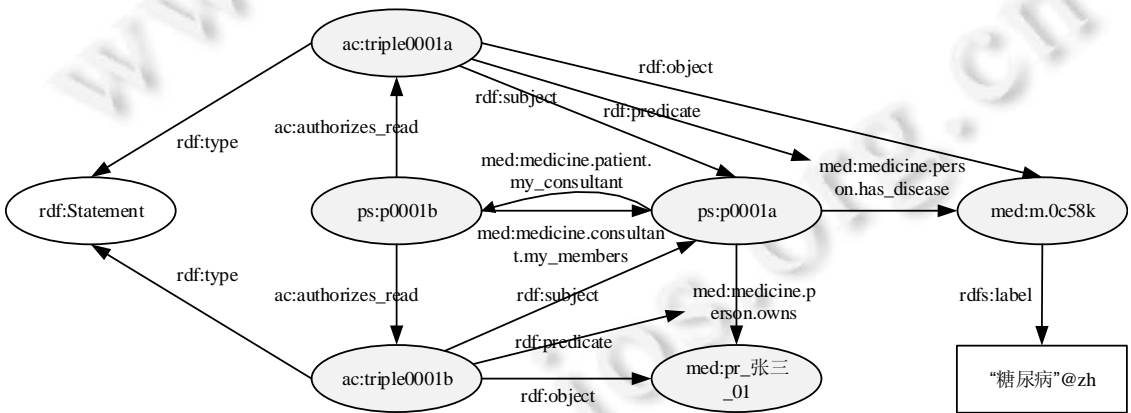


Fig.8 A reification of RDF statement

图 8 RDF 陈述的具体化

为建立授权和许可规则,作出如下的约定.

约定 10. 基于本体模型建立授权和许可访问控制原语,使用 SWRL 作为访问控制策略规范,建立授权和许可规则.

基于本体模型建立如图 9 所示的访问控制原语本体 `Oaccess_control`,即连接人和陈述实例的授权和许可属性.

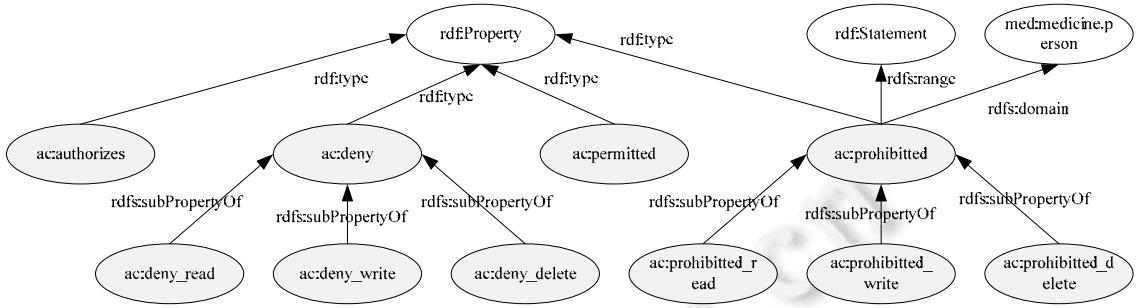


Fig.9 Access control authorization and license primitives

图 9 访问控制授权和许可原语

ac 为访问控制本体的名称空间前缀,授权原语有 ac:authorizes 和 ac:deny,许可原语有 ac:permitted 和 ac:prohibited.建立授权和许可两级访问控制策略模型,分别对应用户级和系统级^[19].在用户级,每个用户可以就个人拥有的资源定制灵活的访问授权策略规则.在系统级的策略规则治理整个系统的安全与隐私策略.由于某些受保护资源可能关联多个授权方用户(比如患者与医生或者顾问之间的关系,如果要访问此关系,需要患者和医生或顾问的同时授权),所以需要通过聚合用户级的授权来判断是否对受保护的资源提供了合适的授权,也即最终的许可需要通过系统级的策略规则来授权.授权和许可同时支持肯定和否定,这样可以简化访问控制策略规则的定义以及解决基于本体知识和规则推理的授权传播问题;如果出现授权冲突,否定授权的优先级高于肯定授权.

3 应用案例

糖尿病、高血压等慢性病患者,除了按医生指导进行日常的用药治疗以外,合理的饮食、适量的运动作为生活方式,对控制血糖、血压等指标同样非常重要.例如,经常有规律的运动不仅可减少 2 型糖尿病的发生率,还可降低血糖,提高胰岛素敏感性,延缓其慢性并发症的发生和发展.但并不是所有慢性病患者都适合做运动,即使适合做运动,还需要考虑适合做多大强度的运动、适合做什么类型的运动、适合什么时间做运动、每次运动的持续时间、运动的注意事项(比如可能需要考虑与药物、饮食的配合,运动环境和装备等)等.运动对糖尿病的防治效果是肯定的,但糖尿病患者运动也有风险,安全有效的运动需要配合调节饮食和用药^[27].著名的 Joslin 纪念章用“胰岛素、运动和饮食三驾马车”来表示成功地治疗糖尿病的 3 个主要方面^[28].

本节采用本文提出的 5 类语义映射方法构建了一个健康医疗领域的语义知识库,用于向慢性病患者推荐个性化的干预方案(饮食处方、运动处方等).当然,这 5 类映射方法并不受限于该特定的领域或应用场景,而是针对几乎在所有领域中都存在的共性问题的解决方案,所以适用于所有类似的应用场景.

3.1 领域本体库的构建

为了支撑向慢性病患者推送个性化的干预方案应用的问题求解需求,这里基于本文提出的语义映射方法约定来将专家知识映射为基于本体和语义 Web 规则语言描述的领域知识.领域专家知识 $EKB_{healthcare}$ 来源于以下几个途径.

- 医疗健康领域的标准术语集,例如 ICD-10,ICD-9-CM,Snomed CT^[29],Mesh^[30],UMLS^[31]等;
- 公开知识库,例如 OpenGallen^[32],Freebase,OBO^[33],OBI^[34]等;
- 公开项目,例如 K4CARE^[35],PIPS^[36]等;
- 论文文献^[36-39]和专业著作^[27,28];
- 网络资料;
- 专家贡献等.

例举的多个标准术语集中,主要都是一些概念分类体系,而没有实例以及实例之间的语义关系.由于不同标

准术语集诞生的背景和应用场景的不同,它们中的一些概念分支存在重叠,但又不完全一样,例如,关于疾病、症状、运动设施等的分类体系有部分重叠.而且部分概念同时出现在多个分支,但表达的语义不一样.为此,由医学专家针对本文的应用场景对其进行了整合.

在列举的几个公开知识库中,除了 Freebase 以外的几个知识库仍然以概念分类体系为主,而 Freebase 正好相反,只有一级概念,以实例关系为主,实例关系主要用于表达领域常识知识,例如某个疾病的主要表现症状、诱因、风险因素等.同时,Freebase 知识库中还包含了度量单位、位置领域的知识.

在例举的几个公开项目中,K4CARE 项目中定义的 CPO(case profile ontology)本体用于描述患者病情概况所需要的知识,比如症状、体征、疾病、综合征、社会问题等及其语义关系,其中,疾病仅仅涵盖了一些慢性病.PIPS 项目中创建了疾病(doid.rdf)、临床记录(PIPSClinicalRecord.owl)、食物(PIPSFood.owl)、菜单(PIPSMenu.owl)、菜谱(PIPSRecipe.owl)、个人概况(PIPSPerson.owl)等本体,但主要是概念分类体系.

列举的论文文献和专业著作主要用于提供运动、饮食相关的专家知识,也从网络上收集了一些资料.

