《智能信息处理》课程考试

本体与知识图谱对比及构建方法研究

李征蔚

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

2021 年 12 月 18 日

本体与知识图谱对比及构建方法研究

李征蔚

(大连海事大学 信息科学与技术学院 辽宁大连 116026)

摘 要 针对当下本体和知识图谱两个专业术语混用严重的问题，本文通过梳理当前已有的相关文献和重要知识资源，对本体和知识图谱以及二者的联系和区别进行了概述，为当前和今后的相关研究和课题的开展提供一定的参考，从而推进本体和知识图谱研究的进一步发展。研究表明，知识图谱与网络空间安全领域业务场景相结合，可实现知识检索、基于算法或数据分析等功能。该方法也可为基于信息检索技术的网络空间安全知识服务提供依据。

关键词 本体；知识图谱；网络空间安全领域；业务场景；知识检索

Comparison of Ontology and Knowledge Graph and Research on Construction Methods

Li Zhengwei

(Dalianmaritimeuniversity,Computerscienceandtechnology,LiaoningDalian,116026)

Abstract In view of the serious problem of the current mixed use of the two professional terms of ontology and knowledge graph, this article summarizes the relationship and difference between ontology and knowledge graph, as well as the current and future The development of related research and topics provides a certain reference, so as to promote the further development of ontology and knowledge graph research. Research shows that the combination of knowledge graphs and business scenarios in the cyberspace security field can realize functions such as knowledge retrieval, algorithm-based or data analysis. This method can also provide a basis for cyberspace security knowledge services based on information retrieval technology.

Keywords Ontology;Knowledge Graph;Cyberspace Security Field; Business scene; Knowledge retrieval

1 引言

随着本体和知识图谱同是重要的知识组织表达形式，目前已经被普遍应用于人工智能、自然语言处理、软件工程、医学信息学以及图书馆学等领域，虽然二者有一定的内在联系，但是它们还是有实质上的差别[1]。为此，本篇论文对本体和知识图谱之间的联系和区别展开相关探索和研究。近几年，知识图谱的语义理解在各行业领域中发挥了具大作用，引入知识图谱解决网络空间安全知识表达、共享、分析和应用等问题，可以推动网络空间安全领域智能化发展。在此背景下，笔者提出了一种基于本体的网络空间安全知识图谱构建方法[2]，从本体层建模、数据层映射和存储层可视化 3 个方面完整阐述知识图谱的构建流程。

2 本体与知识图谱对比

2.1本体

追根溯源，本体（Ontology）概念来源于哲学，在 20 世纪 90 年代被引入到人工智能、图书情报和知识工程等领域，从此本体一直成为众多领域的热门研究话题。关于本体的定义一直是众说纷纭，没有定论。Studer 等人在 1998 年提出本体的定义：本体是共享概念模型的明确的形式化规范说明此定义在学术界具有较大的影响，对于本体研究具有重要意义。在本体研究发展的过程中，描述本体的语言有很多种[3]，其中基于谓词逻辑的本体描述语言和基于 Web 的本体描述语言是最具代表性的两类。通常来说，根据本体的应用领域不同可以将本

体分为领域本体和上层本体两类。

2. 2知识图谱

知识图谱（Knowledge Graph，KG）本质上是一种大规模的语义网络，其概念于 2012 年 5 月由 Google 正式提出，初衷是为了用户能够更快更简单地发现新的信息和知识。知识图谱由节点和边组成，其中节点表示实体或概念，边代表两个实体或概念之间的语义关系，属性是一个键值对，每个实体或关系可以有一个或多个属性[4]，为实体和关系提供信息。

2.3 本体与知识图谱的联系

知识图谱的构建过程如图 1 所示，其中包括息抽取、知识表示、知识融合、知识推理四个部分信息抽取是从结构化、半结构化和非结构化数据中通过自动化或者半自动化的技术抽取有价值的信息，其中包括实体抽取，语义类抽取，属性和属性值抽取，关系抽取；知识表示方法主要是以 RDF 的三元组来符号性描述实体间的关系，近年来采用深度学习技术将实体的语义信息表示为稠密低维实值向量的方法开始兴起[5]。对于本体和知识图谱的联系主要涉及知识融合和知识推理这两个部分。

