

深空探测领域知识库构建与有效载荷配置

阳松^{1,2,3} 牛文龙^{1,2} 彭晓东^{1,3}
李思奇¹ 刘金诚¹ 杨震¹

1(中国科学院国家空间科学中心 北京 100190)

2(中国科学院大学 北京 100049)

3(国科大杭州高等研究院基础物理与数学科学学院 杭州 310024)

摘要 根据国家在深空探测领域任务规划论证战略需求, 针对深空探测领域任务种类繁多、数量巨大、系统复杂, 数据难以直接利用等问题, 提出一种深空探测领域知识库构建方法. 通过对深空探测领域知识分析并结合专家知识自顶向下构建本体模型, 在此基础上提出一种基于 N-gram 和文本语义聚类分析的本体更新策略, 对深空探测领域文本进行语义识别处理, 自底向上动态更新本体; 整合已有结构化、半结构化和非结构化数据, 构建深空探测领域知识库, 共包含实体节点 2140 个, 关系 4608 条, 属性 9079 条; 提出一种基于深空探测领域知识分析的有效载荷配置方法, 对知识库中的科学目标等信息进行关联分析, 将知识抽取模型与语义相似度模型相结合, 综合探测对象与历史有效载荷信息, 推荐出满足科学目标的有效载荷配置, 并利用天问一号任务对所提出的方法进行验证.

关键词 深空探测, 本体, 知识库, 知识抽取, 有效载荷

中图分类号 TP3

Knowledge Base Construction and Payload Configuration in Deep Space Exploration Field

YANG Song^{1,2,3} NIU Wenlong^{1,2} PENG Xiaodong^{1,3}
LI Siqi¹ LIU Jincheng¹ YANG Zhen¹

1(National Space Science Center, Chinese Academy of Sciences, Beijing 100190)

2(University of Chinese Academy of Sciences, Beijing 100049)

3(Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024)

Abstract A method is proposed to address the strategic needs of national mission planning and demonstration in the field of deep space exploration, focusing on challenges such as the diversity and

2024-03-28 收到原稿, 2024-05-10 收到修定稿

©The Author(s) 2025. This is an open access article under the CC-BY 4.0 License
(<https://creativecommons.org/licenses/by/4.0/>)

complexity of missions, the large scale of data, and difficulties in direct data utilization. The construction of an ontology model for the deep space exploration domain is carried out using a hybrid approach that combines top-down and bottom-up methodologies. In the top-down process, domain knowledge is analyzed and expert input is incorporated to build the ontology model. On this basis, a bottom-up ontology updating strategy is introduced, leveraging N-gram analysis and text semantic clustering to dynamically update the ontology through semantic recognition and processing of textual data related to deep space exploration. Structured, semi-structured, and unstructured data are integrated to build a comprehensive knowledge base for the domain, comprising 2140 entity nodes, 4608 relationships, and 9079 attributes. Based on this knowledge base, a payload configuration method is proposed to recommend payloads that meet specific scientific goals. This is achieved through correlation analysis of scientific target data within the knowledge base, combining a knowledge extraction model with a semantic similarity model to synthesize exploration object information and historical payload data. The effectiveness of this approach is validated using the Tianwen-1 mission, demonstrating its capability in supporting payload planning, scientific target analysis, and mission configuration in the field of deep space exploration.

Key words Deep space exploration, Ontology, Knowledge base, Knowledge extraction, Payload

0 引言

根据 2000 年发布的《中国的航天》白皮书中的定义, 深空探测是对地球以外天体开展的空间探测活动^[1]. 深空探测是国家综合国力和创新能力的重要标志, 由科学目标引领的可持续深空探测将成为未来深空科技发展的重要特征^[2].

自 1958 年以来, 人类已开展深空探测任务 260 余次, 探测对象覆盖了太阳系内月球、行星、彗星、太阳等不同类型天体^[3]. 为推动深空探测领域科学探索的前沿突破以及国际合作与引领, 有必要全方位研究梳理深空探测领域科学主题、探测任务、有效载荷及技术、科学成果等进展, 构建知识与决策支撑平台, 支持决策者与实施者把握科学探索全貌、最新进展与未来趋势, 促进全球科学家协同推动前沿研究.

在过去的几十年中, 深空探测任务中的载荷方案等任务主要通过跨学科的团队合作、深入研讨和理论论证得以实现. 然而不同领域的专家可能使用不同的术语体系和方法理论, 这可能导致沟通壁垒, 使得在跨学科团队中, 理解与协调不同学科之间的需求和约束变得复杂化. 随着人工智能技术的迅速发展, 各种自然语言处理技术在知识整合和分析总结方面取得了显著进展. 当前, 知识库技术被应用于各个领域, 知识库的概念来自两个不同的领域, 一是融合了人工智能及其分支领域的知识工程, 二是借鉴了传统数据库领域的理念. 知识库系统的形成和发展是人工智能和

数据库领域相互融合的产物^[4].

随着未来科技的不断发展, 各国都将陆续开展各种深空探测计划, 届时该领域的信息化、数字化进程会不断加快. 面对不断增长的海量相关数据, 高效的数据处理与分析将成为关键, 为深空探测任务的规划设计、关键指标的智能遴选以及辅助评估等方面提供强有力的支持, 进一步促进深空探测领域的创新发展^[5]. 深空探测领域作为知识库技术的应用领域之一, 面对可能出现的复杂数据, 快速地从不同类型文本以及相关文档中挖掘出有价值的信息, 进而为专家辅助决策提供有力的支撑, 具有广阔的应用前景.

深空探测知识库构建是一个全新的领域, 而其他领域的知识库均有发展, 例如综合百科知识库 YAGO, Wikidata^[6] 以及中国法律领域的知识库 CNKI 法库、临床诊疗知识库等. 在航天相关领域, 已有航天术语知识库 ATHowNet^[7]. 然而由于深空探测领域的特殊性以及其知识覆盖的相对匮乏, 现有方法并不完全适用于深空探测领域的知识库构建. 如果能够快速整合现有深空探测领域相关数据, 建立一个囊括各个国家相关计划方案的深空探测知识库以及知识点的深层语义关系, 有助于拓展未来相关领域的知识覆盖面, 能通过更加直观、有条理的知识点来科学、高效地展示深空探测领域各个目标计划, 为未来中国的深空探测技术发展提供辅助支撑. 知识为规范化的数据, 知识之间关系的交织成图, 利用图类型的数据结构对深空探测领域相关事务进行建模, 能够推断、

识别事务之间的领域知识以及复杂关联性. 因此, 本文以深空探测领域相关数据为基础, 探索深空探测领域本体模型和知识库构建方法, 开展深空探测领域知识的相关应用研究.

在深空探测任务中, 有效载荷配置是其中的关键环节之一. 在工程层面, 有效载荷的配置通常涉及多种因素, 包括科学目标、仪器性能、资源限制等. 在遥感领域, 已提出了一种从探测目标出发的方法, 通过分析满足探测要素所需谱段设置、指标要求和有效载荷配置类型, 得出一套有效的载荷配置方法和思路^[8]. 此外, 还有一些基于任务需求和经验规则的配置策略. 尽管这些方法一定程度上能够满足任务需求, 但是仍存在一些局限性, 例如对于任务变化的适应性不足且人工干预成本高. 本文提出一种深空探测领域知识分析的有效载荷配置方法, 通过对知识库中的科学目标等信息进行关联分析, 结合知识抽取模型与语义相似度模型, 综合探测对象与历史有效载荷信息, 推荐出满足科学目标的有效载荷配置, 充分利用深空探测领域内的丰富知识资源, 为载荷配置提供更为智能化和科学化的解决方案.

1 概述

1.1 深空探测领域知识源组成

深空探测领域知识围绕深空探测任务展开, 包括火星、月球、木星、金星、小行星探测任务等. 数据来源主要为《未来 10 年中国学科发展战略: 空间科学》《

中国航天》等期刊杂志, NASA 和 ESA 等权威网站报告、百科知识详解以及人工整理翻译的权威网站报告等, 由此形成知识表格. 根据数据类型可将其分为结构化、半结构化以及非结构化数据^[9]. 具体数据源组成如图 1 所示.