为此,基于本文提出的语义映射方法,由医学专家参与,针对本文的应用场景对几个来源的概念体系进行了整合,实例数据主要参考 Freebase 知识库(并保留了实例数据的 MID)、K4CARE 项目和相关文献.除了 Freebase 知识源以外,其他标准术语集、知识库和公开项目中的知识均为英文描述.Freebase 知识库中的资源对象,其标注属性采用了 20 多种语言分别进行描述.Freebase 知识库中的所有资源对象都具有英文描述,但只有约一半的资源对象具有中文描述.为此,在建立领域语义知识库时,对于参考的英文素材,在保留英文描述的同时,还采用翻译工具对其进行了中文翻译,并由医学专家对其进行了校对.对于医学专家补充的知识,只有中文描述.

领域本体库 $OB_{healthcare}$ 的定义如下:

$$OB_{healthcare} = \{O_{medicine}, O_{people}, O_{food}, O_{exercise}, O_{measurement_unit}, O_{location}\}.$$

各个子本体说明如下.

- 1) $O_{medicine}$ 用于定义医疗健康领域的相关概念(例如疾病、症状、检查、药物、手术、解剖结构、病因、医疗设备等)及其属性(例如疾病概念包含的部分属性有疾病表现的症状、疾病适用的药物、疾病病因、疾病并发症、疾病包含的子疾病、疾病归属的父疾病等);
- 2) O_{people} 用于定义与人相关的概念(例如人、人群、孩子、父母、儿子、女儿、父亲、母亲等)及其属性(例如人的部分属性有性别、出生日期、身高、体重、有身体残疾、有身体损伤、感兴趣/适合/不适合的运动、感兴趣/适合/不适合的饮食等);
- 3) O_{food} 用于定义与食材、食谱相关的概念(例如食材、食谱、中医功效、营养素、坚果种子、奶、水产、水果、油脂、糖品、肉等)及其属性(例如食材包含的营养素、食材功效、食材适宜摄入量、食材推荐摄入量、营养素计量单位、营养素参考值等);
- 4) $O_{exercise}$ 用于定义康复运动相关的概念(例如运动项目、运动设备、运动强度、能量代谢类型、运动益处等)及属性(例如运动项目需要的运动设备、运动项目包含的子运动项目、运动项目归属的父运动项目、运动项目的能量代谢类型、运动项目的运动强度、运动项目的运动益处、运动项目的锻炼部位、运动项目依赖的运动场地等);
- 5) $O_{measurement_unit}$ 包含度量相关的概念(例如长度单位、质量单位、时间单位、货币单位等)及其属性(例如长度计量单位、长度计量单位的标准缩写等);
- 6) $O_{location}$ 包含地理位置、场所相关的概念(例如地址、邮编、洲、国家、省、市、县、镇、场所、部门等)及其属性(例如编码、名称、包含关系、隶属关系等).

领域本体库 $OB_{healthcare}$ 中的主要概念及其属性关系如图 10 所示,其中,椭圆表示概念,带箭头的连线表示属性,箭头源端连接的概念表示该属性的定义域,箭头目的端连接的概念表示该属性的值域.

在定义各个子本体时遵循的约定见表 1.

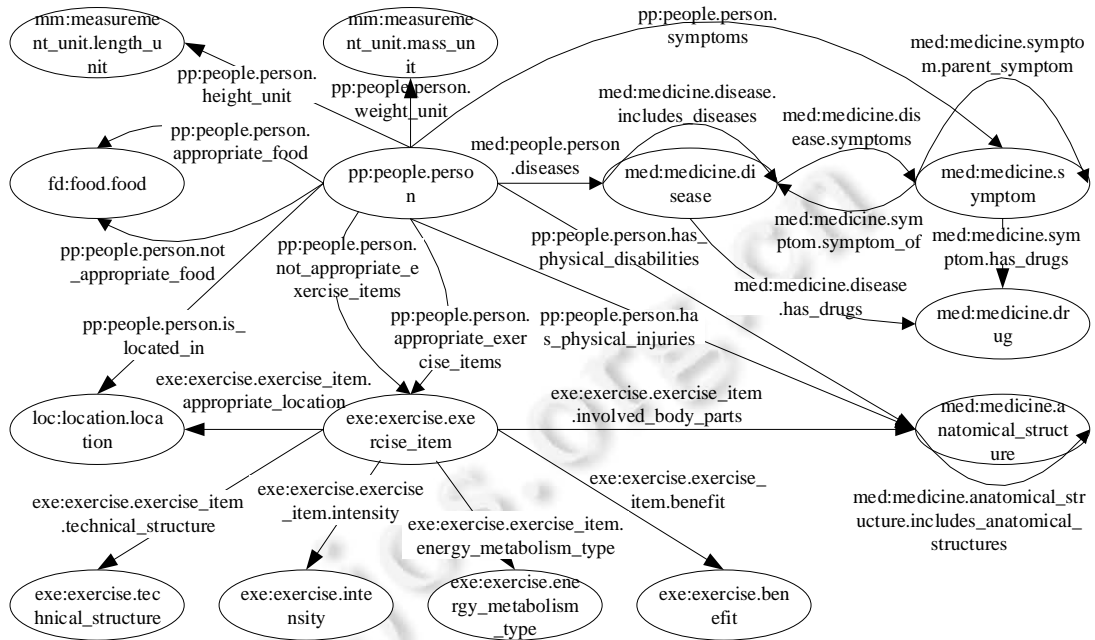


Fig. 10 Main concepts and their relationships

图 10 主要概念及属性关系

Table 1 Conventions followed by the domain ontology library definitions

表 1 领域本体库定义遵循的约定

遵循约定	说明
约定 1	<p>各个子本体中概念和属性的定义都采用了人类可识别的 ID,并遵循约定 1.一方面,从资源名称上便可以识别其为类型还是属性;另一方面,也避免了人类理解上的语义冲突,例如,med:medicine.disease 表示疾病概念,而以下都表示属性:</p> <ul style="list-style-type: none"> med:medicine.disease.parent_disease med:medicine.disease.includes_diseases med:medicine.disease.affiliated_diseases med:medicine.risk_factor.diseases med:medicine.medical_trial.diseases med:medicine.diagnostic_test_signs.disease med:medicine.disease_cause.diseases <p><i>O_{medicine}</i> 本体中来自 Freebase 的实例数据保留了其 mid,来自 ICD-9/10 的实例数据采用在 ICD-9/10 中的编码.来自其他源的实例数据,在与现有 ID 不冲突的情况下,尽可能保留了其原有的编码.手工构建的实例对象,目前采用了人类可识别的 ID.</p>
约定 2	<p>各个子本体中概念的属性定义位置遵循约定 2,一方面确保了领域本体库定义的完整性,另一方面避免了各个子本体属性定义的冗余.同时,针对每个属性都给出了明确的定义域和价值域限定,一方面用于准确表达属性的语义,另一方面避免了语义推理的冲突.</p>

对于表达一般性领域常识知识的实例数据,这里直接定义在对应的领域本体中了,例如疾病实例、症状实例、药物实例、手术实例、解剖结构实例等及这些实例间关系的定义都直接定义在本体 *O_{medicine}* 中了;不同类别的食材实例及其属性的定义(例如包含的营养素、功效等)直接定义在本体 *O_{food}* 中了;不同类别的运动实例及其属性的定义(例如运动项目的强度、能量代谢类型、益处、技术结构等)直接定义在本体 *O_{exercise}* 中了;度量单位实例及其属性的定义直接定义在本体 *O_{measurement_unit}* 中了;位置相关的实例及其属性的定义直接定义在本体 *O_{location}* 中了.在定义领域常识知识时遵循的约定见表 2.

