

《智能信息处理》课程考试

基于本体建模的动态知识图谱构建技术研究

董静阳

考核	到课[10]	作业[20]	考试[70]	课程成绩[100]
得分				

2021 年 12 月 20 日

基于本体建模的动态知识图谱构建技术研究

董静阳

(大连海事大学 计算机科学与技术 辽宁省大连市 中国 116026)

摘 要 : 知识图谱是以图的形式表现客观世界中的概念和实体及其之间关系的知识库,是语义搜索、智能问答、决策支持等智能服务的基础技术之一。目前知识图谱领域构建的图谱多以静态知识图谱为主,数据集较大时,构建出的静态知识图谱的规模会非常大,且存在可读性差、交互性低等问题,非常不利于信息的动态浏览和筛选。为解决这一问题,应以增加知识图谱的动态交互性为目标来完成动态知识图谱的构建。将从知识图谱的体系架构、本体建模等方面进行系统探索,给出基于本体建模的动态知识图谱构建技术流程。最后从知识推理等方面对未来的研究方向进行了展望。

关键词 知识图谱 本体建模 信息可视化

中图法分类号 TP311.20 DOI 号 10.3970/j.issn.1001-3695.2019.11.031

Research on construction technology of dynamic knowledge graph based on ontology modeling

Dong Jingyang

(Computer science and technology, Dalian maritime university, Liaoning Dalian, 116026, China)

Abstract Knowledge graph is a knowledge base that represents concepts and entities in the objective world and the relationships between them in the form of graphs, and is one of the basic technologies for intelligent services such as semantic search, intelligent Q&A, and decision support. However, most of the maps constructed in the field of knowledge graphs are based on static knowledge graphs. When the data set is large, the scale of the constructed static knowledge graphs will be very large, and there will be problems such as poor readability and low interactivity, which are very unfavorable for information dynamic browsing and filtering. In order to solve this problem, the construction of the dynamic knowledge graph should be completed with the goal of increasing the dynamic interactivity of the knowledge graph. This paper systematically explores the architecture and ontology modeling of the knowledge graph, and gives the technical process of dynamic knowledge graph construction based on ontology modeling. Finally, the future research directions are prospected from knowledge reasoning and other aspects.

Key words dynamic knowledge graph; ontology modeling; information visualization

1 介绍

由于信息技术的快速发展，互联网内容规模大以及组织结构松散的特点，给信息和知识的有效获取提出了挑战。谷歌率先提出知识图谱的概念，并将其应用，知识图谱技术很快就得到了广泛的关注。知识图谱(knowledge graph)是以图的形式表现客观世界中的实体(概念、人、事物)及其之间关系的知识库。目前在知识图谱领域，静态知识图谱的发展已经较为成熟，但是动态知识图谱向前发展仍需要解决很多问题，例如：如何构建规范的、可扩展的研究领域数据模型；如何高效、完整地获取研究领域研究数据等，本文着眼于动态知识图谱的构建，从知识图谱的构建架构、本体建模等方面进行系统探索，给出基于本体建模的动态知识图谱构建技术流程。最后从知识推理等方面对未来的研究方向进行展望。

2 知识图谱的定义与体系架构

2.1 知识图谱的定义

知识图谱是以图的形式表现客观世界中的实体(概念)及其之间关系的知识库。知识图谱以实体、概念以及属性作为基本组成单位，通过实体与实体、实体与属性之间的关系，来构成类似于网状的结构化知识库，便于信息的检索和知识的推理。知识图谱本质上是一种语义网络。其节点代表实体(entity)或者概念(concept)，边代表实体或概念之间的关系,可以帮助形式化描述、理解现实世界的事物及其相互关系。随着互联网技术的快速发展，知识图谱的应用也从搜索引擎逐步扩展到各个领域。例如推荐系统、智能问答和文本分析。知识图谱通过对错综复杂的文档数据进行有效的加工、处理、整合，转化为简单、清晰的“实体，关系，实体”的三元组，最后聚合大量知识，从而实现知识的快速响应和推理。