1.2 深空探测领域知识要素分析

深入研究深空探测领域知识体系, 主要包括实体、属性以及实体之间的关系. 所探讨的知识要素涵盖历史上所有规划、实施的深空探测任务, 以及相关科学知识、技术知识、应用知识和外部知识, 包括深空探测科学问题、天体、探测器、有效载荷、科学目标、工程目标等知识要素, 部分知识要素如图 2 所示.

2 深空探测领域本体构建

2.1 深空探测领域本体总体构建方法

知识库的逻辑架构可划分为本体层和数据层两部分^[9]. 本体层在语义层次上描述知识, 可以视为描述某个学科领域知识的一个通用概念模型^[10]. 数据层则是本体层的实例化, 以“实体—关系—实体”或者“实体—属性项—属性值”^[11]形式的三元组表达深空探测知识, 利用“节点—边”构成深空探测知识网络. 深空探测领域本体是包含了深空探测领域关键概念、类别和关系的抽象模型, 旨在以系统性和一致性的方式描述深空探测任务相关知识.

深空探测领域本体采用自顶向下和自底向上融合构建的方式, 具体步骤如图 3 所示.

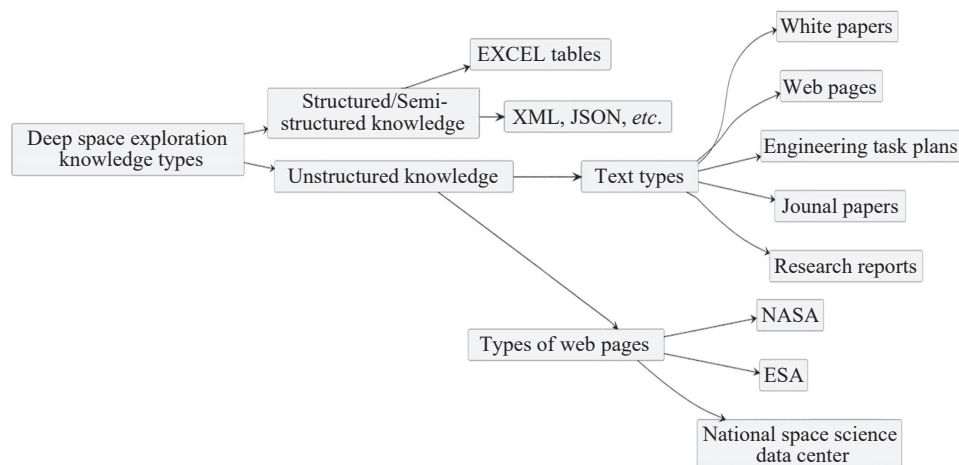


图 1 深空探测领域数据源组成

Fig. 1 Composition of data sources in deep space exploration field

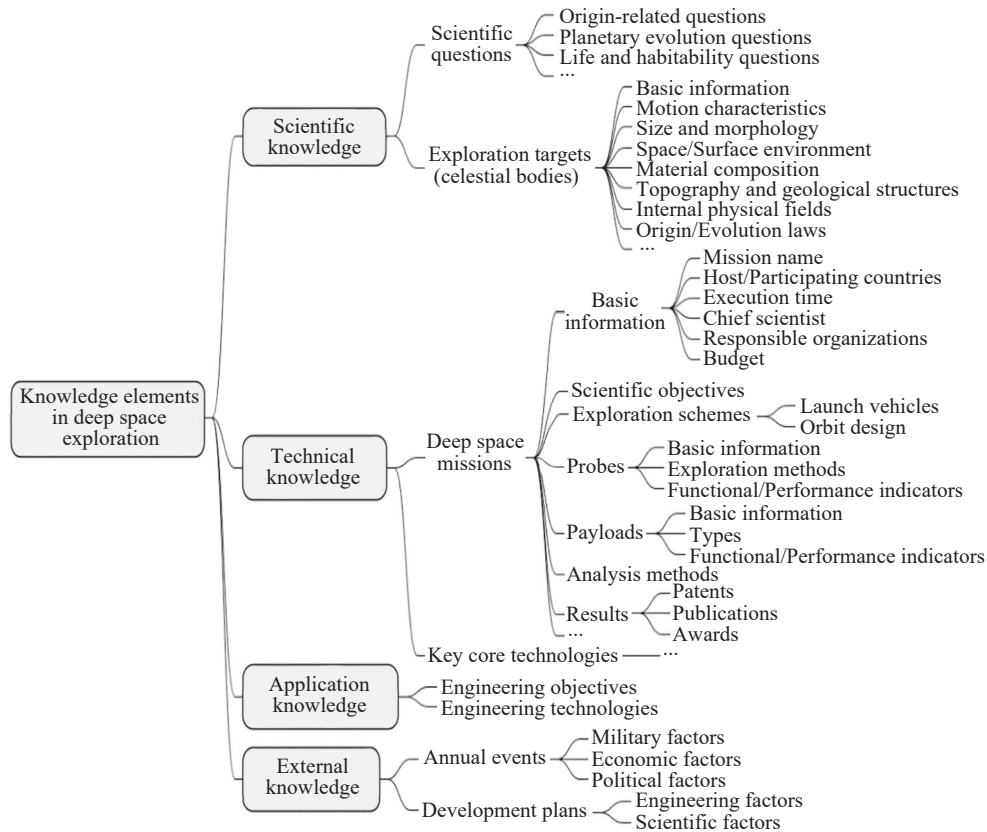


图 2 深空探测领域部分知识要素

Fig. 2 Selected elements of knowledge in deep space exploration field

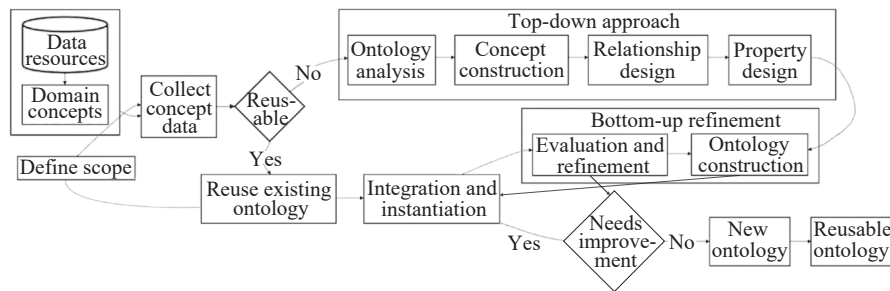


图 3 深空探测领域本体构建流程

Fig. 3 Deep space exploration field ontology construction process

自顶向下的方法是从高层次的概念和抽象开始构建本体架构, 根据深空探测领域知识特点, 建立知识在语义空间中的层次化表达, 具体步骤借鉴斯坦福七步法^[12]构建, 具体如下。

- (1) 确定深空探测领域本体的专业领域和范畴。
- (2) 考查复用现有相关领域本体的可能性。
- (3) 列出深空探测领域本体中的重要术语。
- (4) 定义深空探测领域实体类及其等级体系。

(5) 定义深空探测领域实体类的属性。

(6) 定义深空探测领域实体类的属性的分面。

(7) 创建深空探测领域实例。

自底向上的方法则是在通过自顶向下方式已构建好的本体架构基础上, 在对海量结构化文档、半结构化文档以及非结构化文档进行知识抽取丰富实例层的同时, 继续挖掘已有实例层节点的共性特征, 比较实例层节点之间的共性特征及其在本体要素中的

对应情况, 动态地对本体要素做出相应调整更新^[13], 确保本体架构的准确性和实用性, 丰富整个深空探测领域的本体. 这种融合构建的方式可以促使本体模型在不断迭代和完善中得到进一步的发展和应用. 在此基础上提出一种基于 N-gram^[14] 和文本语义聚类分析的本体动态更新的算法, 通过实时数据整理更新迭代本体.