Table 2 Conventions followed by the domain commonsense knowledge definitions**表 2** 领域常识知识定义遵循的约定

遵循约定	说明
约定 3	<i>O_{medicine}</i> 本体中部分实例资源具有多种类型,例如糖尿病、泌尿道感染等实例既是一种疾病,也是其他疾病的风险因素、病因或表现症状.对于这些实例的定义遵循约定 3,参见图 2 所示.
约定 4	<i>O_{exercise}</i> 本体中所有运动项目都可以按照能量代谢类型、运动的益处、运动的技术结构和运动的强度进行分类,在进行语义映射时遵循约定 4,针对运动项目实例,只建立了一个运动项目一级分类,并将所有运动项目建模为其实例;而将专家视图中的不同特化分类建模为类型实例,同时增加多个对应的分类属性来连接运动项目实例和这些类型实例,参见图 3 所示.
约定 5	在定义疾病、解剖结构、症状、运动项目、食材等概念和实例时,部分资源既是类型又是实例,这时遵循约定 5 的方法建立语义映射,采用 OWL DL 来描述本体视图,参见图 4(b)和图 10,以确保推理计算的完全性和可判定性.

以 Turtle RDF^[40]表示的领域语义知识库的文件(扩展名为.ttl)大小约为 1.8GB,知识库中包含 144 个概念、82 个数据类型属性、166 个对象属性、888 308 个实例和 7 076 316 个三元组.

该知识库的规模和内容可以满足临床科研、临床诊疗、健康管理中一些基本的知识服务需求,例如提供数据的共享与互操作服务.在健康医疗领域中,同一种疾病、同一种症状等在不同医疗机构可能有不同的描述方法,甚至在同一医疗机构也存在许多不同的描述方法.这些不同描述方法与标准描述方法的归一化是一项知识密集型、时间密集型的工作.但此项工作不解决,领域知识库的作用便不能得到有效地发挥.为向领域应用提供智能知识服务,还需要结合特定领域的问题求解需求,补充问题求解知识,知识的形态可以是语义 Web 规则集(参见第 3.3 节)、算法、模型等.

3.2 领域实例库的构建

为了对知识库在问题求解上的可满足性进行验证,这里以推荐个性化运动处方场景为例,分别定义了 3 个个体的实例数据,即个体张三、李四、王二的实例数据,包括他们的基本信息(性别、年龄、身高、体重)、健康状况(患病情况、残疾情况、损伤情况)、兴趣爱好(运动、饮食等)、日常体力活动水平、生活习惯、具有的运动条件(场地、设备、经济条件等).领域实例库 $IB_{healthcare}$ 定义如下:

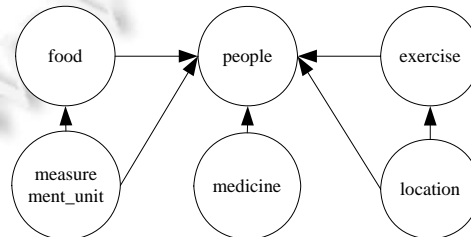
$$IB_{healthcare} = \{I_{zhangsan}, I_{lisi}, I_{wanger}\}.$$

在定义个体实例数据时遵循的约定见表 3.

Table 3 Conventions followed by the individual instance data definitions**表 3** 个体实例数据定义遵循的约定

遵循约定	说明
约定 6 和约定 7	在基于领域本体库 $OB_{healthcare}$ 描述张三、李四和王二的年龄、身高、体重等结构化值时,遵循约定 6 和约定 7,参见图 5 所示.这里,个体的患病情况是一个确定值,作为干预方案推荐时考虑的一个条件,直接使用属性 pp:people.person.diseases 来连接个体及其所患疾病.结构化值中的年龄值、身高值、体重值等作为主值,度量单位作为附加信息.

在使用 Protégé 定义对象属性的值域、实例间的语义关系和个体实例数据时,需要引用或导入(owl:imports)关联本体中的概念和实例,本体间的引用关系如图 11 所示.

**Fig. 11** Reference relations between ontologies**图 11** 本体间的引用关系

在基于专家知识构建的运动推荐模型中,个体的这些实例数据的作用见表 4.

Table 4 Roles of individual instance data in the exercise recommendation model

表 4 运动推荐模型中个体实例数据的作用

个体实例数据	作用
性别、年龄	确定可以参加的运动项目和运动强度
身高、体重	得到身体质量指数(BMI),从而获知体型.体型将决定运动的强度和持续时间
患病情况	是否可以做运动;是否需要做运动耐量实验
残疾情况、损伤情况	判断是否可以参与某些运动项目
兴趣爱好	基于此推荐最佳运动项目
日常体力活动水平	确定运动强度
生活习惯	确定运动项目和强度
具有的运动条件(如场地、设备、经济条件等)	判断是否可以参与某些运动项目

3.3 领域语义规则集的定义

任何基于知识的系统,其知识库至少由两个基础部分组成:静态领域知识和动态推理知识^[1].静态领域知识描述领域明确的静态知识;动态推理知识又称为问题求解知识,问题求解知识从抽象层描述问题求解方法所需要的知识,也就是关于如何满足需求的知识,它描述了要实现的目标、实现这些目标必要的行为、这些行为的激活顺序以及执行这些行为所需要的领域知识.领域知识会影响问题求解知识,而问题求解方法可以用于指导获取静态领域知识.为此,本体视图仅仅在语义上准确反映专家视图还不够,还需要从满足领域应用问题求解的角度,调整本体视图中包含的对象类型、数量及其组织方式的合理性,或者补充问题求解所需要的知识.

基于表 4 可以知道:为了问题求解的需要,即为了描述推荐模型(运动推荐模型、饮食推荐模型等),还需要基于 SWRL 构建动态推理规则集 $F_{healthcare}$.例如在运动推荐模型中,基于个体的实例数据和推理规则,判断患者是否适合做运动;如果适合做运动,是否需要做运动耐量实验,适合多大强度的运动;基于身体残疾、损伤、日常体力活动水平、兴趣爱好、生活习惯、运动条件等因素判断适合个体的最佳运动项目.

SWRL 规则有 3 种不同的语法^[3,20]:抽象语法(abstract syntax)、XML 具体语法(XML concrete syntax)以及 RDF 具体语法,其中,XML 具体语法基于 OWL XML 表示语法和 RuleML XML 语法.但采用这几种 SWRL 语法对规则进行序列化描述时,语句非常冗长,人类不容易阅读.于是,在规则表示中,经常使用一种相对非正式的“人类可读的”语法形式^[20]:

$$antecedent \rightarrow consequent.$$

一个 SWRL 规则包含了一个前提(antecedent)部分和一个结论(consequent)部分,前提又称为规则体(body),结论部分又称为规则头(head).规则体和规则头可以只有一个原子公式(atomic formula),或者是多个原子公式的合取(conjunction),其表示形式如下:

$$atom \wedge atom \wedge \dots \wedge atom \rightarrow atom \wedge atom \wedge \dots \wedge atom.$$

一个 SWRL 规则可以这样理解:如果在前提中的所有原子是真实的,那么结果也必须是真实的.一个原子的表示形式为

$$p(arg_1, arg_2, \dots, arg_n).$$

p 是谓词符号, $arg_1, arg_2, \dots, arg_n$ 是术语或表达式的参数.在 SWRL 中,谓词符号包括 OWL 类、属性、数据类型或内置原语(builtin primitives)等.参数可以是 OWL 实例、数据值或者引用它们的变量.