2.2 知识图谱的体系架构

知识图谱主要有自底向上与自顶向下两种体系架构(构建)方法，有时这两种方法也会结合起来使用。自底向上的构建方法如图 2.1 所示，首先从一些非结构化数据、半结构化数据源中抽取实体、关系和属性，将其加入到知识库中形成数据层。然

后对已经处理好的数据层进行概念抽象，最后形成模式层。此构建方法主要适用于数据量较大的知识图谱的构建。本文就是采用自底向上的构建方法完成研究领域动态知识图谱的构建。

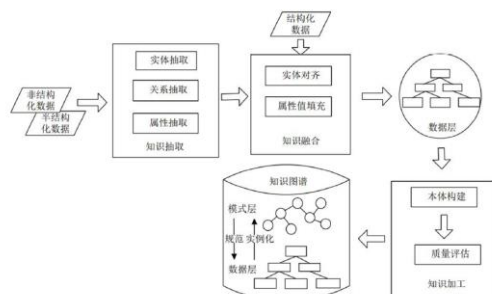


图 2.1 自底向上的知识图谱体系结构

自顶向下的构建方法如图 2.2 所示，其构建过程与自底向上的构建方法相反，先进行顶层设计构建模式层，然后再对本体概念和本体关系进行细分，最后将实体加入到知识库。该构建方法人工依赖性强，适用于数据量较小的知识图谱的构建。

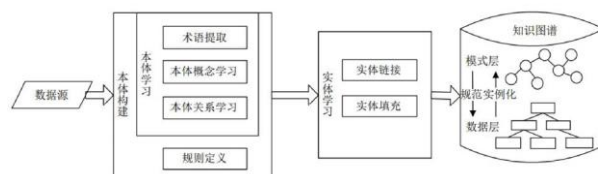


图 2.2 自顶向下的知识图谱体系结构

3 本体模型的构建

本体建模是动态知识图谱构建过程中非常重要的环节，可以很好地梳理实体以及实体间关系。在建立研究领域研究数据的本体模型时应遵从 Gruber 提出的完全性、一致性等原则，合理规范的进行研究领域的本体构建。本文将从研究领域相关数据源出发，参考“七步法”步骤，采用 Protege5.5 本体编辑工具实现领域本体的构建过程。

3.1 本体建模工具

为了方便本体领域的发展，许多科研团队开发了各种类型的建模工具，包括本体的建立、存储、查询、推理、查询等。具体而言，编辑工具应有以下特征：

1) 支持输入输出不同格式的语言，可以定义不同概念间的约束条件；

2) 具有便于用户操作使用的图形界面, 简化编辑类和属性的处理过程;

3) 具有一定的逻辑推理能力, 能够实现简单的案例推理, 并能检查一致性;

4) 支持不同结构语义间的转换, 对语义的维护具有可靠性;

5) 对工具本身的维护操作简单, 有利于后续对本体进行扩展。

由于本体语言具有强大的功能, 例如在可扩展性、灵活性、易用性等方面具有突出优势, 为了展现出本体语言的优点所使用的工具也不同。在众多编辑工具中, 较为著名的有 Ontolingua、WebOnto、OntoEdit、Protégé 等。

因 Protégé 安装便捷、使用简单, 支持创建不同复杂度的本体模型, 可以对不同版本的模型进行操作, 例如输入、编辑和输出等, 同时还可以定义具体类和抽象类。为了给用户更好的使用环境, Protégé 支持用户自定义图形界面。其具有可扩展的结构, 可以根据用户自身需求增加想要的功能模块。

3.2 本体建模方法

本体的构建方法有: 骨架法、TOVE 法、METHONTOLO 法、IDEF5 法、SENSUS 法、KACTUS 工程法、七步法、循环获取法、五步法等。本文在建立研究领域的本体模型时, 将参考七步法完成本体模型的构建, 具体过程如下:

1) 确定本体应用的学科领域和范畴

率先给出限制对于复杂系统的实现是有利的, 因此先明确本体表示的范围和突出优势以及本体的系统运维人员和使用对象。

2) 考虑复用现有本体的可能性

3) 列出本体中的重要术语

将该学科范围内的所有概念名词以及概念名词的具体说明列举出来。在某些学科领域中, 这些概念名词作为重要术语需要将其陈列出来, 这样有助于构建更专业的本体模型。

4) 定义类和类的等级体系

5) 定义类的属性

6) 定义概念之间的关系

对于已经构建好的类进行分析, 总结不同类之间的关系。

7) 创建实例

3.2 本体建模技术路线

3.1 节和 3.2 节分别给出了本体建模的工具以及参照的方法, 本节将给出本体建模的技术路线, 具体如图 3.1 所示。

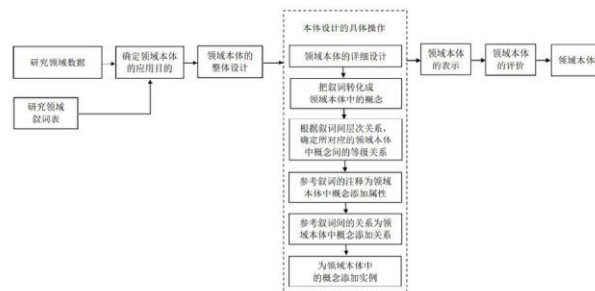


图 3.1 本体建模的技术路线

以相关研究领域的数据为基础, 首先确定构建领域本体应用的目的, 即这个本体模型是要用来做什么; 然后对领域本体的整体进行设计, 分析列举出有哪些类; 接着进行详细设计, 对之前设计的类进行细分, 分析不同类之间的关系以及类的属性; 将设计完整的本体用相关软件以图的形式表示出来; 之后对构建的领域本体进行评价, 判断构建的本体是否存有需要再次改进和完善的地方, 最后完成本体的构建。例如, 图 3.2 展示了某一领域构建好的本体的可视化关系。

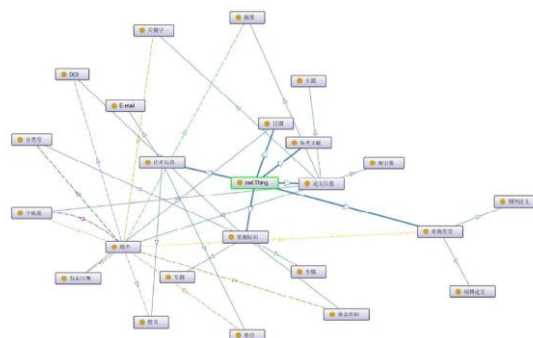


图 3.2 本体可视化

4 动态知识图谱的构建

本文将基于 protégé 本体建模和 Neo4j 图数据库相结合的方法, 应用到相关领域动态知识图谱的构建中。在上一节中已经构建好了研究领域的本体模型, 定义好了研究领域实体和实体间的关系。本节基于构建好的本体结合 Neo4j 图数据库, 且利用 Cypher 语言对相关领域中的实体、关系以及属性进行设计、编译, 从而构建出相关领域的动态知

识图谱。

4.1 Neo4j 图数据库介绍

Neo4j 图数据库是处理图论等关系时常被研究者学习和使用的高性能开源数据库，和关系型数据库一样完全兼容事务的 ACID 特性，即原子性(A)、一致性(C)、隔离性(I)、持久性(D)。由于具有良好的性能，成熟的系统等特点而被用户广泛应用。

基于 Neo4j 的图数据库模型如图 4.1 所示，Neo4j 的信息建模包括图数据节点、节点之间的关系、节点包含的属性以及关系包含的属性。Neo4j 中的节点可以包含多个属性，同时可以与其他任何节点建立相关关系，但是建立的每组关系中必须含有一个初始节点和一个终止节点，且每个关系也可以包含多个属性。同时 Neo4j 可以存储海量数据并且具有较高的查询效率，因此本文将利用 Neo4j 图数据库来构建知识图谱。

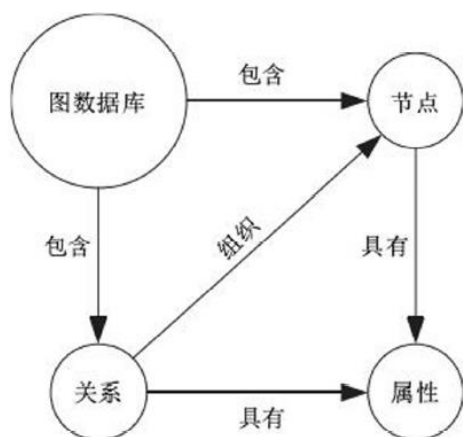


图 4.1 Neo4j 的图数据库模型

使用 Neo4j 图数据库时，其相匹配的查询语言为 Cypher 语言，Cypher 语言入门容易可以快速上手使用，具有非常强大的性能。例如输入“MATCH (n1 { title:“小麦” }) — [] —> (n2) RETURN n1, n2””，最终得到的查询结构如图 4.2。