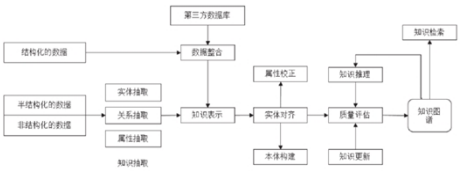

图 1 知识图谱体系架构

2.3.1 知识融合下二者的联系

知识融合指将来自多个数据源的知识进行融合后集成到知识图谱中。就逻辑结构而言，知识图谱可分为模式层与数据层，模式层构建在数据层之上。知识图谱的模式层通常采用本体库来保存，而数据层主要是采用图数据库来管理。知识融合阶段主要是对数据进行本体对齐和实体匹配。本体对齐就是判断和处理来自不同本体的两个实体是否指向一致，达到数据的统一，发生在知识图谱的模式层，涉及的是模式层的融合，包括概念的合并，概念上下位关系合并，概念的属性定义合并。而实体匹配是为了发现来源于多个数据源而具有不同 ID 却代表同一对象的实体，将这些实体融合为一个具有全局唯一 ID 的实体[6]，然后添加到知识图谱中，主要发生在数据层，更多涉及的是数据层的融合，匹配结果类型分为一对一，一对多和多对一 3 种。由于知识图谱的构建为了保证模式层的可靠性，模式层基本上通过人工校验。因此，知识融合的主要任务是数据层的融合。

2.3.2 知识推理下二者的联系

知识推理是在现有知识图谱的基础上通过各种方法进一步挖掘隐含的知识、结论或识别出知识图谱中错误的知识，从而丰富和拓展知识图谱，主要分为本体推理和规则推理，推理的对象不仅仅局限于实体层面还涉及本体库中概念的层次结构等。基于本体的推理，体现在本体层面，主要是通过预定义的本体公理进行推理，例如对于性别男、女是交集为空的两个类，那么一定不会存在一个人的性别既是男又是女。基于自定义规则的推理，可以根据特定的场景制定规则，来实现自定义的推理过程。推理关系规则，定义父亲的母亲是祖母，已知 a 是 b 的父亲，b 是 c 的母亲，则可以推出 a 是 c 的祖母。

2. 4本体与知识图谱的区别

基对于本体和知识图谱表达的信息方面而言，本体表达的是领域内共同认可的概念和概念间的关系，它反映的是常识或相对恒定的知识，不具备情报价值。譬如，Wordnet、Hownet 和 Cyc 都是国内外主要的通用本体库，是由众多行业专家经过多年手工编制的结果，其知识具备稳定性而不具备情报性，通常知识图谱则是情报挖掘的结果[7]。知识图谱构建过程的知识抽取环节，从结构化、半结构化和非结构化的数据中进行信息抽取，形成知识存入知识图谱中。谷歌知识图谱中所涉及的实体、实体间的关系以及其他相关信息并不是相对恒定的知识，具有流动性。

对于自然语言理解而言，语义消歧是其中的基础问题，是研究热点也是研究难点。在句法知识或者单独的句法不能实现消歧的情况下，本体作为一个支撑性的知识，有助于实现语义区分，实现对语句的正确理解。语言理解之后的信息抽取，涉及哪些实体以及实体间的发生何种关系，都可以从知识图谱中得到。对于结构而言，本体描述了知识图谱的模式层，提供对相关领域知识的共同理解，突出和强调概念以及概念之间的关联关系[8]。知识图谱则是在本体构建的模式层的基础上添加更多实体的信息，不断丰富和扩充。

3 网络空间安全知识图谱构建

3.1本体层建模

本体层建模本体层是知识图谱的核心层次，分析和细化网络空间安全知识内部的概念及关系，形成具有良好结构的概念层次树，并以本体语义关系形式表达，作为构建网络空间安全知识图谱的实体及关系的结构框架。

3.2 网络空间安全知识概念分类

网络空间安全知识体系结构复杂，主要涉及基础设施安全、系统安全、应用安全以及数据安全等。本文主要面向网络攻防知识构建图谱，因此，将网络空间安全知识划分为资产、术语和方式 3 类概念，又细分为 12 类子概念[9]。

(1) 资产:主要指网络空间安全中相关对象，包括主体、目标等子概念，例如“网络管理员”、“服务器”。

(2) 术语:主要指网络空间安全中相关名称，包括软件、硬件、系统、协议、算法、语言、病毒、漏洞等

子概念，例如“IE 浏览器”、“交换机”、“Windows 操作系统”、“ARP 协议”、“AES 算法”、“C 语言”、“黑色星期五病毒”、“代码注入漏洞 Apache Struts”。

(3) 方式:主要指网络攻防的具体技术，包括攻击方法与防御措施等子概念，例如“DDoS”、“限制 SYN 流量”。

3.3数据层映射

实体可以看作是本体的实例，通过基于 Bi － LSTM 模型与 CRF 模型相结合的实体识别方法，从结体化和非结构化文本中识别网络空间安全领域中 12 类实体，并采用三元组表达实体及实体间的关系，通过本体与知识图谱的映射匹配机制填充数据层。