2.2 深空探测领域本体更新方法

2.2.1 算法设计

深空探测领域的专业术语通常是特定任务、设备或技术的名称, 通过多种词汇搭配组合而成, 例如环绕器次表层探测雷达、火星探测器等, 语句则是专业术语的组合搭配. N-gram 模型通过分析和处理文本中的 n 个连续项 (单词或短语) 序列, 可以有效捕获和理解这些组合术语的结构和意义. 基于 N-gram 和知识聚类分析的本体更新算法流程见表 1.

首先将数据源根据 N-gram 分为不同的数据集, 接着将数据集转化为词向量. 词嵌入选用 bge-large-zh 模型, bge 是当前中文任务下最强语义向量模型, 各项语义表征能力全面超越同类开源模型^[15].

采用 K-Means++^[16] 算法对数据进行聚类分析, K-Means++ 是 K-Means 聚类算法的改进版本, 旨在改善 K-Means 初始聚类中心点的选择. 其具体步骤如下.

(1) 从数据集中随机选择 1 个数据点作为第 1 个

簇中心 C_1 .

(2) 计算每个样本与当前已有聚类中心之间的最短距离, 用 $D(x)$ 表示; 然后计算每个样本点被选为下一个聚类中心的概率 $P(x)$, 最后选择最大概率值 (或者概率分布) 对应的样本点作为下一个簇中心, 有

$$C_1 = \{x_i\}, x_i \in X; \quad (1)$$

$$P(x) = \frac{D(x)^2}{\sum D(x_i)^2}. \quad (2)$$

(3) 重复第 2 步, 直至选出 k 个聚类中心.

采用簇内平方和 (WCSS, 定义数学符号 S_{WCSS})^[17] 与轮廓系数 (Silhouette Score, 定义数学符号 C_{SS})^[18] 双重方法确定聚类分析中的 k 值.

$$S_{WCSS} = \sum_{i=1}^k \sum_{x \in C_i} |x - \mu_i|^2, \quad (3)$$

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}, \quad (4)$$

$$C_{SS} = \frac{1}{n} \sum_{i=1}^k s(x_i). \quad (5)$$

其中, k 为聚类的数量, C_i 为第 i 个聚类中点的集合, μ_i 为簇 C_i 的中心点, $\|x - \mu_i\|$ 为点 x 到质心 μ_i 的欧几里得距离的平方; $a(x)$ 为样本 x 到其自身簇中所有其他点的平均距离, 代表簇内不相似度; $b(x)$ 为样本 x 到最近簇 (样本不属于此簇) 中所有点的平均距离,

表 1 基于 N-gram 和深空探测领域知识聚类分析的本体更新算法
Table 1 Ontology updating algorithm based on N-gram and deep space exploration field knowledge clustering analysis

算法步骤	功能描述
Input: 某类实体或者实体属性数据源	
Output: 可选择更新本体中的实体、关系或者属性	
1. Dtest ← Doriginal, Dunigram, Dbigram, Dtrigram	// 选择需要更新的数据源
2. Xembedding ← Dtest	// 数据源向量化
3. Initialize the number of cluster k	// 初始化聚类 k 值
4. For $i = 2$ to k	
score = max (score, wcss + silhouette score)	// 综合两类得分确定聚类更新 k 值
$i = i + 1$	
6. SVD (XEmbedding, k)	// SVD 降维分析
7. Calculae hot words and identify themes	// 计算热词和确定主题
8. Updating the ontology	// 更新本体

代表簇间不相似度; n 为样本点的总数, $s(x_i)$ 为第 i 个样本点的轮廓数.

2.2.2 实验数据

在进行本体自底向上动态更新验证时, 采用的数据源为 NASA 和 ESA 等网站报告形成的表格、中国相关探测任务文献资料以及相关百科知识, 涉及已收

集整理的有效载荷共 715 条, 载荷功能共 541 条, 功能描述去重后共 383 条. 分别选取原文 (original), unigram, bigram, trigram^[14] 作为相应的数据源, 部分数据示例见表 2 (表中采用 “#” 分隔).

2.2.3 实验结果与分析

基于簇内平方和的聚类结果如图 4 所示, 基于轮

表 2 有效载荷的功能描述文本数据源
Table 2 Functional description of the payload text data source

original	unigram	bigram	trigram
探测宇宙辐射强度及其在地球、太空中随着与太阳的距离增加以及火星周围的变化	探测#宇宙#辐射强度#地球#太空#太阳#距离#增加#火星#周围#变化	探测宇宙#宇宙辐射强度#辐射强度地球#地球太空#太空中#中太阳#太阳距离#距离增加#增加火星#火星周围#周围变化	探测宇宙辐射强度#宇宙辐射强度地球#辐射强度地球太空#地球太空中#太空中太阳#中太阳距离#太阳距离增加#距离增加火星#增加火星周围#火星周围变化
获取火星表面图像	获取#火星#表面#图像	获取火星#火星表面#表面图像	获取火星表面#火星表面图像
分析火星岩石和土壤中含铁矿物质的构成	分析#火星#岩石#土壤#中含铁#矿物质#构成	分析火星#火星岩石#岩石土壤#土壤中#中含铁#含铁矿物质#矿物质构成	分析火星岩石#火星岩石土壤#岩石土壤中#土壤中含铁#中含铁矿物质#含铁矿物质构成
拍摄冥王星及其卫星和柯伊伯带其他天体的图像	拍摄#冥王星#卫星#柯伊伯#带#天体#图像	拍摄冥王星#冥王星卫星#卫星柯伊伯#柯伊伯带#带天体#天体图像	拍摄冥王星卫星#冥王星卫星柯伊伯#卫星柯伊伯带#柯伊伯带天体#带天体图像

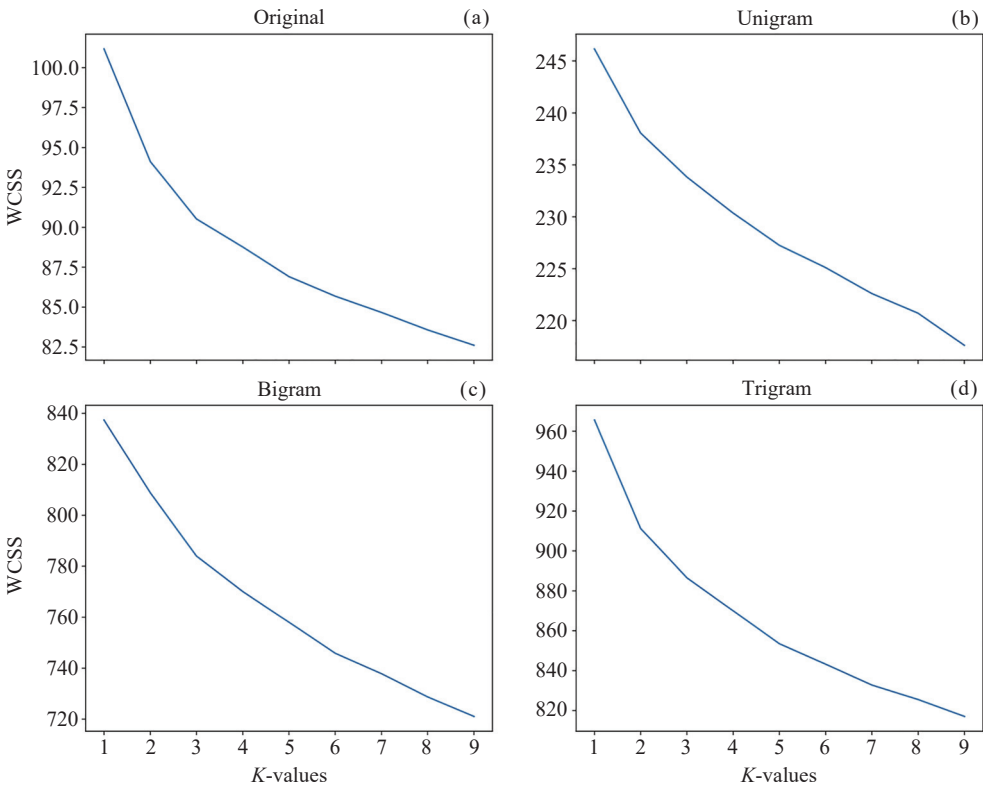


图 4 基于簇内平方和确定最优 K 值
Fig. 4 Determination of optimal K value based on WCSS

廓系数的聚类结果如图 5 所示. 通过簇内平方和观察拐点位置确定 K 值的合理范围, 再结合轮廓系数评估各 K 值下聚类的凝聚度和分离度, 选择轮廓系数最高的 K 值以达到最优聚类效果, 综合得出在 original, unigram, bigram, trigram 下的最优 K 值分别为 2, 2, 3, 2.