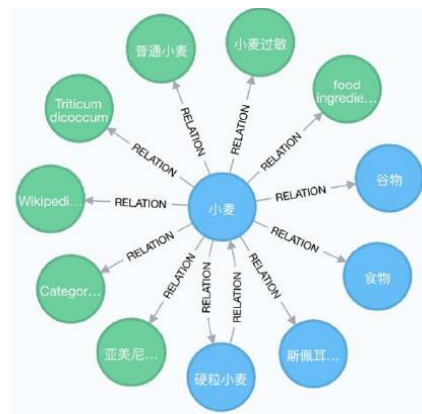


图 4.2 Neo4j 的图数据库模型

4.2 本体到Neo4j的映射

在建立好本体模型后，可以将数据实例按照此前定义的和类间关系存入图数据库中。图数据库存储的知识表示形式和基于本体的知识表示形式有所不同，Neo4j 中的数据等价于本体中类的具体实例。因此为了实现数据的存储，需要把这两种表示模型进行相互映射，即将本体模型中的知识表示形式映射为可以存储在 Neo4j 数据库中的知识表示形式。依据 Neo4j 数据库的存储方式，本文将研究领域的本体模型和 Neo4j 图模型中的基本元素按如图 4.3 所示映射，具体解释说明如下：

1) 节点映射

Neo4j 图数据库中存储的节点是本体模型中不同类的具体数据实例。例如，对农业信息化文献而言，在本体中是类，在 Neo4j 中是实例，将这些含义清晰的、具体的实例作为 Neo4j 中的节点。

2) 关系映射

Neo4j 图数据库中的关系用于连接本体模型中不同的类，包括类之间的“Sub Class Of”关系。构建本体模型时自定义的两个类之间存在的某种关系作为类间关系，如“前驱关系”等，因此将本体中类间关系转换为图数据库中节点与节点间的关系。

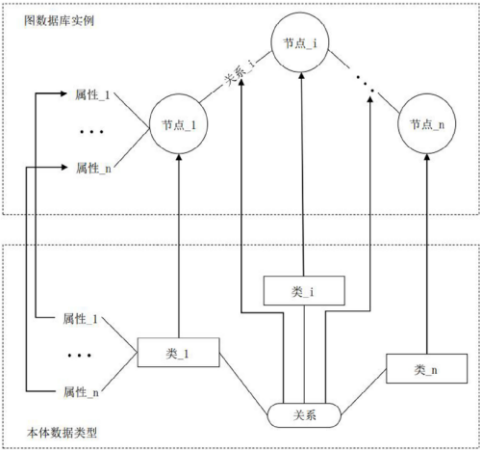


图 4.3 数据映射图

3) 属性映射

在本体模型中，数据属性是对不同类特征的描述，如果在图数据库中将其存储为节点的形式，在展示的过程中会有很多冗余数据出现，影响最终可视化显示的效果，在 Neo4j 图数据库中，可以将数据的特征描述作为节点的属性存储。Neo4j 的数据类型由本体确定，按照本体模型中研究对象的类名称、属性和类间关系分别转化为 Neo4j 中的节点标签、节点属性和节点关系，在 Neo4j 中的节点表示本体模型中不同类型的实例。在具体实现方式上，可通过 Python 语句选取 MongoDB 数据库中相应的数据自动存储到 Neo4j 中，利用 Cypher 中的 CREATE 语句创建图结构，将数据三元组转化成图谱中节点和边，从而实现知识的可视化。

4.3 可视化平台设计

将数据存储到 Neo4j 图数据库，进行数据库设计后，动态知识图谱就构建完成。因 Neo4j 自带可视化查询界面，所以可直接使用 Cypher 语句进行知识图谱的动态查询。但是并不是所有人都会使用 Cypher 语句，因此为了方便使用者使用，还需开发一个可视化平台，图 4.4 为可视化平台的总体架构。

