

基于知识图谱的航天器数字资源建设

陈 恒^{1,2}, 马鑫君¹, 梁艺多¹, 李正光^{1,2}

(1. 大连外国语学院软件学院, 辽宁 大连 116044; 2. 大连外国语学院语言智能研究中心, 辽宁 大连 116044)

摘要: 随着我国航天事业的蓬勃发展, 航天技术已居世界前列。以资源数字化与知识图谱工具的使用为驱动, 实现航天器数据的深度挖掘与展示, 将航天器数据资源构建成可共享的立体数字资源。以公开的航天器数据为研究对象, 采用深度学习模型进行信息抽取, 借助 Protege 构建航天器本体, 结合 Neo4j 数据库构建航天器知识图谱。航天器知识图谱能够提高航天器数据资源的数字化再利用, 同时也能利用航天器知识图谱来对数据进行查询服务, 能够实现未来航天器数据资源的数字化、智能化发展, 进一步推动航天领域的发展和创新。

关键词: 航天器; 本体; 知识图谱

中图分类号: G203 **文献标志码:** A **文章编号:** 1671-1807(2024)02-0042-08

随着中国航天技术的不断发展, 其研制航天器的技术水平也在不断提高。在此基础上, 航天器的数据量也在快速增长。航天企业需要对这些数据进行有效的管理, 以便以后的研究使用。如果在管理过程中没有形成有效的存储与查询, 将导致航天器设计人员很难快速找到有参考价值的数据, 这不利于航天企业的持续性发展。航天需要以服务利用为目的开展数字化资源建设工作, 为航天科研生产活动提供资源保障和服务支撑^[1]。

随着社会的信息化程度提高, 知识图谱在企业知识管理中有着举足轻重的地位。知识图谱是通过关系边来连接各个实体, 通过可视化技术对知识资源进行描述, 知识图谱的构建和分析可以让各实体之间的联系变得更加紧密, 从而形成一个信息网络。这样既可以观察到不同数据之间的联系, 又可以对知识形成有效的管理, 从而让企业资源充分调动起来, 促进企业的持续性发展。

本文提出通过建立航天器知识图谱来提升数据管理效率, 图谱主要对航天器的发射时间、基本参数、设计者等信息进行储存。首先, 对已公开的航天器数据进行搜集, 并对数据进行分析处理, 再通过 Protege 软件实现领域本体的可视化构建, 最后再将构建好的本体导入 Neo4j 数据库中完成航天器知识图谱的构建。以此来实现对航天器数据的

有效管理和查询, 也为航天领域知识管理提供新思路。

1 相关工作

1.1 本体构建

本体可以看作一种对领域知识进行描述的概念化模型, 通过构建本体可以将知识进行细粒度划分。通过构建航天器本体模型使其知识的形式化表达更准确, 同时也使航天器的知识结构更加清晰, 从而为构建航天器知识图谱打下坚实的基础。

本体的构建也被应用到各个领域。例如, Wu 等^[2]提出气候分析本体论来模拟遥感分析师使用的气候数据集。并利用美国海洋大气管理局公布的数据, 进一步探索如何使用本体建模来促进气候数据处理领域; Jarrar^[3]基于本体建立阿拉伯语义网, 进行了全面的评估, 证明该本体可以涵盖大多数阿拉伯语义。何春雨和滕春娥^[4]运用七步法构建了赫哲族非遗资源本体, 为非物质文化遗产知识建模提供了思路。陈琦等^[5]通过分析中医临床试验文献的结构和语言特点, 确定了领域概念和属性关系, 并将文献质量评价信息映射到本体中, 构建了中医临床试验文献知识本体, 基本实现了临床试验文献的结构化表达为后续证据强度及诊疗信息推理的实现奠定了良好的基础。

收稿日期: 2023-10-10

基金项目: 辽宁省应用基础研究计划(2022JH2/101300270); 大连外国语学院科研基金(2021XJYB16)

作者简介: 陈恒(1982—), 男, 安徽阜阳人, 副教授, 博士, 研究方向为知识图谱; 通信作者马鑫君(1999—), 男, 新疆乌鲁木齐人, 硕士研究生, 研究方向为知识图谱; 梁艺多(1985—), 男, 辽宁盘锦人, 副教授, 博士, 研究方向为开放数据; 李正光(1980—), 男, 四川资阳人, 副教授, 博士, 研究方向为自然语言处理。

