计算机科学
COMPUTER SCIENCE

基于企业数据空间的数据资源组织方法与数据资产管理

李敏波, 王少华, 吴大臻

引用本文

李敏波, 王少华, 吴大臻. 基于企业数据空间的数据资源组织方法与数据资产管理[J]. 计算机科学, 2026, 53(5): 119-128.

LI Minbo, WANG Shaohua, WU Dazhen. Data Resource Organization Method Based on Enterprise Dataspace and Data Asset Management [J]. Computer Science, 2026, 53(5): 119-128.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.250600019

基于企业数据空间的数据资源组织方法与数据资产管理

李敏波[1]　王少华[2]　吴大臻[1]
1 复旦大学计算与智能创新学院　上海 200433
2 浪潮云洲工业互联网有限公司　济南 250101
(limb@fudan.edu.cn)

摘　要　针对军工科研单位因保密分级分块管理导致的数据资源孤岛问题,以及数据资源智能检索与知识复用困难,提出了多源异构企业数据资源与数据资产的治理方案。通过企业数据空间属性图模型实现数据资源之间的图节点关联映射,构建融合BOM树型结构的数据资源知识图谱,涵盖研发工艺、生产制造与质检数据的层级关系、属性信息及关联关系。具体地,提出了一个新颖的RAG框架HireRAG,建立了基于C-HNSW的知识图谱社群化分层索引,低层保留细粒度的知识单元,高层社群提供全局摘要,以处理不同层次的检索;提出了一种图增强聚类算法,使得C-HNSW更好地捕捉知识图谱中的语义信息。实验表明,HireRAG相比现有的一些先进RAG框架更适合处理企业数据空间BOM关联数据,可以在实现最高准确率的同时保证检索效率优势。数据资产管理系统确保数据资产全过程合规入表。

关键词:企业数据空间;数据资源;知识图谱;数据检索;HNSW
中图分类号　TP391

Data Resource Organization Method Based on Enterprise Dataspace and Data Asset Management

LI Minbo[1], WANG Shaohua[2] and WU Dazhen[1]
1 College of Computer Science and Artificial Intelligence, Fudan University, Shanghai 200433, China
2 Inspur Yunzhou Industrial Internet Co., Ltd., Jinan 250101, China

Abstract　Aiming at the problem of data resource islands caused by confidentiality hierarchical and block management in military research institute, and the difficulty of intelligent retrieval and knowledge reuse of data resources, a governance solution for multi-source heterogeneous enterprise data resources and data assets is proposed. The graph node association mapping between data resources is realized through the attribute graph model of enterprise data space, and a data resource knowledge graph integrating the BOM tree structure is constructed, covering the hierarchical relationship, attribute information and association relationship of R&D process, production and manufacturing, and quality inspection data. This paper proposes a novel RAG framework—HireRAG, and establishes a community-based hierarchical index of knowledge graph based on C-HNSW. The low-level retains fine-grained knowledge units, and the high-level community provides a global summary to handle retrieval at different levels. A graph-enhanced clustering algorithm is proposed to enable C-HNSW to better capture the semantic information in the knowledge graph. Experiments demonstrate that HireRAG is more adapt at processing bill of materials(BOM) related data within enterprise data spaces compared to several existing advanced retrieval-augmented generation(RAG) frameworks. Furthermore, it achieves superior performance metrics in both retrieval recall and accuracy. The data asset management system ensures that the data assets are entered into the table in compliance with the whole process.

Keywords　Enterprise dataspace, Data resources, Knowledge graph, Data query, Hierarchical Navigable Small World Graphs

1　引言

进入数字经济时代,数据成为一种新型生产要素,基于数据信息的价值创造成为社会经济发展的活力源泉。数据空间是一种由数据治理框架定义的分布式系统,该框架在确保数据主权的同时支持安全可信的数据共享。

企业数据通常分散存储于多个异构系统中,如产品生命周期管理(PLM)系统、制造执行系统(MES)、质量管理系统(QMS)以及企业资源计划(ERP)系统等,导致数据孤岛现象严重。这种分散且异构的数据环境使得数据的统一管理、高效检索和智能应用面临巨大挑战。传统的数据集成方法(如数据仓库和ETL技术)在灵活性、实时性和语义一致性方面

到稿日期:2025-05-29　返修日期:2025-08-11
基金项目:国家重点研发计划(2023YFC3304400)
This work was supported by the National Key R&D Program of China(2023YFC3304400).
通信作者:王少华(wangshaohua03@inspur.com)