另外,在本应用场景中,将个体的年龄、患病情况、残疾情况、损伤情况、经济条件等识别为个体的敏感和隐私数据,为实现对这些数据的访问控制,首先需要建立用于授权和许可规则的访问控制原语本体 $O_{access_controls}$,参见图 9.然后将识别出的敏感和隐私保护内容映射为领域实例库中的三元组,如图 7 所示.然后在访问控制本体中将这受保护的三元组具体化为陈述对象,如图 8 所示.最后,基于访问控制需求、陈述对象和访问控制原语建立相应的访问控制策略规则 $F_{access_controls}$.这个过程中遵循的约定见表 5.

Table 5 Conventions followed by the definitions of the data security and privacy protection model**表 5** 数据安全和隐私保护模型定义遵循的约定

遵循约定	说明
约定 8	在本应用场景中,将个体的年龄、患病情况、残疾情况、损伤情况、经济条件等识别为个体的敏感和隐私数据,为实现对这些数据的访问控制,将个体与这些数据间的关系实例(即三元组)识别为被保护对象,即访问客体.
约定 9	为在访问控制策略规则中引用这些三元组,采用 RDF 具体化的方法将这些三元组具体化为陈述对象.
约定 10	基于本体模型建立授权和许可访问控制原语,使用 SWRL 作为访问控制策略规范,基于领域安全专家、信息拥有者或信息来源者的访问控制需求建立授权和许可规则.

最终形成的领域语义知识库 $SKB_{healthcare}$ 定义如下:

$$SKB_{healthcare}=(OB_{healthcare}, IB_{healthcare}, F_{healthcare}, O_{access_control}, F_{access_control}).$$

3.4 案例实现

Protégé 和 Jena 都提供了对本体标准描述语言很好的支持,应用中的本体视图采用 Protégé 4.3 进行构建,将每个本体和实例数据分别存储到不同的文件中,并以 Turtle 语法格式进行表示,文件扩展名为 .ttl.

原型系统基于 Jena 2.12.0 语义 Web 编程框架实现,使用 Jena 提供的命令行工具 tdbloader 将领域本体库和实例库加载到名为 healthcare 的 TDB^[41]存储中,TDB 版本为 TDB 1.1.0.

Jena TDB 是一个高性能的 RDF 专用存储.

使用 Jena 通用规则引擎^[42]提供的规则语法格式来描述领域规则集,并存储到扩展名为 .rules 的规则文件中,运动处方推荐模型包含 51 条规则,部分规则如下.

```
@prefix medicine:<http://www.neusoft.com/ontologies/2017/6/medicine.owl#>.
```

```
@prefix location:<http://www.neusoft.com/ontologies/2017/6/location.owl#>.
```

```
@prefix measurement_unit:<http://www.neusoft.com/ontologies/2017/6/measurement_unit.owl#>.
```

```
@prefix exercise:<http://www.neusoft.com/ontologies/2017/8/exercise.owl#>.
```

```
@prefix food:<http://www.neusoft.com/ontologies/2017/12/food.owl#>.
```

```
@prefix people:<http://www.neusoft.com/ontologies/2017/12/people.owl#>.
```

```
@prefix xsd:<http://www.w3.org/2001/XMLSchema#>.
```

```
@include <RDFS>
```

```
...
```

```
#1 型糖尿病患者,如果存在高血糖和酮症,应当避免运动.
```

```
[r9:(?person people:people.person.has_diseases medicine:m.0849qj),
```

```
(?person people:people.person.has_symptoms medicine:高血糖),
```

```
(?person people:people.person.has_symptoms medicine:酮症)→
```

```
(?person people:people.person.can_do_exercise "false"^^xsd:boolean)]
```

```
#空腹血糖在 13.9mmol/L(250mg/dl)以上且尿酮体阳性,或尿酮体虽阴性但空腹血糖在 16.7mmol/L  
#(300mg/dl)以上的病人,运动不仅不能促进肌肉对糖的利用,而且能促进肝脏对葡萄糖的输出,使脂肪分解  
#增加,易诱发酮症酸中毒.
```

```
[r10:(?person people:people.person.has_diseases medicine:m.0849qj),
```

```
(?person people:空腹血糖?value),
```

```
ge(?value,13.9),
```

```
(?person people:尿酮体 medicine:阳性)→
```

```
(?person people:people.person.can_do_exercise "false"^^xsd:boolean)]
```

```
[r11:(?person people:people.person.has_diseases medicine:m.0849qj),
```

```
(?person people:空腹血糖?value),
```

```
ge(?value,16.7),
```

```

(?person people:尿酮体 medicine:阴性)→
(?person people:people.person.can_do_exercise "false"^^xsd:boolean)]
...
#如果患者大于 35 岁、2 型糖尿病(T2DM)持续 10 年以上、1 型糖尿病(T1DM)持续 15 年以上、存在发生
#冠状动脉疾病的危险、有微血管并发症(增生性视网膜病或肾病,包括微量蛋白尿)、有外周血管疾病和自
#主神经病变的,符合上述任何一条者都需要做运动耐量实验.
#患者大于 35 岁
[r15:(?person people:people.person.has_diseases ?disease),
  (?disease medicine:disease.parent_disease medicine:m.0c58k),
  (?person people:people.person.has_age ?age),
  greaterThan(?age,"35"^^xsd:integer)→
(?person people:people.person.need_do_exercise_tolerance_test "true"^^xsd:boolean)]
#T2DM 持续 10 年以上
[r16:(?person people:people.person.has_diseases medicine:m.0146bp),
  (?person people:people.person.dm_time ?years),
  greaterThan(?years,"10"^^xsd:integer)→
(?person people:people.person.need_do_exercise_tolerance_test "true"^^xsd:boolean)]
#T1DM 持续 15 年以上
[r17:(?person people:people.person.has_diseases medicine:m.0849qj),
  (?person people:people.person.dm_time ?years),
  greaterThan(?years,"15"^^xsd:integer)→
(?person people:people.person.need_do_exercise_tolerance_test "true"^^xsd:boolean)]
...
#糖尿病患者,年龄大于 50 岁,适合低强度或轻微强度的运动
[r25:(?person people:people.person.has_diseases ?disease),
  (?disease medicine:disease.parent_disease medicine:m.0c58k),
  (?person people:people.person.has_age ?age),
  greaterThan(?age,"50"^^xsd:integer)→
  (?person people:people.person.appropriate_exercise_intensity exercise:low_intensity),
  (?person people:people.person.appropriate_exercise_intensity exercise:light_intensity)]
...
#因为身体部位有损伤或残疾而不适合的运动项目
[r42:(?person people:people.person.has_physical_disabilities ?body_part),
  (?person rdf:type people:person),
  (?body_part rdf:type medicine:anatomical_structure),
  (?exercise_items exercise:exercise.items.involved_body_parts ?body_part)→
(?person people:people.person.not_appropriate_exercise_items ?exercise_items)]
...

```

- 规则文件中,“@include <RDFS>”表示导入规则推理所需要的 RDFS 规则集;
- 每条规则中,“:”前为规则名,可选;“→”表示规则的方向,这里为前向规则(forward rule),方向的左边表示规则体,即推理条件,右边表示规则头,即推理结论;
- 规则体和规则头中的多个项(term)之间使用逗号分隔,每个项可以是一个三元组模式或者 Jena 规则引

擎的内置原语,这里用斜粗体表示内置原语,用于提供一些通用的计算功能;

- 三元组模式中包含变量(*?person,?value* 等)、本体描述语言语义组件、*SKB_{healthcare}* 中定义的语义组件.