1.2 知识图谱构建

知识图谱的概念是谷歌于2012年首次提出的,并在信息搜索中得到应用。知识图谱是一种以实体与关系为基础的图库,其实质是一个大型的语义网络图^[6]。同样,知识图谱的应用也被推广到各个领域。例如,Xu等^[7]提出了PubMed数据库构建了医学领域知识图谱;Kim等^[8]基于深度学习模型构建了新型冠状病毒知识图谱。李书钦以科研人员、项目、论文、专著、获奖、专利和标准等为核心要素,建立高校科研管理知识图谱,并通过可视化方式展示高校科研管理数据,为科研管理的科学决策提供参考^[9];彭晖^[10]探讨如何利用史馆已有的档案条目数据和第三方在线数据构建明清专题档案的知识图谱及系统平台。通过知识图谱的建模和系统平台的设计,可以更好地满足用户对明清专题档案的需求;赵雪芹等^[11]将Protege工具与Neo4j相结合,构建了基于万里茶道数字资源知识图谱;佟淑玲等^[12]通过领域分析和数据收集,设计本体模型,通过处理将实例数据导入到知识图谱中,完成声像档案知识图谱的构建,对于有效管理和利用声像档案数据具有重要意义。

综上所述,本体和知识图谱在多个领域已经取得了显著的应用成果,然而在航天器领域的应用却相对较少。在航天器数据数字化管理方面,本体模型可以为其设计顶层结构,通过设计本体

模型,可以对航天器的数据进行分类和组织,建立起一个清晰而有层次的数据结构,为航天器数据的数字化管理提供方向。另外,知识图谱的应用可以将航天器的数据以可视化的形式展示出来。通过将数据进行关联和可视化呈现,可以更加直观地展示航天器的相关信息,为航天企业进行数据挖掘和智能决策提供基础。因此,本文的研究具有较强的现实意义,主要对知识获取、知识抽取、知识建模、知识储存、知识应用方面进行详细论述(图1)。

2 航天器知识图谱构建过程

2.1 数据搜集与处理

选取百度百科和中国航天科普网上的航天器数据作为研究实证,并选取Python作为爬取工具。爬虫主要是将请求通过HTTP发送到网站,然后等待服务器的回应。当服务器做出回应后会得到一个Response,之后需要对Response进行解析,最后将数据储存起来。本文主要是对网站中的xpath标签进行爬取,共爬取数据732条,对爬取数据进行去重去除异常值后进行保存。

2.2 本体构建方法

当前,建立本体的方式有很多种,典型的有骨架法、IDEF5法、七步法等^[13]。七步法主要用于构建领域本体,因此本文选择七步法来构造知识本体。具体构建流程如图2所示。