三维空间中的可视化有助于直观理解数据分布情况, 包括聚类的形状、大小和相互之间的关系, 这在高维空间中是难以实现的, 采用 SVD^[19] 技术将 1024 维的词向量降维到 3 维表示, 降维后聚类向量空间如图 6 所示.

统计判断各簇内离中心最近的前 50 个点对应的原始数据源分布情况, 采用词云图、词频柱状图可视化形式辅助确认聚类主题, 并确定是否进行本体更新.

当数据源为原文, $K=2$ 时, 其中一个簇相关可视化分析如图 7 所示; 当数据源为 bigram, $K=3$ 时, 其中一个簇相关可视化分析如图 8 所示.

对深空探测载荷功能描述文本进行分析, 根据原文数据情况可以看出, 所收集整理的数据与探测天体

的关系密切, 在初始本体的制定中, 探测载荷只与探测器直接关联, 与探测天体间接关联, 可以选择新增探测载荷与探测天体的直接关联; 根据 bigram 数据情况可以看出, 所收集整理的数据与探测对象的关系密切, 可新增探测对象的概念, 以详细描述深空探测领域中的目标对象, 例如大气、电离层、磁场、等离子体等. 同时新增探测载荷与探测对象的关系. 通过此方法, 可以对领域内新扩充或者补充收集的数据进行实时分析, 分析实体之间的关联, 牵引出概念之间的联系, 实现动态更新领域本体, 丰富领域内的本体架构.

2.3 深空探测领域本体

从深空探测领域专业知识和公共领域知识两类知识出发, 结合目前规划的应用场景和需求, 包括有效载荷配置、科学目标推荐与遴选应用、科学成果分析等应用, 融合深空探测领域专业术语知识, 自顶向下构建了探测计划、探测任务、探测器、有效载荷、运载火箭、科学方面、工程方面、技术方面等领域专业术语知识, 同时融合公共领域知识, 包括国家、机构、地区、时间、人物等. 以自底向上的方式更新本体, 新