Jena 通用规则引擎可用于实现 RDFS,OWL 推理以及通用目的的推理,提供了正向链接推理、反向链接推理和混合执行模式.本体推理、规则推理和应用的执行逻辑如图 12 所示.本体推理基于领域本体库中的语义组件(即资源)进行推理,例如基于类和属性的分类关系、类和属性的约束(例如复杂类构造中的实例值约束、存在限定、完全限定;属性的取值约束、基数约束、定义域和值域约束)^[3,6,10]、属性的特征(传递性、对称性、非对称性、自反性、非自反性、函数属性、逆函数属性)^[3,6,10]等;规则推理基于语义知识库和规则体中的条件模式进行推理,如果满足规则体的条件,那么得出规则头的陈述结论,作为蕴含的知识和领域本体库、领域实例库中的显式知识一起作为可查询的领域语义知识库.

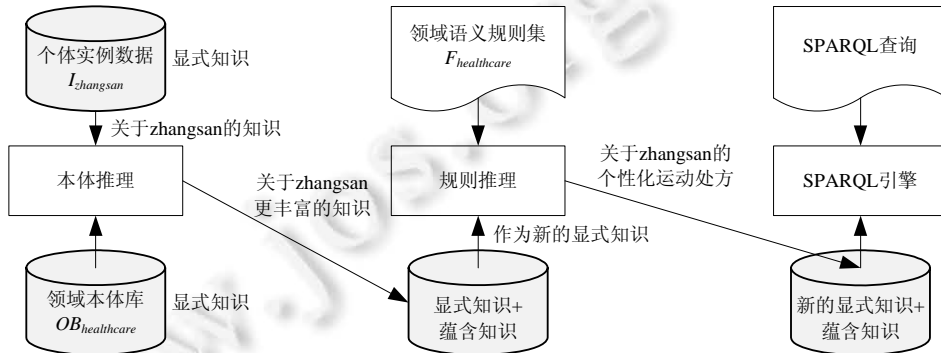


Fig.12 Ontology inference, rule inference and query logic

图 12 本体推理、规则推理和查询逻辑

基于领域本体库和领域实例库中定义的显式知识和基于本体推理与规则推理得到的蕴含知识,执行 SPARQL 查询得到的张三的个性化运动处方推荐示例如图 13 所示.首先基于个体的实例数据和运动评估模型,判断个体是否适合做运动;如果适合做运动,进一步判断是否需要做运动耐量实验;然后,基于个体实例数据、评估结果和领域语义知识库推荐个性化的运动处方,包括适合的运动项目、适合的运动强度、运动频率、适合的运动进行时间、适合的运动持续时间和运动注意事项.

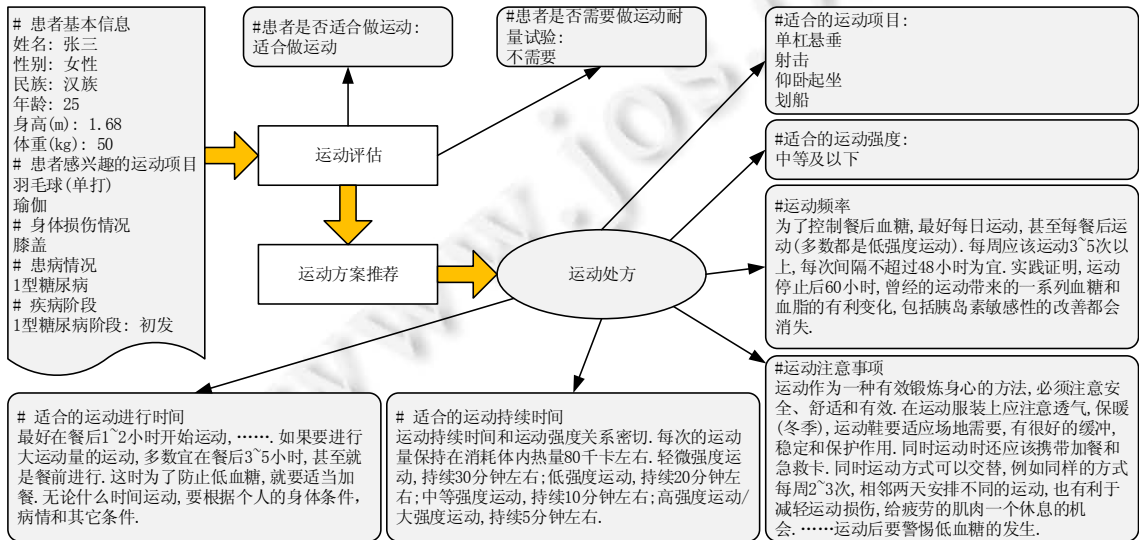


Fig.13 Exercise evaluation and personalized exercise prescription recommendation

图 13 运动评估和个性化运动处方推荐

4 方法评价

目前,相关研究工作还没有深入讨论本文提出的语义映射问题,也没有给出相应的语义映射方法.本文的研究动机和内容完全是基于实践中遇到的一部分问题,这些问题在相同领域的其他实践中以及其他领域中也都会遇到,属于共性问题.这些问题如果不解决,本体工程也得不到很好的实施和落地.

第3.1节的应用案例覆盖了本文提出的5类语义映射方法,当然,针对其他应用案例可能还会遇到其他需要解决的语义映射问题.基于本节提出的领域应用需求和本文提出的语义映射方法建立了领域语义知识库.基于领域应用的问题求解需求补充了领域语义规则集.通过向规则引擎传递个体实例数据(已知条件,也是显式知识)、领域语义知识库(显式知识)和领域语义规则集,获得推理结果(隐式知识).最后,基于 SPARQL 查询以获得个体的运动、饮食等干预方案.案例实现结果可以成功获取个体的运动处方.从推荐结果上看,干预计划完全符合个体的身体状况和干预需求.当然,推荐结果的好坏不仅仅是一个方法的问题,还涉及到领域语义知识库的规模和质量,尤其是面对具体应用场景的问题求解知识的科学性和正确性.基于知识的智能决策支持系统依赖领域专家和知识工程师的密切配合,以及适当的方法、手段和工具的支撑.

5 相关工作

目前,相关工作主要集中在以下几个方面.

1) 为知识工程师或者领域专家提供方便的本体编辑工具.

近年来,随着本体的应用范围越来越广泛,已经出现了很多本体构建工具,例如 Protégé, OntoEdit, KAON 等,本体构建工具为用户提供了友好的图形化操作界面,并通过集成语义推理引擎提供了本体构建过程中的语义一致性检查机制.借助这些工具,用户不必了解本体描述语言的细节,就可以方便地构建本体.而且,通过本体构建工具集成的语法和语义一致性检查机制,避免了很多错误的发生^[43].

以上这些工具仅仅提供了本体的辅助编辑功能,可以方便本文讨论的语义映射结果的创建和输入,但并未提供具体的本体建模方法.或者说,在使用这些本体编辑工具之前,需要事先在纸面上或者头脑中形成需要的本体模型,然后借助这些工具进行方便地形式化.

2) 从本体工程或本体论的角度提出本体建模的一些方法学.

文献[44]介绍了用于创建企业本体的 IDEF5 方法、骨架法以及用于创建领域本体的七步法以及 TOVE, Methontology 等通用本体创建方法.这些本体建模方法只提供本体建模的基本步骤和指导方针.

本体与面向对象对客观世界的认识方法有许多相似之处,可以借助面向对象的建模方法来建模本体.文献[11]从 UML 的静态模型和动态模型两个方面研究了本体的面向对象建模方法.文献[12]通过 UML 的类图来描述环境领域的本体模型,并提出了一种从 UML 类图各元素到本体 OWL 描述各元素的转换方法,详细阐述了从本体模型到 OWL 描述逻辑的转换过程和转换规则.