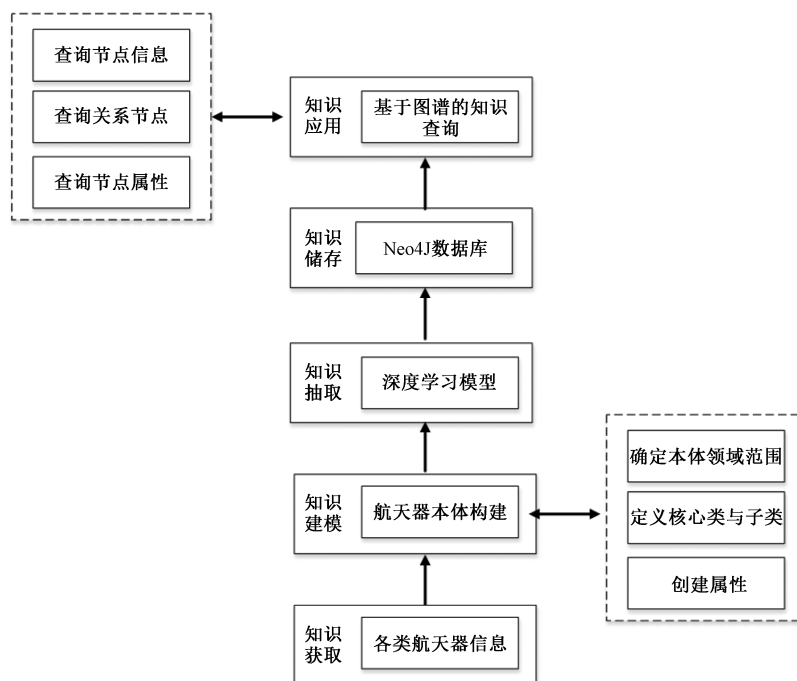


图1 航天器知识图谱构建流程

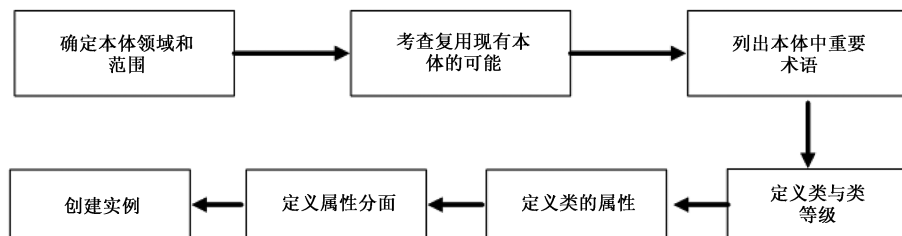


图2 本体构建流程

接下来主要对确定本体领域和范围、定义类与类等级、定义类的属性、创建实例四个方面进行详细论述。

常用的本体建模工具有 Apollo、Ontostudio、TopBraid、Protege 等。Protege 是一款开源的本体编辑器,可以自由下载,而且 Protege 还支持多种继承,可以自定义文件的保存格式,包括 XML、RDF 等。故选择 Protege 作为建模工具。

2.3 确定领域范围及核心类

在构建本体模型的过程中,划定本体领域的范围是至关重要的步骤。这样能提升本体模型在解决特定问题时的有效性和专业性,从而实现更良好的领域适应性。航天企业内部拥有丰富的知识资源,这对航天器的研制发挥着重要作用,如何对航天数据形成有效的管理也是目前需要解决的问题。本文的本体构建就是在此问题的基础上进行的,主要以公开的航天器数据为范围进行本体构建,构建出来的本体可以解决两方面问题。一方面,我国航天事业正处在上升时期,航天器型号更新换代较快,出色的航天设计师层出不穷,对于航天器数据来说,用纸张或者表格来记录不容易体现数据之间的联系,无法对知识形成有效的储存,从而可能会影响航天企业内部知识资源效能的发挥。在引入本体的概念后,能直观看见数据之间的联系,从而对知识形成有效的管理。另一方面,平时航天企业在知识管理过程中,对于知识管理的精度不高,在本体构建的过程中,通过设定类与子类,将知识进行细分,从而使知识管理的精度有所提升,达到资源的有序组织和整合,实现资源利用最大化。

在确定了本体领域范围后,需要对核心类和子类进行定义。航天器的种类有很多,往往一个大类下还有许多小类,那么在对类进行定义时,可以将大类作为核心类名,小类作为子类名,如卫星这个大类下有定位卫星、气象卫星等子类。针对这一特点,确定载人航天器、运载火箭、空间站、卫星、空间探测器、人物等核心类;构建对地观测卫星、通信广播卫星、火星探

测器、总设计师、航天员等子类。之后,用本体建模工具 Protege 进行类的创建,如图3所示。

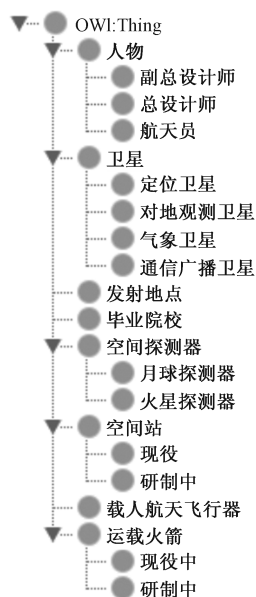


图3 航天器中心类

2.4 属性的创建

在本体构建的过程中,属性的设计是关键。属性设计需遵循以下原则:第一,属性能清楚地表现出类之间的关系;第二,属性应是可传递的,也就是说,该子类可以继承其父类的全部属性;第三,应根据实际问题需求去设计属性,而不是盲目地去覆盖所有属性。