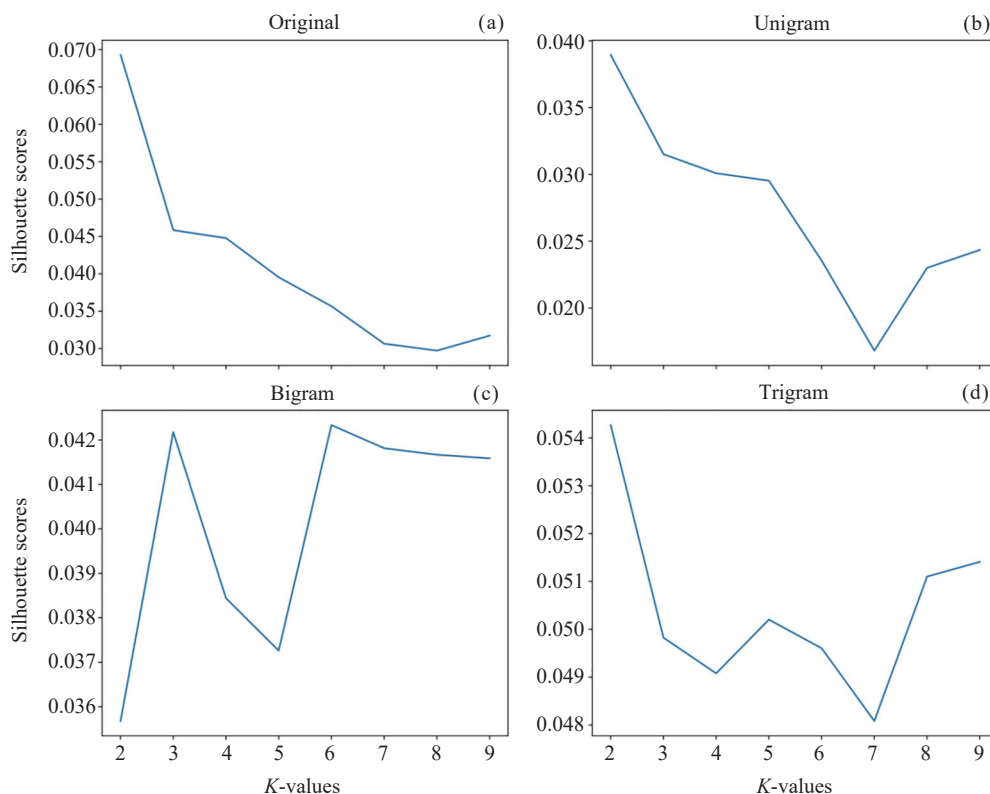


图 5 基于轮廓系数确定最优 K 值

Fig. 5 Determination of optimal K value based on Silhouette Score

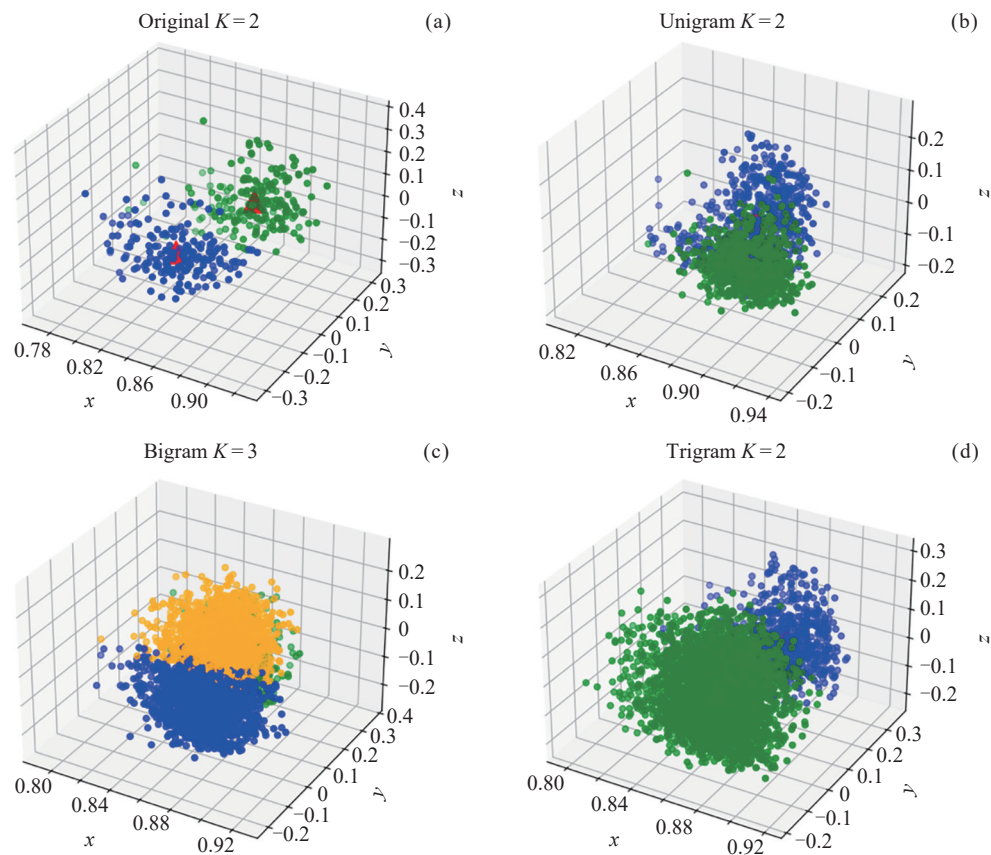


图 6 功能文本描述聚类向量空间

Fig. 6 Functional text description clustering vector space

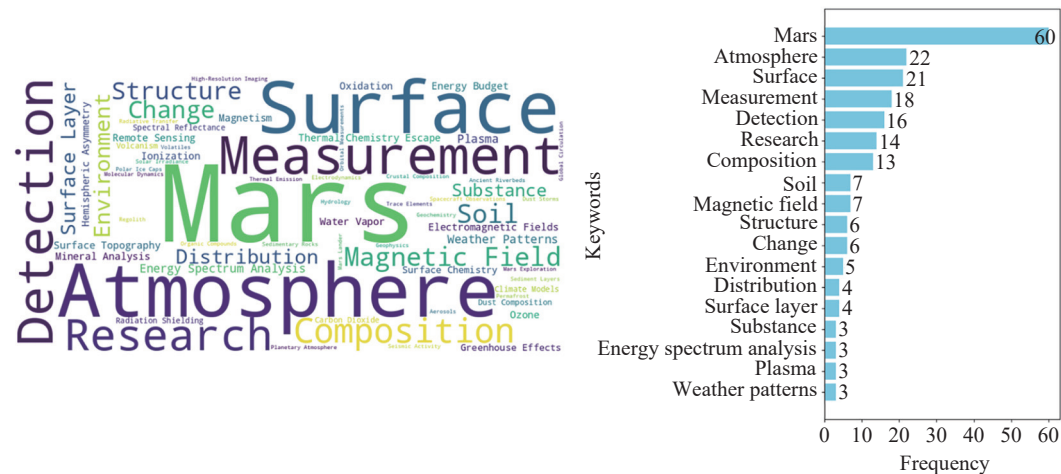


图 7 数据源为原文且 $K=2$ 时的可视化分析结果

Fig. 7 Visualization of the analysis when the data source is the original text and $K=2$

增了探测对象以及探测对象与有效载荷的关联信息, 构建了深空探测领域本体架构. 分别从概念的名称、含义、属性以及概念之间的关系来描述本体架构. 深

空探测领域的具体概念名称列于表 3, 概念之间关系列于表 4. 目前共有 17 类概念、17 类概念间关系以及 90 条属性项.

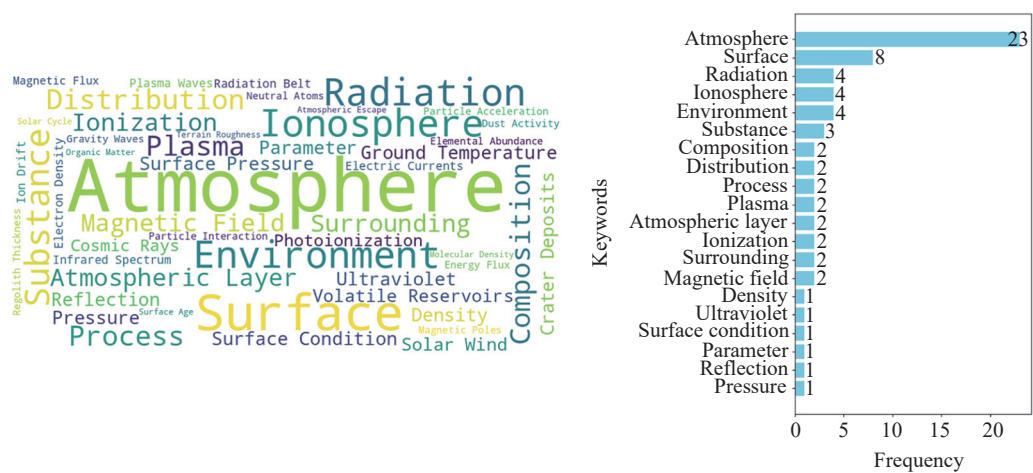


图 8 数据源为 bigram 且 $K=3$ 时的可视化分析结果

Fig. 8 Visualization and analysis when data source is bigram and $K=3$

表 3 深空探测领域概念名称、含义及属性说明

Table 3 Description of concept names, meanings and attributes in the field of deep space exploration

序号	概念名称	描述说明	属性
1	国家	表示各个国家, 用来标识特定国家相关信息	中文名称、英文名称
2	天体	代表宇宙中的天体, 如行星、卫星	中文名称、英文名称、结构、自转轨道、生命潜力、磁层、形状、尺寸、公转轨道、天体类型、基本情况介绍、大气和气候、质量、与太阳的距离
3	科学主题	描述科学研究中的不同主题	中文名称
4	科学问题	涉及科学研究中的具体问题或未解之谜	中文名称
5	科学目标	表明科学研究的具体目标或愿望	中文名称
6	科学要素	科学人员对科学目标关键词的凝练	中文名称
7	工程目标	表示工程项目的具体目标或任务	中文名称
8	探测任务	描述针对宇宙的各种科学探测任务	中文名词、英文名称、任务别名、发射场、当前状态、任务负责人、研制单位、发射质量、发射时间、着陆日期、结束日期、任务来源、寿命、失败类型和具体故障情况、任务简述、首席科学家、任务结果、成本控制措施、分析手段、任务经费、轨道设计、标识符、研发机构
9	探测计划	代表一系列相关的探测任务, 可能隶属于一个国家或机构	中文名称、英文名称
10	探测器	表示空间中的探测器、巡视器等	中文名称、英文名称、主要参数、负责人、研制单位、寿命、发射时间、探测器类型、探测方式、当前状态、性能指标、分析手段、功能指标
11	有效载荷	描述探测器上携带的科学仪器或设备, 用于执行特定的科学任务	中文名称、英文名称、载荷类型、载荷当前状态、对应学科、分析手段、主要参数、研制机构、功能描述
12	运载火箭	搭载探测器的运载火箭	中文名称、英文名称、研发机构、运力、服役情况
13	探测方式	科学探测的不同方法或技术	中文名称、探测类别
14	探测对象	表示探测任务的目标对象, 如火星大气等	中文名称、英文名称
15	区域	特指火星、月球上不同地区, 包括撞击坑、山脉等、经纬度等	中文名称、英文名称、描述
16	关键技术	描述在科学探测任务或工程项目中需使用的关键技术或创新技术	中文名称、技术来源、技术描述、技术挑战、技术指标
17	科学成果	科学研究或探测任务取得的具体成果, 如专利、研究论文和奖项	名称、申请人、文献来源、类型、DOI、公开号、来源库、作者、发表时间、单位

表 4 深空探测领域概念与概念关系
Table 4 Description of concepts and conceptual relationships in deep space exploration field

序号	关系名称	主体类型	客体类型
1	属于	探测任务	国家
		探测任务	探测计划
		有效载荷	国家
		有效载荷	探测器
		探测器	探测任务
2	探测	探测任务	天体
		有效载荷	探测对象
		探测器	天体
3	研究	科学主题	科学问题
		科学问题	科学要素
4	科学要素	探测任务	科学要素
5	科学目标	探测任务	科学目标
6	工程目标	探测任务	工程目标
7	运载火箭	探测任务	运载火箭
8	包括	探测任务	探测技术
9	探测方式	探测任务	探测方式
10	着陆区域	探测器	区域

3 深空探测领域知识抽取

3.1 结构化、半结构化知识抽取

结构化、半结构知识抽取专注于从具有相关明确格式和严格组织的数据源中提取信息. 这些数据源通常是数据库和表格, 其中包含行和列以及与之相关的表头信息.