面向对象的本体建模方法主要是通过类图描述本体中的概念、属性和关系,但是由于领域知识体系的复杂性,类图这种简单的结构难以建模概念属性繁多、关系错综复杂的本体.因此,研究者们利用层次结构中层次结构清晰简洁的特点,通过层次建模方法对复杂本体进行建模.文献[13]提出了领域本体建模的4层模型,通过领域层、分类层、类层和实例层这4个层次构建领域本体.其中:领域层表示领域本体的领域名称,由领域专家定义的领域分类组成;分类层包括该领域的具体分类项;类层由各应用领域的类组成,每个类包括类名、属性集合和操作集合;实例层由各应用领域的实例组成,每个实例包括实例所属类的名称、属性集合和操作集合.文献[14]中提出了领域本体自顶向下建模的5层模型,用于自适应 Web 系统的本体建模.5层框架中包括数据层、概念层、用户层、自适应层和表示层,其中:数据层包括所有松散的 Web 数据单元;概念层包括由数据层的数据单元抽象而来的概念与概念之间的关系;用户层包括用户访问 Web 的方式及行为的信息;自适应层包括 Web 中的自适应规则和逻辑规则;表示层包括 Web 系统内容和结构的信息.

除了以上较通用的一些方法学以外,还有一些针对具体问题域或领域的领域语义知识库构建的方法学研

究.文献[45]从本体论的角度出发,探讨了几何学知识的获取及表示方法,然而仅简单描述了几何学本体的结构,列举了部分属性和关系,识别了一些简单的公理,并未清晰完整地给出几何学知识中的每类对象的本体描述方法,也未针对几何学本体知识在用于问题求解时可能需要补充的语义规则集进行讨论.文献[46]提出了基于 OWL 本体与 Prolog 规则的平面几何知识库的构建方法,用 Protégé 和 Prolog 构建了一个基于本体和规则的平面几何知识库,但还都是一些简单对象的语义映射.文献[47]设计了一个本体制导的基于问题框架方法的需求建模过程,为需求分析员提供建模指导并规范其建模活动.

文献[15,16]对几种面向具体学科领域或工程问题的本体建模方法进行了比较.以上本体建模方法或者只是提供基本的步骤和简单的指导原则,或者基于面向对象和层次化建模的思想提供了识别领域中类、类属性及类之间关系的方法,或者讨论了本体在特定领域中的一些应用,但仅仅构建了一些简单的类和实例对象.均未提及本文讨论的深层次语义映射问题.

3) 如何从领域数据源中半自动或自动地构建垂直领域语义知识库或知识图谱.

目前公认领域本体的开发需要领域专家的参与,但由于领域知识体系的复杂性,完全由人工构建几乎是不可能的,并且在时间上也是不可接受的.因此,如何利用知识获取技术来降低本体构建的开销,已成为一个非常活跃的研究方向,相关技术被称为本体学习(ontology learning)技术^[43,48-50].本体学习,又称为本体获取(ontology acquisition),即采用自然语言处理、统计分析、机器学习等技术自动或半自动地从领域数据中获取领域知识,并基于本体进行描述.领域数据的类型多种多样,不同类型的数据可能需要采用不同的本体学习方法,可以按领域数据的结构化程度将现有的本体学习技术大致分为基于结构化数据的本体学习技术、基于半结构化数据的本体学习技术和基于非结构化数据的本体学习技术.文献[43]对现有的本体学习技术和学习工具进行了调查.由于实现完全自动的知识获取技术还不现实,整个本体学习过程还是在用户指导下进行的一个半自动的过程.本体学习中的很多技术都依赖于对自然语言的处理,所以本体学习工具具有很强的语言特征,目前还没有一个能够很好地支持中文的本体学习工具.虽然目前已经提出了很多本体学习方法,但大部分方法都不理想.目前的本体学习工具的功能都非常有限,它们都仅能处理某些类型的数据源,获取某些本体学习对象,例如,将关系数据库中的表映射为本体模型中的概念,将表字段映射为概念的属性,将字段值映射为属性值等.文献[3]也探讨了如何将基于 XML 的 Web 服务、关系数据库以及其他类型数据源映射为 RDF 模型,但也是提取一些简单的对象和关系.

近年来,对知识图谱技术^[51,52]的研究越来越受到关注.知识图谱是结构化的语义知识库,其描述方法同基于本体的知识表示模型.知识图谱可以用于支撑语义搜索、智能问答、知识推理等智能化、精准化的应用.各个垂直领域存在大量结构化、半结构化和非结构化的数据,如何发挥这些数据的价值,一种最可行的解决方法就是建立面向垂直领域的知识图谱,用于支撑垂直领域的语义搜索、数据集成、数据分析等智能化的应用.华东理工大学的阮彤教授就垂直领域的知识图谱,面向图书馆、证券、医疗等行业做了部分探索^[53-56],其研究工作聚焦在如何迭代式构建领域本体(即领域本体 Schema 定义)、如何从行业拥有的多种中文数据源(例如关系数据库、文本、网页等)中自动或半自动地抽取出结构化的领域知识(即领域实例数据定义)、如何实现异源数据的融合(即本体映射)等.

本体学习中的领域数据源往往并不是专家整理后的知识,而是一些面向业务过程的数据或者面向领域教育的素材.所以,现有的本体学习技术和工具还没有深入探讨本文提出的语义映射问题.

4) 通过本体集成和本体映射对现有的本体进行复用.

由于本体具有的强大的知识表示和推理能力,已经出现了很多基于本体构建的领域语义知识库.现有的本体构建方法^[15-17]都强调在基于本体构建领域语义知识库之前,考虑集成和复用已经存在的领域本体库.一方面,因为构建领域本体库的目的本身是为了领域知识的共享、集成和复用,通过集成和复用已有的领域本体库,既体现了这个目的的价值,也有利用所构建的领域知识库的共享、集成和复用;另一方面,通过集成和复用已有的领域本体库,在此基础上进行修订和扩展,可以帮助快速构建满足领域应用问题求解需求的新的领域本体库,例如通过本体集成(ontology integration)^[57]和本体映射(ontology mapping)^[58]的方法来快速构建所需要的领域本

体知识库。本体集成是指在建立一个新本体时重用其他现有的本体,大多数本体的创建都重用了已有的其他本体;本体集成通过本体扩展(ontology extension)构建新的本体。

本体映射又称为本体对准(ontology alignment)^[3,59],本体映射是指通过对两个本体进行语义的联系,实现将源本体的实体(即资源)映射到目标本体实体上的过程;本体映射通过本体比较(ontology comparison)将两个或多个本体归并或合并为一个本体。本体集成和本体映射企图复用已有的本体快速构建新的领域本体,以满足领域应用的知识需求。本体集成过程会基于领域知识对复用的本体内容进行调整、修正和扩展,本体映射过程也会参考领域知识实现资源间语义的对准和修正,但现有的研究更多的强调对复用本体的处理,并未涉及本文的研究工作。

6 总结和进一步的工作

本文从具体实践中识别了5类语义映射的共性问题,对其进行了讨论,提出了对应的解决方案。然后,通过一个完整的应用案例对这5类语义映射方法进行了验证,证明了其可用性。这5类语义映射问题在所有领域中都存在,这里提出的语义映射方法也适用于各种类似的应用场景。语义映射的结果最终会用于领域问题的求解,所以还需要根据具体应用场景补充问题求解知识。问题求解结果的好坏取决于很多因素,例如语义映射方法的正确性、领域语义知识库的规模和质量、面对具体应用场景的问题求解知识的科学性和正确性等。