本体的属性分为对象属性和数据属性。对象属性是描述类与类之间的关系,比如,在本研究中,总设计师与空间站这两个类之间的对象属性为设计。与对象属性不同,数据属性是描述类的基本特征,比如发射时间是卫星类的数据属性。利用 Protege 完成对象属性和数据属性的创建。创建好的本体属性如图4所示。

2.5 实体抽取及实例导入

实体抽取是构建知识图谱的基础性工作^[14]。通过爬取得来的数据大多以半结构化为主,需要对

数据进行实体抽取,为知识图谱数据层构建打好基础。而航天器数据通常都是长短不一,没有规律可循,需要使用深度学习模型进行实体抽取。

采用 BERT-BiLSTM-CRF 模型识别实体。模型包括 BERT 预训练语言层、BiLSTM 层和 CRF 层三个部分,总体结构如图 5 所示。



图 4 本体的对象属性与数据属性

BERT-BiLSTM-CRF 模型的主要思想是利用 BERT 模型作为特征提取器,将输入序列转换为词向量表示。这些词向量通过双向 LSTM 进行序列建模,从而捕捉上下文信息。然后,通过条件随机场对序列标注结果进行全局优化,并输出最终的标签序列。

在使用模型抽取前,使用 BIOES 的标注方式对数据进行标注,将标注好的数据按照 7:3 的比例划分为训练集和测试集。采用三种衡量标准, P (准确率)、 R (召回率)、 F_1 值三种,计算方法如式(1)~式(3)所示。

$$P = \frac{\text{识别出正确的实体数量}}{\text{识别出所有的实体数量}} \quad (1)$$

$$R = \frac{\text{识别出正确的实体数量}}{\text{标注所有的实体数量}} \quad (2)$$

$$F_1 = \frac{2PR}{P+R} \quad (3)$$

共标注实体 2 103 个,模型共抽取出实体 2 084 个,抽取正确的实体数量为 1 754 个,经计算 $P=0.8416$ 、 $R=0.834$ 、 $F_1=0.8377$,抽取效果良好。