在结构化知识抽取过程中, 表头字段用于确定数据元素与知识本体中相应的概念、属性和关系的映射. 例如, 一个表格的列标题可直接对应于本体中一个属性名称. 在处理结构化数据时, 需要特别注意表格中对同一对象的不同表示方式, 例如, 同一个实体在不同数据中可能会有不同的名称或标符. 为保证数据的一致性和准确性, 需采用唯一的标识符或名称与之对应和映射. 与结构化数据相比, 半结构化数据的格式表现更为灵活但仍然包含一些组织结构, 如 XML, JSON 和 HTML 文档. 提取的信息通常根据其网页上的表现形式 (如表格或段落标题) 而将其转换为键值对 (Key-Value) 形式的 JSON 数据. 这种格式的优点在于其灵活性, 在不牺牲结构化的同时接

受数据的多样性和复杂性. 数据一旦被转换为 JSON 格式, 就可以被进一步处理, 与知识本体中的实体、属性和关系相对应. 这样, 数据就从原始的半结构化形式转化为更加结构化的知识表示. 收集和整理的结构化、半结构化知识来自 NASA 和 ESA 等网站报告形成的表格、百科知识形成的 JSON 数据.

3.2 非结构化知识抽取

非结构化知识抽取是自然语言处理中的一个重要研究方向, 包括实体抽取和关系抽取, 实体抽取是抽取文本中的相关实体及其类型, 关系抽取是抽取文本中的主体 (Subject)、谓语 (Predicate)、宾语 (Object), 即 SPO 三元组方面^[20].

深空探测领域非结构化文本, 例如相关新闻文章、项目文件、书籍和科学文献等, 蕴含着丰富的信息, 但是这些信息的表达形式多样且不规则, 给自动化提取带来了挑战. 通过对工程任务方案、研究报告、网页、期刊文献等数据的整理和优化, 基于构建好的本体架构, 综合构建了部分知识抽取数据集. 目前已设计 10 类深空探测领域实体, 包含 13 类关系. 具体数据集实体和关系分布如图 9 所示, 数据集

Schema 如图 10 所示. 该数据集包含 1490 条实例、10 种实体种类、6755 条实体、13 种关系种类以及 4616 条三元组.

深空探测领域知识抽取模型采用 UIE^[21] 信息抽取模型, UIE 是一种针对非结构化文本的通用抽取模型, 可以针对不同的抽取任务灵活进行抽取. 为了描述抽取任务目标, UIE 模型设计了一套模式化的提示符号, 即结构模式指导器 (Structural Schema Instructor, SSI), 用于控制需要抽取的实体、关系等抽取目标. SSI 具体包含如下三种片段.

(1) 实体名称 SPOTNAME.

(2) 关系名称 ASSONAME.

(3) 在每个实体名称、关系名称和输入文本序列前添加的特殊符号 ([spot], [asso], [text]).

$$y = \text{UIE}(s \oplus x), \quad (6)$$

$$s \oplus x = [s_1, s_2, \dots, s_{|s|}, x_1, x_2, \dots, x_{|x|}] = \begin{bmatrix} [\text{spot}], \dots, [\text{spot}], \dots, [\text{asso}], \dots, \\ [\text{asso}], \dots, [\text{text}], x_1, x_2, \dots, x_{|x|} \end{bmatrix} \quad (7)$$

上述公式给出了 SSI 的具体结构. UIE 将给定的结构模式指导器 (s) 和文本序列作为输入, 并根据模式 s 生成包含从 (x) 中提取的信息线性化 SEL(y), SEL 即结构化提取语言, 对需要抽取的任务目标定义统一的抽取结构.

指标评估采用了 P (Precision), R (Recall), F_1 (F1 Score) 值. 这里 F_1 值是精确率 (Precision) 与召回率 (Recall) 的调和平均数, 用于评估知识抽取任务的准确性; 其中, 精确率衡量系统抽取结果的准确性, 召回率衡量系统抽取结果的完整性. 理想情况下, 一个高效的知识抽取系统应同时具有高精确率和高召回率. P 表示被模型预测为正类别的样本中, 真正属

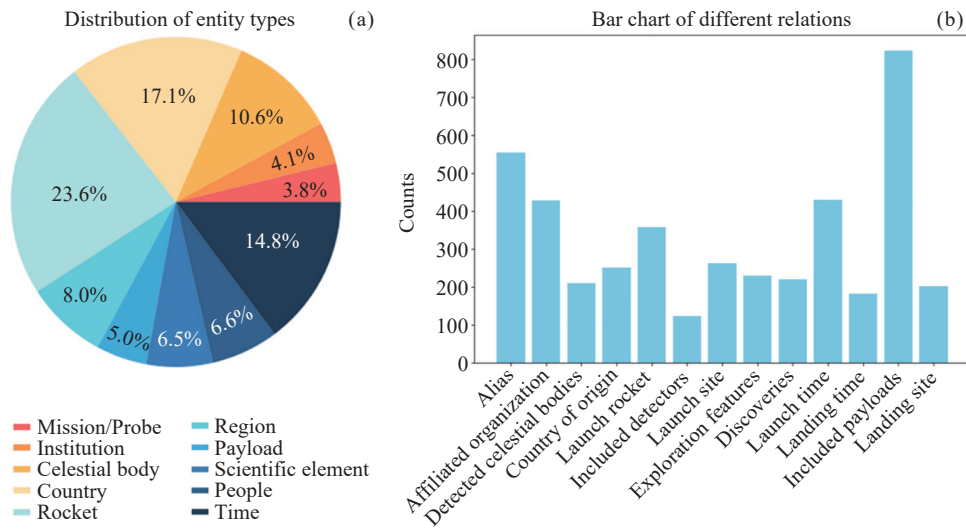


图 9 实体类型和关系类型分布情况

Fig. 9 Distribution of entity types and relation types

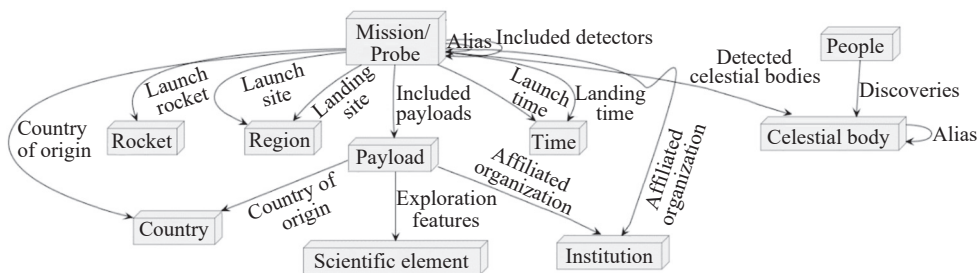


图 10 数据集 Schema

Fig. 10 Dataset Schema

于正类别的比例. R 表示真正属于正类别的样本中, 被模型预测为正类别的比例. N_{TP} 为正确抽取的信息数, N_{FP} 为错误抽取的信息数, N_{FN} 是未被抽取但是应该被抽取的信息数.

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}}, \quad (8)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}}, \quad (9)$$

$$F_1 = \frac{2PR}{P + R}, \quad (10)$$

$$F_{1-\text{Macro}} = \frac{1}{n} \sum_{i=1}^n F_{1-i}. \quad (11)$$

类别综合评价采用 Macro-F1 (定义符号 $F_{1-\text{Macro}}$)^[22], 其是一种多类别分类评估指标, 用于衡量分类模型在处理不平衡数据集时的性能. 其知识抽取结果见表 5.

非结构化知识抽取可作为后续知识的继续挖掘工具, 丰富深空探测领域知识库. 综合两类知识抽取方法, 以结构化、半结构知识为主, 在此基础上采用

表 5 知识抽取结果

Table 5 Knowledge extraction results

类别	Marco-P	Marco-R	Marco-F1
实体	0.891	0.800	0.841
关系	0.867	0.843	0.854

人工监督的方式融合非结构化知识, 构建了一套深空探测领域知识库, 目前知识库共有实体节点 2140 个, 实体之间关系 4608 条, 实体属性 9079 条.