3.3.1 实体识别

根据网络空间安全领域特征，基于网络空间安全教程、漏洞数据库、网络安全术语词典以及科学文献，经人工提取语料共计 5 000 条。对语料分析后编写网络空间安全领域词典，使用 O = { per，net} 表示主体和目标，E = { soft，hard，sys，prot，alg，prog，vir，vul} 表示软件、硬件、系统、协议、算法、语言、病毒和漏洞，M = { att，def} 表示攻击方法和防御措施。在 Jieba 中使用自定义的网络空间安全词典对语料句进行分词，词性标注采用 BIO 标注法，B 代表实体开头，I 代表实体中间，O 代表其他非实体，例如 B －per 和 I － per 代表主体的开头与中间。将语料的词向量矩阵与领域词典输入实体识别模型，模型架构如图 2 所示，包括 look － up 层、双向 LSTM 层和 CRF 层[10]，模型输入是网络空间安全领域语料词向量，加载训练集与测试集后，使用 Bi － LSTM 模型对输入的词向量序列进行特征提取，再将特征供给 CRF 模型选择出最适合的 tag 序列完成实体识别。

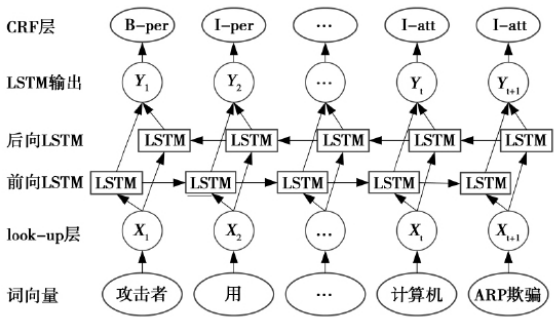

图 2 实体识别模型架构

3.3.2 本体与知识图谱映射匹配机制

网络空间安全知识图谱基本结构包括本体层关系图 OG 和数据层关系图 DG，即 KG = ＜ OG，DG＞ 。其中，本体层关系图表示概念间的层次关系; 数据层关系图表示实体及实体间的关系。本体层关系图 OG = ＜ CO，RO ＞ ，其中 CO 表示概念节点，RO 表示概念之间的关系边。数据层关系图 DG = ＜ ED，RD ＞ ，其中 ED 表示实体节点，RD 表示实体之间的关系边。网络空间安全知识图谱的图结构由节点和边构成，即 KG = { ＜N＞ ，＜R＞ } 。＜N＞ 表示节点集合，且 N∈(CO∪ED) ; ＜R＞ 表示边集合，且 R∈(RO∪RD) 。

本体与知识图谱的映射是树与图之间的映射，如图 3 所示，将本体的实例及其语义关系完整的映射到知识图谱中，形成网络空间安全知识图谱的数据层关系图。

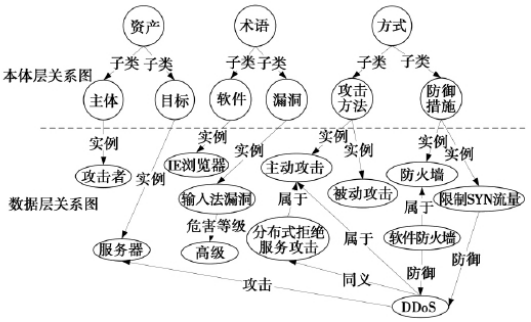

图 3 网络空间安全知识图谱数据层关系

3.3.3 存储层可视化

Neo4j 是 Neo technology 研发的图数据库，可将数据存储于图模型中，实现处理数据间关系的存储与查询。Neo4j 模型中的数据源实例名作为实体节点，如“DDoS”、“输入法漏洞”等; 关系边连接数据源中不同类型的实体节点，如“DDoS”与“服务器”之间的“攻击”关系; 属性作为实例性质的基本描述，如“危害等级为高级的输入法漏洞”等。网络空间安全实体、关系和属性以列的形式写入到 Excel

文档后存储为 CSV 格式文件，利用 Cypher 语句将文件中数据导入到 Neo4j 中构建知识图谱，分别使用 Node、Relationship 和 Property 数据类型建立节点、关系和属性。图 4 为部分网络空间安全知识图谱，从图 4 可以看出，每一个节点会根据它的类型定义不同的特征，实体在 Neo4j 中可视化为圆图，节点关系通过边表示，对象属性则单独展示。