完成知识抽取工作后,根据之前创建好的类与属性依次导入卫星、人物、运载火箭、探测器等实例

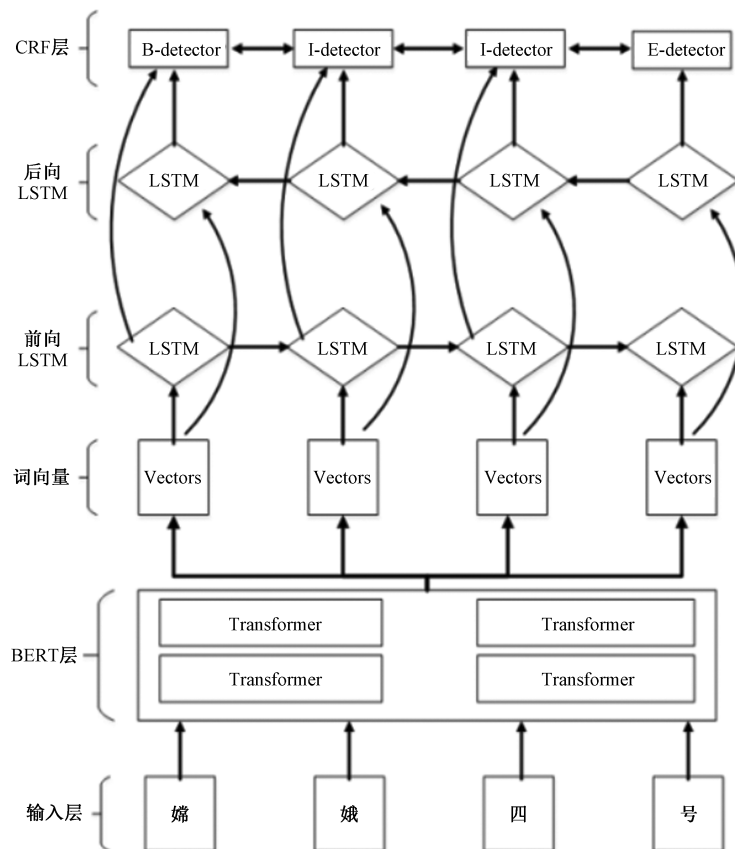


图 5 BERT-BiLSTM-CRF 模型示意图

数据,形成具体的领域应用本体。为进一步表达实例之间的关系,将构建好的本体模型保存为 RDF 格式,为下一步研究做好准备。

2.6 映射匹配与知识储存

Neo4j 作为图形数据库的典型代表,可高效地完成对知识图谱的储存、更新、管理与检索^[15],因此本文选用 Neo4j 存储航天器知识图谱。通过使用 Protege 建立了本体模型,并将模型保存为 RDF 格式。下一步将模型导入 Neo4j 中,完成知识图谱的构建。但是由于 Neo4j 与本体的知识表示在数据结构上有所不同,因而若要将本体模型进行导入,需要设计两种数据结构之间的映射规则,这样能最大限度保证本体模型的完整性,也有利于后期图谱的可视化展示。

本体模型的数据类型有类、对象属性、数据属性、实例,而在 Neo4j 中数据类型有节点、关系、属性。因为是要将数据导入到 Neo4j 当中,所以应当以 Neo4j 的数据类型为标准。在此基础上,制定映射规则。

Neo4j 中的节点代表一个具体的实例,这与本体中的实例表示相类似。故可以将本体中的实例映射为 Neo4j 中的节点。例如,在本体中月球探测器的某一个实例为嫦娥五号,映射到 Neo4j 中,嫦娥五号就变成一个独立的节点。本体中的类名与 Neo4j 中节点的标签名类似,可以将类名映射为标签名,如嫦娥四号、嫦娥五号都属于子类月球探测器,映射到 Neo4j 中嫦娥四号、嫦娥五号的标签名都为月球探测器。特别要注意的是,为了保证数据在 Neo4j 中的细粒度划分,这里只将子类名映射为标签名,这样也可以保证知识的精细度。具体的节点信息如表 1 所示。

表 1 航天器知识图谱节点信息

节点类型	举例
总设计师	叶培建
副总设计师	于登云
航天员	杨利伟
定位卫星	北斗一号卫星
对地观测卫星	资源一号卫星
气象卫星	风云一号气象卫星
通信广播卫星	东方红二号卫星
发射地点	中国文昌航天发射场
毕业院校	北京航空航天大学
月球探测器	嫦娥一号
火星探测器	萤火一号
空间站	天宫一号
载人航天飞行器	神舟五号
运载火箭	长征一号

Neo4j 中的关系主要是表达节点与节点之间的关系,对应到本体中对象属性也是表示类与类之间的关系。由此可以将本体中的对象属性映射为 Neo4j 中的关系类型。在上一条规则中,已经将子类名转化为标签名,但 Neo4j 中标签与标签之间并没有建立关系,所以这条规则只适用于本体中实例之间的对象属性的映射。具体的关系信息如表 2 所示。

表 2 航天器知识图谱关系信息

关系类型	举例
设计	＜叶培建,设计,嫦娥一号＞
发射于	＜嫦娥五号,发射于,中国文昌航天发射场＞
运载	＜长征三号甲运载火箭,运载,嫦娥一号＞
乘坐	＜张晓光,乘坐,神舟十号飞船＞
上一代	＜嫦娥二号,上一代,嫦娥一号＞
毕业于	＜叶培建,毕业于,浙江大学＞

Neo4j 中的属性主要是描述节点的相关信息,对应到本体中,数据属性也是对类的相关信息进行描述,所以可以将本体中的数据属性映射为 Neo4j 的属性信息。同上可得,只将本体中的实例数据属性映射为 Neo4j 的属性。

Neo4j 本身并不支持 RDF 数据格式,因此无法直接将其导入 Neo4j 中。需要安装插件并在 Neo4j 运行输入相对应的语句,就能完成数据导入操作,完成知识图谱的构建。构建好的部分知识图谱结构部分如图 6 所示,共构建节点 896 个、关系 2 094 条、属性 858 个。