4 基于深空探测知识的有效载荷配置

4.1 应用设计

在深空探测任务中, 针对既定科学目标下的载荷选择提出了一种有效载荷配置的应用场景, 根据用户输入的科学目标文本, 通过分析科学目标中涉及的探测天体和探测对象以及科学目标语义信息, 基于知识库中已存储的有效载荷类别信息及其关联的探测对象、有效载荷功能描述文本, 得出当前输入科学目标下的有效载荷配置信息, 其流程如图 11 所示, 具体包括以下步骤.

(1) 文本预处理. 清洗数据, 移除无关字符和停用词, 包括对输入文本中探测对象的识别.

(2) 选择知识库中的元路径. 天体→**→有效载荷 (功能描述、载荷类型)→探测对象. 前面括号中内容是有效载荷的实体属性, **表示多跳路径.

(3) 基于 Sentence-Bert^[23] 的语义相似度查询. 分别查询和计算某类型载荷的功能描述与输入的科学目标, 载荷关联的探测对象与输入文本中识别的科学

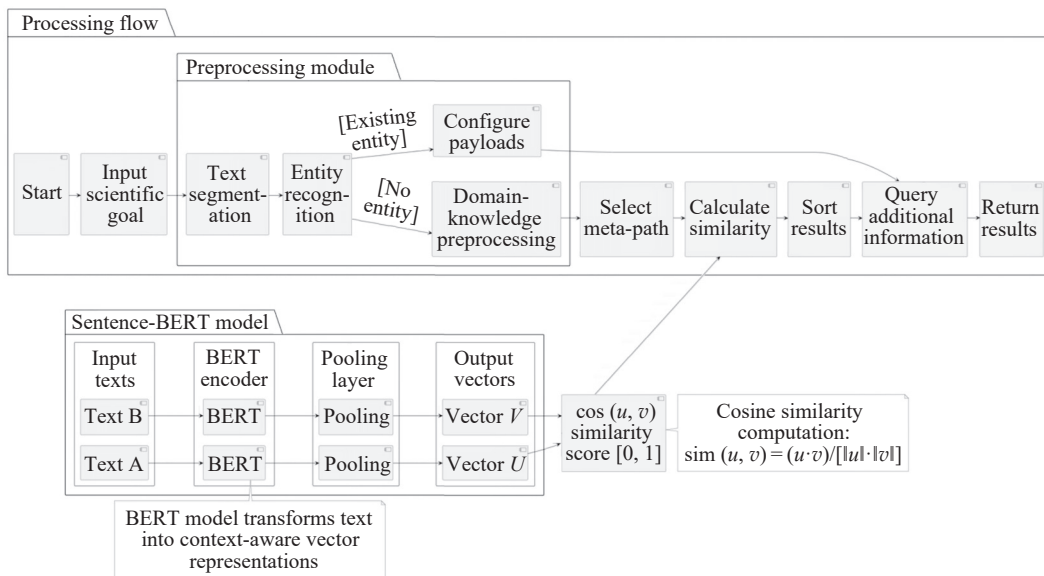


图 11 基于深空探测科学目标的有效载荷配置流程

Fig. 11 Payload configuration process based on scientific objectives of deep space exploration

要素两种相似度值, 进行线性加和, 得到每类载荷在该科学目标下的最大相似度。

(4) 载荷类型推荐. 根据查询结果取前 3 个推荐载荷类型为候选答案, 设定相似度阈值, 大于阈值的载荷类型即为推荐结果. 进行综合查询, 返回目前知识库中包含该类型载荷的所有信息。

(5) 参数配置信息推荐. 根据知识库中探测对象与所需载荷能力信息的关联情况, 结合已识别的探测对象名称, 推荐和配置出当前载荷类型下的能力指标。

4.2 实际结果与分析

以天问一号探测任务^[24]作为测试样例, 删除知识库中与天问一号任务相关的信息, 以历史上其他火星探测任务信息经过查询综合整理得到的有效载荷、有效载荷类型、探测对象等关联信息构建知识库. 输入天问一号探测任务的科学目标^[25], 进行模型配置, 结果见表 6. 可以看出在多数情况下模型能够正确推荐与实际载荷类型相关的载荷类别, 但在某些情况下存在一些差异, 这意味着模型可能没有充分捕捉到不同

有效载荷类型之间的复杂关系和细微差异, 尤其是那些与特定科学目标紧密关联的载荷. 其次, 由于输入的科学目标文本中对原始载荷功能描述的多样性, 对于科学目标的表述存在多样化, 当前测试使用的火星探测任务的知识库数据来源于其他国家已实施的科学任务, 载荷功能文本等是翻译整理得到的中文描述, 且表达和来源不一, 容易造成模型在语义相关上的误解, 从而出现部分误判等情况. 后续将引入更丰富的关系抽象机制, 增强模型对不同载荷类型之间复杂关系的理解能力, 继续扩充更加多样化的数据, 包含更多样化的科学目标描述和载荷功能描述, 提高模型对语言多样性的处理能力, 更好地理解不同描述之间的相互关联, 从而提升载荷配置能力. 此外, 还将继续整理知识库, 优化载荷与探测对象之间的具体参数指标关联等信息, 实现更加精准的载荷参数配置。

5 结语

从深空探测任务的应用规划需求出发, 以深空探

表 6 天问一号火星探测任务有效载荷实际选取与模型配置情况
Table 6 Actual selection and modelling of payloads for the Tianwen-1 Mars exploration

科学目标	科学任务	实际载荷	实际载荷类型	识别的科学要素	推荐载荷类型
研究火星表面土壤特征与水冰分布, 获取火星生命生存的条件和环境信息	火星表面和地下水冰的探测	环绕器次表层探测雷达	雷达	火星表面和地下水冰	雷达、*温度计
	火星土壤类型分布和结构探测	环绕器次表层探测雷达、火星矿物光谱分析仪	雷达、光谱仪	火星土壤类型分布和结构	雷达、光谱仪、*结构探测仪
	火星巡视区土壤结构(剖面)探测和水冰探查	火星车次表层探测雷达	雷达	水冰	*温度计、雷达、*望远镜
研究火星大气电离层及表面气候与环境特征, 研究火星物理场与内部结构, 获取太阳活动与火星相互作用的规律	火星大气电离层分析及行星际环境探测	火星磁强计、火星离子与中性粒子分析仪、火星能量粒子分析仪、甚低频射电接收设备(次表层探测雷达一体化设计)	磁强计、离子分析仪、粒子分析仪、雷达	行星际环境	磁强计、*雷达
	火星巡视区大气物理特征与表面环境探测	火星表面磁场探测仪、火星气象测量仪	气象仪、磁强计	火星巡视区大气物理特征与表面环境	气象仪、*重力仪、*雷达
研究火星形貌与地质构造特征及其变化, 获取火星演化历史的基础信息	火星地形地貌特征及其变化探测	中分辨率相机、高分辨率相机、环绕器次表层探测雷达	相机、雷达	火星地形地貌特征及其变化	相机、雷达
	火星巡视区形貌和地质构造探测	地形相机	相机	火星巡视区形貌和地质构造	相机、*雷达
研究火星表面物质组成, 获取火星化学演化的基础信息	火星表面物质成分的调查和分析	火矿物光谱分析仪	光谱仪	火星表面物质成分	光谱仪、*成分探测仪
	火星巡视区表面元素、矿物和岩石类型探测	火星表面成分探测仪、多光谱相机	相机、成分探测仪	火星巡视区表面元素、矿物和岩石类型	相机、成分探测仪、*雷达

注 *表示与实际情况的差异.