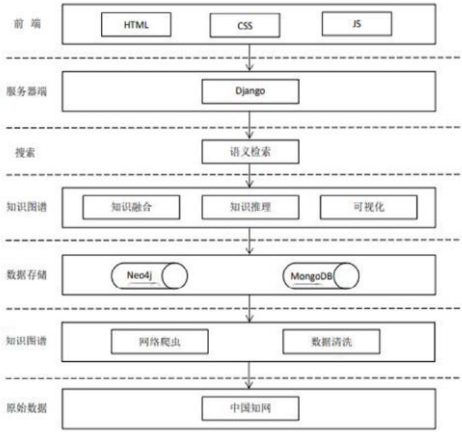


图 4.4 可视化平台总体架构

该平台采用 B / S 架构，平台从下端往上端依次是：原始数据模块；数据采集处理模块；数据存储模块；知识图谱模块；搜索模块；服务器模块；前端设计模块。

结束语

本文通过比较与分析给出了基于本体建模的动态知识图谱构建技术，将基于 protégé 本体建模和 Neo4j 图数据库相结合的方法，应用到动态知识图谱的构建中。改进了单纯的依靠 protégé 本体建模和单纯的依靠 Neo4j 图数据库构建知识图谱的方法。同时秉承交互友好性原则进行了平台的总体设计，给出了基于知识图谱的信息可视化搜索查询平台的构建流程，构建了基于知识图谱的信息资源智能检索可视化应用平台，达到了知识图谱的动态交互性。

但是本文仅对动态知识图谱的构建进行了研究，对于构建好的图谱后期的知识推理过程并没有详细阐述，目前的知识推理过程多数还是依赖人工参与。后期研究的重点应该放在知识推理 [15] 过程中，将知识图谱与深度学习领域 [16] 有机结合。

参 考 文 献

- [1]朱冬亮,文奕,万子琛.基于知识图谱的推荐系统研究综述[J/OL].数据分析与知识发现:1-16[2021-12-10].
- [2]张吉祥,张祥森,武长旭,赵增顺.知识图谱构建技术综述[J/OL].计算机工程:1-16[2021-12-10].<https://doi.org/10.19678/j.issn.1000-3428.0061803>.
- [3]姜强,药文静,赵蔚,李松.面向深度学习的动态知识图谱建构模型及评测[J].电化教育研究,2020,41(03):85-92.
- [4]冯俐. 基于 Neo4j 图数据库构建中学语文诗词知识图谱[D].陕西师范大学,2019.
- [5]桑盛田,杨志豪,刘晓霞,王磊,赵迪,林鸿飞,王健.融合知识图谱与深度学习的药物发现方法[J].模式识别与人工智能,2018,31(12):1103-1110.
- [6]焦爽. 基于小规模知识图谱的语义相似度计算[D].聊城大学,2018.
- [7]陆丽娜,王萍,于啸.农业科学数据监管平台构建研究[J].图书情报工作,2017,61(10):68-73.DOI:10.13266/j.issn.0252-3116.2017.10.009.
- [8]宫云宝. 面向知识图谱构建的知识验证方法研究[D].国防科学技术大学,2016.
- [9]王红,张青青,蔡伟伟,姜洋.基于 Neo4j 的领域本体存储方法研究[J].计算机应用研究,2017,34(08):2404-2407.
- [10]王余蓝.图形数据库 NEO4J 与关系数据库的比较研究[J].现代电子技术,2012,35(20):77-79.DOI:10.16652/j.issn.1004-373x.2012.20.045.
- [11]章勇,吕俊白.基于 Protege 的本体建模研究综述[J].福建电脑,2011,27(01):43-45.
- [12]徐国虎,许芳.本体构建工具的分析与比较[J].图书情报工作,2006(01):44-48.
- [13]顾慧翔,俞勇.基于领域本体和知识推理的语义互联网应用[J].上海交通大学学报,2004(04):583-585.DOI:10.16183/j.cnki.jsjtu.2004.04.022.