3 基于航天器知识图谱的应用探讨

3.1 航天器知识图谱为航天器电子档案顶层设计打好基础

目前,航天企业在数字档案资源建设方面存在一些问题。首先,缺乏顶层设计与统筹规划,以及科学、系统的建设思路和管理对策,导致一些有价值且具备经济和社会效益的航天数字资源可能会被遗漏或淹没在大量的信息数据中。这导致科研人员和公众无法及时了解和利用这些资源。此外,航天企业在电子文件归档、元数据管理和传统载体档案数字化等方面尚未制定统一的标准,给数字档案资源建设带来了一定的盲目性^[1]。而本文构建的航天器知识图谱能在一定程度上解决这些问题,在构建该领域知识图谱前,会先设计该领域的本体模型,航天器本体模型构建可以为航天企业的数字档案管理提供顶层设计和统筹规划。通过定义和组织航天器知识的重要概念、属性和

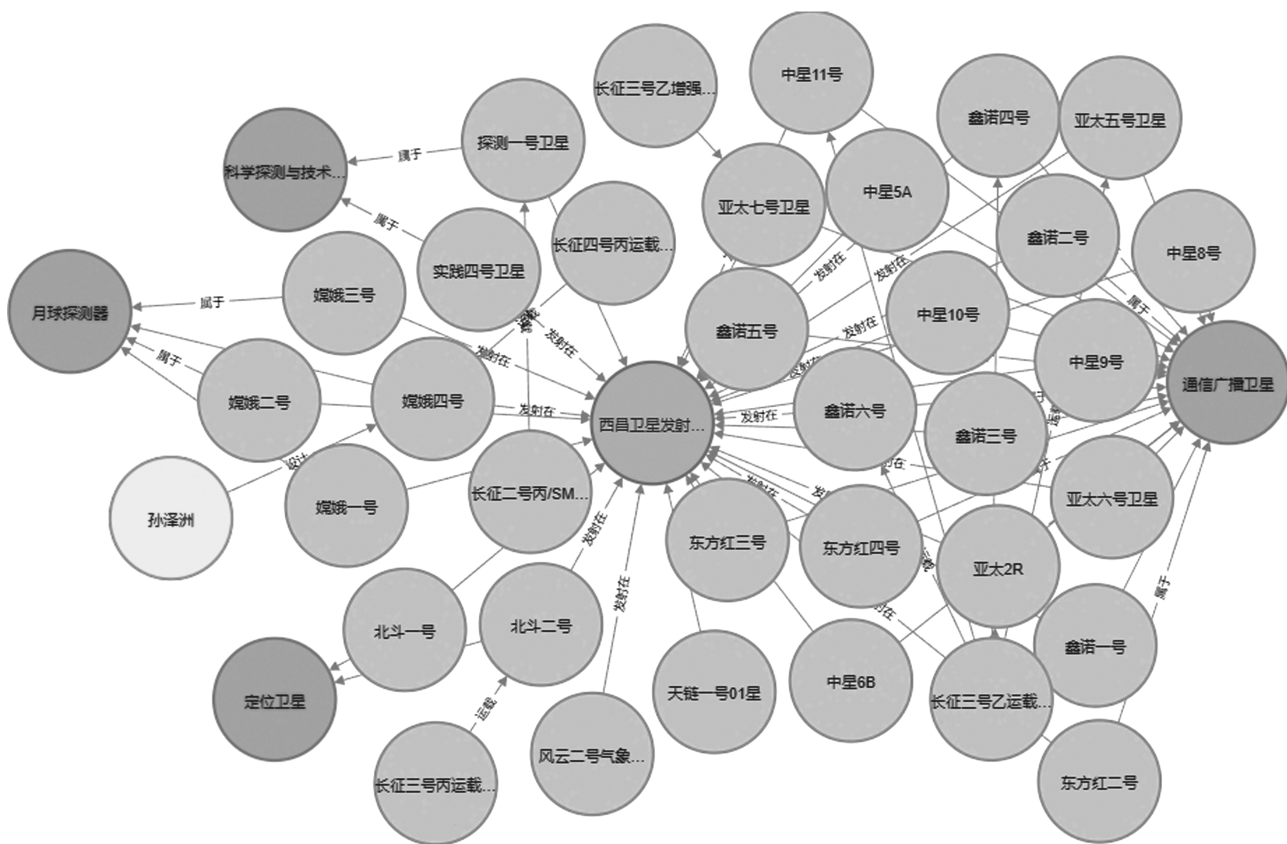


图6 航天器知识图谱(部分)