测任务科学目标下的有效载荷配置为切入点, 基于顶层知识和领域知识特点, 定义了深空探测领域的实体、关系及属性, 提出了深空探测领域本体模型构建方法, 并结合实时收集整合的数据提出一种本体动态更新算法. 基于本体架构及所采集的数据, 构建了深空探测领域知识抽取数据集, 针对结构化数据、半结构化数据和非结构化数据, 提出了不同的知识抽取方法, 构建了一套能够支撑当前应用和后续规划应用的知识库, 实现深空探测领域实体、属性及关系的可视化表示, 解决了深空探测领域知识多源异构、难以直接表达等问题. 从深空探测领域科学目标下的有效载荷配置场景出发, 提出了一种基于知识抽取模型和语义相似度综合的有效载荷配置应用方法. 通过这种方法, 可以根据既定科学目标得出所需载荷类型, 进而选择参考已有载荷能力, 为任务规划、决策制定等工作提供有力支撑, 缩短有效论证周期. 后续研究将继续扩大深空探测领域知识的覆盖度及深度, 完善已规划的深空探测领域知识应用场景, 如科学目标智能推荐与遴选、科学成果分析等, 研究更多可实际工作的应用场景, 构建更加丰富的深空探测领域知识库, 也将更新迭代算法管理, 扩展前沿算法在深空探测领域的应用, 助力深空探测领域科学研究的进一步发展.

参考文献

- [1] WU Weiren, YU Dengyun. Development of deep space exploration and its future key technologies[J]. *Journal of Deep Space Exploration*, 2014, **1**(1): 5-17 (吴伟仁, 于登云. 深空探测发展与未来关键技术 [J]. 深空探测学报, 2014, **1**(1): 5-17)
- [2] WU Weiren, WANG Chi, LIU Yang, *et al.* Frontier scientific questions in deep space exploration[J]. *Chinese Science Bulletin*, 2023, **68**(6): 606-627 (吴伟仁, 王赤, 刘洋, 等. 深空探测之前沿科学问题探析 [J]. 科学通报, 2023, **68**(6): 606-627)
- [3] YU Dengyun, MA Ji'nan. Progress and prospect of deep space exploration in China[J]. *Science and Technology Foresight*, 2022, **1**(1): 17-27 (于登云, 马继楠. 中国深空探测进展与展望 [J]. 前瞻科技, 2022, **1**(1): 17-27)
- [4] NECHES R, FIKES R E, FININ T, *et al.* Enabling technology for knowledge sharing[J]. *AI Magazine*, 1991, **12**(3): 36-56
- [5] LIU Jizhong, HU Chaobin, PANG Fuchuan, *et al.* Strategy of deep space exploration[J]. *SCIENTIA SINICA Technologica*, 2020, **50**(9): 1126-1139 (刘继忠, 胡朝斌, 庞涪川, 等. 深空探测发展战略研究 [J]. 中国科学: 技术科学, 2020, **50**(9): 1126-1139)
- [6] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th International Conference on World Wide Web. Banff Alberta: ACM, 2007: 697-706
- [7] WANG Peiyan, ZHANG Guiping, CAI Dongfeng. Constructing of ATHowNet: a semantic knowledge base for aviation terms[J]. *China Terminology*, 2021, **23**(3): 26-32 (王裴岩, 张桂平, 蔡东风. 航空术语语义知识库 ATHowNet 的构建 [J]. 中国科技术语, 2021, **23**(3): 26-32)
- [8] ZHAO Qichang, GUO Lingling, YANG Chunyan, *et al.* Payload allocation method of quantitative remote sensing and its development[J]. *Aerospace Shanghai*, 2022, **39**(6): 125-134 (赵其昌, 郭玲玲, 杨春燕, 等. 定量遥感有效载荷配置方法与发展 [J]. 上海航天, 2022, **39**(6): 125-134)
- [9] SINT R, STROKA S, SCHAFFERT S, *et al.* Combining unstructured, fully structured and semi-structured information in semantic wikis[C]//Proceedings of the 4th Semantic Wiki Workshop (SemWiki) at the 6th European Semantic Web Conference. Hersonissos, Greece: Heraklion, 2009: 73-87
- [10] WACHE H, VÖGELE T, VISSER U, *et al.* Ontology-based integration of information-a survey of existing approaches[C]//Proceedings of the IJCAI-01 Workshop on Ontologies and Information Sharing. Seattle: CEUR-WS.org, 2001
- [11] WANG Y M, WANG X L, FENG Y Q, *et al.* Strategic response for ease of doing business using case-based reasoning[J]. *Expert Systems with Applications*, 2022, **210**: 118514
- [12] LI Aihua, XU Yize, CHI Yuxue. Review of ontology construction and applications[J]. *Information Studies: Theory & Application*, 2023, **46**(11): 189-195 (李爱华, 徐以则, 迟钰雪. 本体构建及应用综述 [J]. 情报理论与实践, 2023, **46**(11): 189-195)
- [13] GE Yinchi, ZHANG Hui, SONG Wenyan, *et al.* Scientific and technology resource clustering based on domain ontology[J]. *Frontiers of Data & Computing*, 2020, **2**(5): 13-22 (葛胤池, 张辉, 宋文燕, 等. 基于领域本体的科技资源聚类方法研究 [J]. 数据与计算发展前沿, 2020, **2**(5): 13-22)
- [14] CAVNAR W B, TRENKLE J M. N-gram-based text categorization[C]//Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval. Las Vegas, 1994: 161-175
- [15] XIAO S T, LIU Z, ZHANG P T, *et al.* C-pack: Packaged resources to advance general Chinese embedding[OL]. arXiv preprint arXiv: 2309.07597, 2023
- [16] ARTHUR D, VASSILVITSKII S. k-means++: The advantages of careful seeding[C]//Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms. New Orleans: Society for Industrial and Applied

- Mathematics, 2007: 1027-1035
- [17] CUI M Y. Introduction to the k-means clustering algorithm based on the elbow method[J]. *Accounting, Auditing and Finance*, 2020, 1(1): 5-8
- [18] SHAHAPURE K R, NICHOLAS C. Cluster quality analysis using silhouette score[C]//2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA). Sydney: IEEE, 2020: 747-748
- [19] KALMAN D. A singularly valuable decomposition: the SVD of a matrix[J]. *The College Mathematics Journal*, 1996, 27(1): 2-23
- [20] ZAMAN G, MAHDIN H, HUSSAIN K, *et al.* Information extraction from semi and unstructured data sources: A systematic literature review[J]. *ICIC Express Letters*, 2020, 14(6): 593-603
- [21] LU Y J, LIU Q, DAI D, *et al.* Unified structure generation for universal information extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: ACL, 2022
- [22] GOWDA T, YOU W Q, LIGNOS C, *et al.* Macro-average: rare types are important too[C]//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. ACL, 2021
- [23] FENG F X Y, YANG Y F, CER D, *et al.* Language-agnostic BERT sentence embedding[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin: ACL, 2022
- [24] ZHU Yan, BAI Yunfei, WANG Lianguo, *et al.* Integral technical scheme of payloads system for Chinese mars-1 exploration[J]. *Journal of Deep Space Exploration*, 2017, 4(6): 510-514,534 (朱岩, 白云飞, 王连国, 等. 中国首次火星探测工程有效载荷总体设计 [J]. 深空探测学报, 2017, 4(6): 510-514,534)
- [25] LI Chunlai, LIU Jianjun, GENG Yan, *et al.* Scientific objectives and payload configuration of China's first mars exploration mission[J]. *Journal of Deep Space Exploration*, 2018, 5(5): 406-413 (李春来, 刘建军, 耿言, 等. 中国首次火星探测任务科学目标与有效载荷配置 [J]. 深空探测学报, 2018, 5(5): 406-413)

作者简介



阳 松 男, 1998 年 6 月出生于湖北省荆州市, 现为中国科学院国家空间科学中心硕士研究生, 计算机技术专业.

E-mail: yangsong211@mailsucas.ac.cn



牛文龙(通信作者) 男, 现为中国科学院国家空间科学中心副研究员, 硕士生导师, 主要研究方向为运动目标检测、智能目标识别、图像处理、信号处理、航天系统仿真与评估.

E-mail: niuwenlong@nssc.ac.cn