除了本文讨论的几个语义映射的通用问题以外,在不同领域的各种应用场景中还存在很多其他需要解决的语义映射问题或建模方法问题,例如各个领域中都存在大量模糊的和不确定性的知识,如何对这些知识进行准确的语义建模和语义推理,是一个值得研究的共性问题。对于此类问题,现有的研究主要分为两类:从逻辑层面通过引入模糊逻辑对描述逻辑进行扩展,形成模糊描述逻辑;从本体描述语言层面对领域本体进行模糊扩展,形成模糊本体。但这两类研究都存在很多不足,需要深入探讨。

References:

- [1] Studer R, Benjamins VR, Fensel D. Knowledge engineering: Principles and methods. *Data and Knowledge Engineering*, 1998, 25(1/2):161–197. [doi: 10.1016/S0169-023X(97)00056-6]
- [2] Berners-Lee T, Hendler J, Lassila O. The semantic Web: A new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*, 2001,284(5):34–43.
- [3] Hebel J, Fisher M, Blace R, Perez-Lopez A. *Semantic Web Programming*. Indianapolis: Wiley Publishing, 2009.
- [4] Manola F, Miller E, McBride B. RDF 1.1 primer. W3C, 20140225, 2014. <https://www.w3.org/TR/rdf11-primer/>
- [5] Brickley D, Guha RV. RDF schema 1.1. W3C, 20140225, 2014. <https://www.w3.org/TR/2014/REC-rdf-schema-20140225/>
- [6] Smith MK, Welty C, McGuinness DL. OWL Web Ontology Language Guide. W3C, 20040210, 2004. <https://www.w3.org/TR/2004/REC-owl-guide-20040210/>
- [7] Protégé ontology editor. <http://protege.stanford.edu/>
- [8] Sure Y, Angele J, Staab S. OntoEdit: Multifaceted inferencing for ontology engineering. *Journal on Data Semantics I*, 2003,2800: 128–152. [doi: 10.1007/978-3-540-39733-5_6]
- [9] KAON: The karlsruhe ontology and semantic Web tool suite. <http://www.aifb.kit.edu/web/KAON>
- [10] Antoniou G, Harmelen FV. *A Semantic Web Primer*. 2nd ed., Cambridge: MIT Press, 2008.
- [11] Kim IW, Lee KH. A model-driven approach for describing semantic Webservices: From UML to OWL-S. *IEEE Trans. on Systems, Man, and Cybernetics Part C: Applications and Reviews*, 2009,39(6):637–646. [doi: 10.1109/TSMCC.2009.2023798]
- [12] Iribarne L, Padilla N, Asensio JA, Criado J, Ayala R, Almendros J, Menenti M. Open-Environmental ontology modeling. *IEEE Trans. on Systems, Man and Humans*, 2011,41(4):730–745. [doi: 10.1109/TSMCA.2011.2132706]
- [13] Lee CS, Kao YF, Kuo YH, Wang MH. Automated ontology construction for unstructured text documents. *Data & Knowledge Engineering*, 2007,60:155–176. [doi: 10.1016/j.datak.2006.04.001]

- [14] Raufi B, Ismaili F, Zenuni X. Modeling a complete ontology for adaptive Web based systems using a top-down five layer framework. In: Proc. of the 31st Int'l Conf. on Information Technology Interfaces (ITI 2009). 2009. 511–518. [doi: 10.1109/ITI.2009.5196136]
- [15] Li J, Meng LS. Comparison of seven approaches in constructing ontology. *New Technology of Library and Information Service*, 2004,7:17–22 (in Chinese with English abstract).
- [16] Subhashini R, Akilandeswari J. A survey on ontology construction methodologies. *Int'l Journal of Enterprise Computing and Business Systems*, 2011,1(1).
- [17] Noy NF, McGuinness DL. Ontology development 101: A guide to creating your first ontology. Technical Report: KSL-01-05. Stanford Knowledge Systems Laboratory, 2001. [doi: 10.1016/j.artmed.2004.01.014]
- [18] Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. *Int'l Journal of Human and Computer Studies*, 1995,43(5/6):907–928. [doi: 10.1006/ijhc.1995.1081]
- [19] Chen DY, Zhao H, Zhang X, Zhao LJ, Ping A. Study on semantic Web security. *Journal on Communications*, 2012,33(Z2):45–59 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-436x.2012.z2.007]
- [20] Horrocks I, Patel-Schneider PF, Boley H, Tabet S, Grosz B, Dean M. SWRL: A semantic Web rule language combining OWL and RuleML. W3C, 20040521, 2004. <https://www.w3.org/Submission/SWRL/>
- [21] Garlik SH, Seaborne A. SPARQL 1.1 query language. W3C, 20130321, 2013. <https://www.w3.org/TR/sparql11-query/>
- [22] Tan SS, Chiarello P, Quentin W. Knee replacement and diagnosis-related groups (DRGs): Patient classification and hospital reimbursement in 11 European countries. *Knee Surgery, Sports Traumatology, Arthroscopy*, 2013,21(11):2548–2556. [doi: 10.1007/s00167-013-2374-6]
- [23] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proc of the KDD. New York: ACM, 2008. 1247–1250. [doi: 10.1145/1376616.1376746]
- [24] Noy N, Rector A, Hayes P, Welty C. Defining *N*-ary relations on the semantic Web. W3C, 20060412, 2006. <https://www.w3.org/TR/swbp-n-aryRelations/>
- [25] Apache Jena. <http://jena.apache.org/>
- [26] Chen DY, Zhao H. Data security and privacy protection issues in cloud computing. In: Proc. of the 2012 Int'l Conf. on Computer Science and Electronic Engineering (ICCSEE 2012). 2012. 647–651. [doi: 10.1109/ICCSEE.2012.193]
- [27] Nagi D. Exercise and Sport in Diabetes. 2nd ed., Wiley, 2005.
- [28] Zhou SH, Fan ZH. Practical Rehabilitation Medicine. Nanjing: Southeast University Press, 1989. 351–355 (in Chinese).
- [29] SNOMED CT. <http://www.ihtsdo.org/snomed-ct/>
- [30] MeSH (medical subject headings). <https://www.ncbi.nlm.nih.gov/mesh/>
- [31] Unified medical language system (UMLS). <https://www.nlm.nih.gov/research/umls/>
- [32] OpenGallen. <http://www.opengalen.org/>
- [33] The OBO foundry. <http://www.obofoundry.org/>
- [34] OBI ontology. <http://obi-ontology.org/>
- [35] Campana F, Moreno A, Riaño D, Varga LZ. K4Care: Knowledge-based homecare e-services for an ageing europe. In: Annicchiarico R, Cortés U, Urdiales C, eds. Proc. of the Agent Technology and e-Health. Whitestein Series in Software Agent Technologies and Autonomic Computing. Switzerland: Birkhäuser Verlag Basel, 2007. 95–115. [doi: 10.1007/978-3-7643-8547-7_6].
- [36] Cantais J, Dominguez D, Gigante V, Laera L, Tamma V. An example of food ontology for diabetes control. In: Proc. of the Int'l Semantic Web Conf., 2005 Workshop on Ontology Patterns for the Semantic Web. 2005.
- [37] Kostopoulos K, Chouvarda I, Kokonozi A, Gils MV. An ontology-based framework aiming to support personalized exercise prescription: Application in cardiac rehabilitation. In: Proc. of the 33rd Annual Int'l Conf. of the IEEE EMBS. Boston: IEEE, 2011. 1567–1570. [doi: 10.1109/IEMBS.2011.6090456]
- [38] Pramono D, Setiawan NY, Sarno R, Sarno M. Physical activity recommendation for diabetic patients based on ontology. In: Proc. of the 7th ICTS. Bali, 2013. 27–32.