关系,可以建立一个整体框架,明确数字档案的组织结构和内容,使得数字档案管理更有条理和系统性。本体模型和知识图谱的建立也可以为航天企业提供科学、系统的数字档案资源建设思路。通过对航天器知识的整合和组织,可以识别和填补航天器数字档案资源的空白点,提升航天企业数字档案的完整性和覆盖范围。航天器本体模型为企业在电子文件归档时制定统一标准,为后续数字化档案管理打下基础。

3.2 推进航天企业智能化研究进程

3.2.1 知识查询

航天器知识图谱的构成可以使航天企业更加清晰地看到数据之间的关联,实现对知识的有效管理。同时,在 Neo4j 数据库中,使用者也可以通过使用 Cypher 语言实现对数据的查询服务,分别可以对图谱中的节点、关系、属性进行查询。在图谱中可以根据标签名来搜索标签名下的所有节点,比如航天企业需要统计目前月球探测器有哪些型号,可以在 Neo4j 输入:“match(n:月球探测器) return n”就会显示出符合条件的所有节点:嫦娥一号、嫦娥二号、嫦娥三号等。图谱也支持根据已知节点信息

和关系类型去查找关系边的另一节点的信息。比如航天企业想要设计新型探测器,想查询之前某一探测器的总设计师是谁,则可以在 Neo4j 中输入:“match(n:总设计师)-[r:设计]-(m:月球探测器) where m.Name = ‘嫦娥四号’ return n”,就会显示出符合条件的节点:孙泽洲。在 Neo4j 中,也支持去查询节点的某个属性,比如航天企业想设计新型载人航天飞船,需要参考神舟十四号飞船的总高度,则可以在 Neo4j 中输入:“match(n:载人航天器 {Name: ‘神舟五号’}) return n”总高度,就可以得到结果:9 m。图谱的查询可以让航天企业在管理数据时更加高效,对推动航天企业的数据管理有着积极作用,使查询数据变得更加灵活,图谱的构建也使得数据之间的关系更加清晰,对航天企业的持续性发展有着积极作用。

3.2.2 智能化决策

基于航天器档案知识图谱的信息,可以开发智能推荐系统和决策支持系统。通过分析航天企业的设计需求,智能推荐系统可以推荐相关的航天器档案信息,帮助用户快速获取所需的数据。决策支持系统可以利用知识图谱中的关系和属性信息,为

航天企业的决策制定提供数据支持和决策建议。同时,航天器档案知识图谱也可以作为一个共享平台,促进决策协同和知识共享。不同部门和团队可以共享和访问同一个知识图谱,通过知识图谱中的数据和信息,实现决策协同和共同决策。这样,航天企业可以更加快速、准确地做出智能化决策。

3.2.3 知识发现

通过航天器档案知识图谱,可以对大量的航天器数据进行挖掘和分析。知识图谱中的关系和属性信息可以帮助识别数据之间的关联性和潜在的模式。这样,航天企业可以利用知识图谱中的数据,进行数据挖掘和分析,从中发现隐藏的知识和规律。

3.3 航天器知识图谱有助于完善航天企业信息化系统建设

在航天企业中,电子文件管理体系的应用相对滞后,目前的电子文件管理体系的覆盖面很小,总体上的使用程度不高,无法支持大量文件数据的管理需求。当前,大多数航天企业的档案部门,其档案业务仍处在信息化的初期,与各个业务系统的整合程度不高,还没有形成一个能够涵盖整个业务过程的数字档案综合管理平台,无法达到数据的贯通。

随着信息化水平的不断提高,信息的生成、管理、储存、使用等都呈现一种数字化、网络化、智能化的趋势。知识图谱技术的开发,为航天公司的数据存储、分析和处理提供了一种新的方法。在长期的航天器研发过程中,航天企业已经积累了大量的成功经验以及与之相关的数据,建立航天器知识图谱能够更有效地挖掘数据之间的关联,为企业经营和管理提供更多的有效信息,从而指导企业制定更加科学的管理决策,促进企业的发展。另外,通过使用知识图谱技术,航天企业可以构建一套完善的知识管理系统,使其能够自动地进行知识的检索,从而提高企业的管理水平,增强企业的竞争能力。