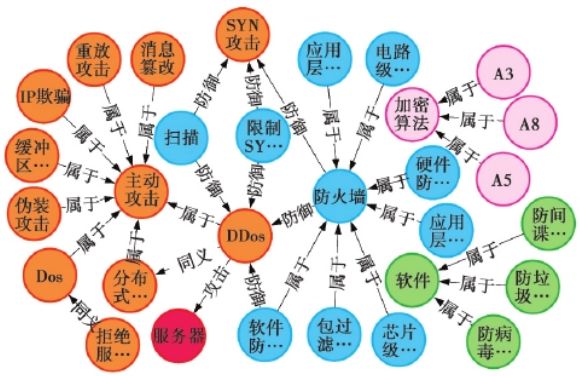

图 4 网络空间安全知识图谱

4 知识图谱应用研究

知识图谱与网络空间安全领域业务场景相结合，可支持知识检索、智能问答和数据分析等任务。

(1) 知识检索。基于知识图谱优化领域知识管理模式，根据实体、属性、关系等关键字段进行多维度查询，以图谱形式展示检索结果，也可依据检索结果进行知识推理，生成辅助安全防御决策信息[11]。

(2) 智能问答。结合网络空间安全语境意图模型，使用知识图谱实现语义理解，分析用户自然语言提问，将非结构化的语义表述转化为结构化的语义表述，实现人机交互问答。

(3) 数据分析。基于网络空间安全领域多源数据，通过知识图谱增强数据之间的语义关联，对数据进行关联分析与挖掘。

结束语

随着互联网、人工智能等行业的迅猛发展，本体和知识图谱作为重要的知识组织表达手段，不仅可以将海量数据表达成更接近人类认知现实世界的形式，还提供一种更好的组织、管理和利用信息的方式。加之，本体和知识图谱相辅相成的紧密关系，只有将二者共同发展强大才能满足人类对海量数据管理和利用的需求。

(1) 基于本体与知识图谱的映射匹配机制，实现了一种网络空间安全知识图谱构建方法。面向网络攻防知识进行资产、术语和方式 3 类本体构建，定义了通用语义关系和自定义语义关系。

(2) 自建网络空间安全语料库，使用 Bi －LSTM + CRF 模型进行实体识别，在 Neo4j 图数据库中对知识图谱进行存储与展示。本文的研究将对网络空间安全知识图谱的构建提供一定的参考，但仍有进一步改进的空间，未来工作主要在以下两个方面: 一是网络空间安全知识图谱关系抽取采用人工标注完成，未来可进行关系识别技术研究，实现知识图谱的自动化构建。二是基于已经构建的网络空间安全知识图谱，用户还不能快速、方便地检索信息，需要探索以知识网络为支撑的智能应用，实现辅助决策信息的有效输出。

参 考 文 献

[1]Lu Ruqian,Zhang Songmao.PANGU — An agent-oriented knowledge base.In Processing of Conference on Intelligent Information Processing(16th WCC 2000)：486-493

[2]金芝，知识工程中的本体论研究。世纪之交的知识工程与知识科学。清华大学出版社 2001：468-477

[3]Kyung-Yong Jung,Kee-Wook Rim,and Jung-HyunLee.Automatic Preference Mining through Learning User Profile with Extracted Information[C].SSPR&SPR2004,LNCS3138,2004:815–823.

[4]Yue feng Li,Y.Y. Yao.User Profile Model:A View from Artificial Intelligence[C].RSCTC 2002,LNAI 2475,200:493–49.

[5]ugene Santos,Hien Nguyen.Empirical Evaluation of Adaptive UserModelingina Medical Information Retrieval Application[C].UM 2003,LNAI 2702:292–296.

[6]徐焕良.企业知识资源计划及其关键技术研究[D].南京航空航天大学博士论文.2003.10.

[7]徐焕良,李绪荣,丁秋林.基于角色模型的业务过程再工程(BPR)的研究.[J].计算机科学.2003 Vol.30 No.1：154—157.

[8]张磊，谢强，王金栋，丁秋林.基于业务过程的知识需求研究[J].吉林大学学报（信息科学版）.2005.Vol 23 No.5.

[9]顾丹阳,李明倩,权冀川,刘勇,罗晨.基于本体的主战武器装备知识图谱构建[J].指挥控制与仿真,2021,43(06):14-20.

[10]肖宇,郑翔文,宋伟,佟凡,毛逸清,刘圣,赵东升.新冠肺炎领域本体构建及应用[J/OL].军事医学:1-6[2021-12-16].http://kns.cnki.net/kcms/detail/11.5950.R.20211029.1155.002.html.

[11]张金福,刘雪.我国高校管理知识图谱构建与应用研究[J].实验室研究与探索,2021,40(09):237-241+276.DOI:10.19927/j.cnki.syyt.2021.09.053.