- [39] Izumi S, Kuriyama D, Itabashi G, Takahashi K. An ontology-based advice system for health and exercise. In: Proc. of the 10th IASTED Int'l Conf. on Internet and Multimedia Systems and Applications (IMSA 2006). 2006. 95–100.
- [40] Beckett D, Berners-Lee T. Turtle-Terse RDF triple language. W3C, 20110328, 2011. <https://www.w3.org/TeamSubmission/turtle/>
- [41] Jena TDB. <http://jena.apache.org/documentation/tdb/index.html>
- [42] Reasoners and rule engines: Jena inference support. <http://jena.apache.org/documentation/inference/>
- [43] Du XY, Li M, Wang S. A survey on ontology learning research. Ruan Jian Xue Bao/Journal of Software, 2006,17(9):1837–1847 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/17/1837.htm> [doi: 10.1360/jos171837]
- [44] Gan JH, Jiang Y, Xia YM. Ontology Method and Its Application. Beijing: Science Press, 2011 (in Chinese).
- [45] Zhong XQ, Fu HG, She L, Huang B. Geometry knowledge acquisition and representation on ontology. Chinese Journal of Computers, 2010,33(1):167–174 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2010.00167]
- [46] Zhong XQ, Liu Z, Ding PP. Construction of knowledge base on hybrid reasoning and its application. Chinese Journal of Computers, 2012,35(4):761–766 (in Chinese with English abstract). [doi: 10.3724/SP.J.1016.2012.00761]
- [47] Chen XH, Yin B, Jin Z. Ontology-Guided requirements modeling based on problem frames approach. Ruan Jian Xue Bao/Journal of Software, 2011,22(2):177–194 (in Chinese with English abstract). <http://www.jos.org.cn/1000-9825/3755.htm> [doi: 10.3724/SP.J.1001.2011.03755]
- [48] Sivasankari S, Jacob SG. An automated ontology learning for benchmarking classifier models through gain-based relative-non-redundant (GBRNR) feature selection: A case-study with erythemato. Int'l Journal of Business Intelligence & Data Mining, 2018,1(1).
- [49] Wang Y, Gao W. Ontology learning algorithm using distance measure computation. Journal of Southwest China Normal University, 2018,43(1):40–46 (in Chinese with English abstract). [doi: 10.13718/j.cnki.xsxb.2018.01.006]
- [50] Lin Z. Terminological ontology learning based on LDA. In: Proc. of the Int'l Conf. on Systems and Informatics. IEEE, 2018. 1598–1603. [doi: 10.1109/ICSAI.2017.8248539]
- [51] Wang HF. Large scale knowledge graph technology. Communications of the CCF, 2014,10(3):64–68 (in Chinese with English abstract).
- [52] Liu Q, Li Y, Duan H, Liu Y, Qin ZG. Knowledge graph construction techniques. Journal of Computer Research and Development, 2016,53(3):582–600 (in Chinese with English abstract). [doi: 10.7544/issn1000-1239.2016.20148228]
- [53] Ruan T. Vertical knowledge graph making tools and its application in industry. 2014 (in Chinese). <http://www.cipsc.org.cn/kg2/>
- [54] Fang ZJ, Wang HF, Gracia J, Gil JB, Ruan T. Zhishi.lemon: On publishing zhishi.me as linguistic linked open data. In: Proc. of the 15th Int'l Semantic Web Conf. (ISWC 2016). Kobe: Springer-Verlag, 2016. [doi: 10.1007/978-3-319-46547-0_6]
- [55] Wang HF, Wu TX, Qi GL, Ruan T. On publishing Chinese linked open schema. In: Proc. of the 15th Int'l Semantic Web Conf. (ISWC 2014). Trentino: Springer-Verlag, 2014. [doi: 10.1007/978-3-319-11964-9_19]
- [56] Ruan T, Xue LJ, Wang HF, Hu FH, Zhao L, Ding J. Building and exploring an enterprise knowledge graph for investment analysis. In: Proc. of the 15th Int'l Semantic Web Conf. (ISWC 2016). Kobe: Springer-Verlag, 2016. [doi: 10.1007/978-3-319-46547-0_35]
- [57] Jadhav SB, Pardeshi SN. Ontology intergration with semantic similar entity classes amongst different ontologies for enhanced information retrieval. Int'l Journal of Recent Trends in Engineering, 2009,2(3):132–134. [doi: 10.1.1.381.7277]
- [58] Zaiß K, Schlüter T, Conrad S. Instance-Based ontology matching using different kinds of formalisms. Int'l Journal of Computer, Electrical, Automation, Control and Information Engineering, 2009,3(7):1716–1724. [doi: 10.1.1.193.3112]
- [59] Lambrix P, He T. Ontology alignment and merging. Computational Biology, 2008,6:133–150. [doi: 10.1007/978-1-84628-885-2_6]

附中文参考文献:

- [15] 李景,孟连生.构建知识本体方法体系的比较研究.现代图书情报技术,2004,7:17–22.
- [19] 陈德彦,赵宏,张霞,赵立军,平安.语义 Web 安全研究.通信学报,2012,33(Z2):45–59. [doi: 10.3969/j.issn.1000-436x.2012.z2.007]
- [28] 周士仿,范振华.实用康复医学.南京:东南大学出版社,1989.351–355.
- [43] 杜小勇,李曼,王珊.本体学习研究综述.软件学报,2006,17(9):1837–1847. <http://www.jos.org.cn/1000-9825/17/1837.htm> [doi: 10.1360/jos171837]
- [44] 甘健侯,姜跃,夏幼明.本体方法及其应用.北京:科学出版社,2011.

- [45] 钟秀琴,符红光,余莉,黄斌.基于本体的几何学知识获取及知识表示.计算机学报,2010,33(1):167-174. [doi: 10.3724/SP.J.1016.2010.00167]
- [46] 钟秀琴,刘忠,丁盘苹.基于混合推理的知识库的构建及其应用研究.计算机学报,2012,35(4):761-766. [doi: 10.3724/SP.J.1016.2012.00761]
- [47] 陈小红,尹斌,金芝.基于问题框架的需求建模:一种本体制导的方法.软件学报,2011,22(2):177-194. <http://www.jos.org.cn/1000-9825/3755.htm> [doi: 10.3724/SP.J.1001.2011.03755]
- [49] 王燕,高伟.基于距离测度计算的本体学习算法.西南师范大学学报(自然科学版),2018,43(1):40-46. [doi: 10.13718/j.cnki.xsxb.2018.01.006]
- [51] 王昊奋.大规模知识图谱技术.中国计算机学会通讯,2014,10(3):64-68.
- [52] 刘峤,李杨,段宏,刘瑶,秦志光.知识图谱构建技术综述.计算机研究与发展,2016,53(3):582-600. [doi: 10.7544/issn1000-1239.2016.20148228]
- [53] 阮彤.垂直知识图谱构造工具与行业应用.2014. <http://www.cipsc.org.cn/kg2/>



陈德彦(1977-),男,博士,正高级工程师,主要研究领域为自然语言处理,语义 Web,知识工程,数据挖掘,机器学习,网络与信息安全.



张霞(1965-),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为软件架构,软件工程,数据库,大数据,人工智能.



赵宏(1954-),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为下一代网络,网络与信息安全,网络管理,图像处理.