4 总结

本文提出了航天器知识图谱构建方法,以公开的航天器数据为基础,构建了航天器知识图谱。这为学者研究相关领域知识表示提供参考,航天器知识图谱通过可视化展示可以看出数据之间的联系,使数据利用率提高,同时对图谱的查询操作,可以

使航天企业的工作效率提高。

在航天器的研制过程中,采用数字化技术对数据进行管理已成为一种趋势。但是,随着型号开发对数字化技术提出的新要求越来越多,数字化技术自身又处在一个高速发展期,这就导致了数字化技术的使用是一个边规划、边实施、边修改的过程,所以航天器知识图谱的构建应该紧密围绕航天产品的研制而同步进行、同步发展。

参考文献

- [1] 田红,何正文,杨琴茹. 航天企业数字档案资源建设的问题与对策研究[J]. 北京档案, 2020(12): 25-27.
- [2] WU J, ORLANDI F, O'SULLIVAN D, et al. An ontology model for climatic data analysis[C]//2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Brussels: IEEE, 2021: 5739-5742.
- [3] JARRAR M. The Arabic ontology-An Arabic wordnet with ontologically clean content[J]. Applied Ontology, 2021, 16(1): 1-26.
- [4] 何春雨,滕春娥. 非物质文化遗产知识本体构建——以赫哲族非遗资源为例[J]. 情报科学, 2021, 39(4): 140-147.
- [5] 陈琦,李芹,张君冬,等. 中医临床试验文献知识本体的构建[J]. 中国实验方剂学杂志, 2023, 29(20): 190-197.
- [6] SINGHAL A. Introducing the knowledge graph: things, not strings[J]. Official Google Blog, 2012, 5: 16.
- [7] XU J, KIM S, SONG M, et al. Building a PubMed knowledge graph[J]. Scientific Data, 2020, 7(1): 1-15.
- [8] KIM T, YUN Y, KIM N. Deep learning-based knowledge graph generation for COVID-19[J]. Sustainability, 2021, 13(4): 2276.
- [9] 李书钦. 面向高校科研管理的知识图谱构建研究[J]. 科技和产业, 2020, 20(12): 107-110.
- [10] 彭晖. 明清档案知识图谱的构建探索[J]. 历史档案, 2022(3): 134-141.
- [11] 赵雪芹,李天娥,曾刚. 基于 Neo4j 的万里茶道数字资源知识图谱构建研究[J]. 情报资料工作, 2022, 43(5): 89-97.
- [12] 佟淑玲,王越文,李泽坤. 基于本体的声像档案知识图谱构建研究[J]. 档案管理, 2022(6): 52-56.
- [13] 胡汗林,邓三鸿. 知识图谱在青铜器数字馆藏建设中的应用[J]. 数字图书馆论坛, 2023, 19(4): 1-8.
- [14] 杨媛媛. 领域知识图谱的基本概念与构建特点[J]. 辽宁工业大学学报(社会科学版), 2022, 24(3): 57-61.
- [15] 胡卫,赵文龙,李石磊,等. 军事装备管理数据知识图谱构建及应用[J]. 火力与指挥控制, 2022, 47(10): 125-131.

Construction of Spacecraft Digital Resources Based on Knowledge Graph

CHEN Heng^{1,2}, MA Xinjun¹, LIANG Yiduo¹, LI Zhengguang^{1,2}

(1. Software School of Dalian University of Foreign Languages, Dalian 116044, Liaoning, China;

2. Center for Language Intelligence of Dalian University of Foreign Languages, Dalian 116044, Liaoning, China)

Abstract: China's space industry is thriving and space technology is world class. Driven by digitization of resources and use of knowledge mapping tools, deep mining and demonstration of spacecraft data was carried out, and spacecraft data resources were constructed into shared stereoscopic digital resources. Taking open spacecraft data as the research object, using deep learning model to extract information, using Protege to construct spacecraft ontology, combining Neo4j database to construct spacecraft knowledge map. The spacecraft knowledge graph can improve the digital reuse of spacecraft data resources, at the same time, it can also use the spacecraft knowledge graph to query the data, which can realize the digital and intelligent development of spacecraft data resources in the future, and further promote the development and innovation of the aerospace field.

Keywords: public data; spacecraft; ontology; knowledge graph