存在局限性，难以满足企业日益增长的数据协同与价值挖掘需求。然而，企业在构建数据空间时面临着诸多挑战。首先，企业内部存在各类多源异构的数据，包括产品研发数据、经营管理数据、生产数据以及海量的知识文档。这些数据来源广泛，数据类型多样，如何有效地组织和管理这些异构数据是一个关键问题。军工科研单位，其科研数据与知识文档由于保密分级分块管理，难以实现单位内部及跨组织的智能检索与知识协同复用。其次，现有的文档管理与知识管理系统在处理这些数据时存在明显的不足，例如难以自动化识别提取文档摘要、图表信息，导致智能检索水平较低，无法满足企业对数据高效利用的需求。此外，数据空间的动态性要求系统能够实时响应数据源的变化，包括数据的更新、数据源的加入与退出等。数据空间的安全性和隐私保护也是需要重点关注的问题。

随着大语言模型(LLM)与检索增强生成(Retrieval Augmented Generation，RAG)技术的融合应用，本文提出了一种基于企业数据空间框架的数据资源组织方法，旨在高效存储企业异构信息系统的数据文件与海量知识文档属性、摘要与数据资源间的关系，基于产品 BOM 树型结构构建知识图谱，实现数据资源的语义关联属性图，基于 C-HNSW 建立知识图谱社群化分层索引，低层保留细粒度的知识单元，高层社群提供全局摘要，以处理不同层次的检索。设计图增强聚类算法，使得 C-HNSW 能够更好地捕捉知识图谱中的语义信息，从而提高检索的准确性和效率。开发数据资产入表管理系统，实现数据资产的精细化管理。本研究将为企业数据资源组织与智能检索提供新的思路和方法，推动企业数据价值的深度挖掘与应用。

2 相关工作

2.1 数据空间

数据空间(Data Space)作为一种新兴的数据管理范式，旨在通过动态关联多源异构数据，构建灵活、可扩展的数据服务体系，用于解决分散、异构数据的共享与数据集成问题。文献[1-2]最早提出数据空间的概念，将其定义为由异构资源对象及其关联关系组成的集合，并通过带标签的图模型进行可视化表征，底层支持关系模型、XML 数据等多种数据模型。这一概念突破了传统数据库的刚性模式，为动态数据集成提供了理论基础。文献[3-4]提出了用图刻画的数据空间模型 iDM，他们用一种统一资源视图的概念和形式化表示方法，实现对各种数据类型(如文档、目录、关系表、数据流等)的统一表示。数据空间采用数据集成、数据虚拟化、语义建模和元数据管理等技术统一组织管理数据，提供数据编目和浏览、搜索和查询、更新和监控、事件检测和支持复杂工作流等服务。数据空间强调以一种“按需付费”(Pay-as-You-Go)的方式动态整合数据，更加注重数据的价值应用，而非数据的物理存储形式。

在企业应用场景中，数据空间的构建需要解决信息孤岛问题，例如工业领域需整合 ERP，MES，CRM 等系统的数据，形成统一视图以支持决策。德国弗朗霍夫应用研究促进协会(Fraunhofer Gesellschaft)提出，国际数据空间 IDS(International Data Space)是基于连接器构建的一个分布式存储、中心化认证的数据流通环境，参考架构模型(IDS-RAM)由 5 层组成，即业务层、功能层、过程层、信息层、系统层，还包括需要在 5 个层中实现的视角，即安全、认证和治理。IDS 为供应链上下游企业搭建数据共享流通平台，强化主体身份认证和数据使用控制。

企业数据空间(Enterprise Data Space，EDS)则是以整个企业为主体，以企业中各个部门的信息系统中的数据及其关联关系为管理对象的数据空间[4]。EDS 中，数据关联基于异构多源的数据对象，即任何数据对象之间都可以建立关联，因此其可以提供按需、即时、灵活的数据服务。与 IDS 不同，企业数据空间重点是面向企业内部各部门的各类数据资源的关联管理与数据服务。

中国人民大学孟小峰团队提出了数据空间系统框架[5]，包括数据集成引擎、数据空间引擎、数据演化引擎和数据输出引擎。数据空间应支持用户查询数据空间的任何数据，涉及查询优化、查询转换、查询接口等多方面。

数据空间可以看作一个以数据对象为节点、以数据关系为边的图，基于图的数据搜索技术可以应用到数据空间查询，提升查询效率和准确性，包括图索引结构[6-7]、图相似性度量[8]、图搜索优化[9]等方面。Van Bruggen 等[10]利用属性图数据模型来描述数据空间中存在的各种异构数据。企业数据空间利用属性图模型将所有数据描述并关联起来，形成一个与企业相关的属性图。这种模型不仅能够有效管理数据，还能支持复杂的数据关联和查询。Elsayed 等[11]提出了基于本体的数据模型，用来管理数据源之间的关系。他们应用 RDF 实现了对科学数据的组织，将不同数据源的元数据用三元组(实体、属性、值)进行描述，从而实现数据的语义化管理。Yang 等[12]进一步提出了数据空间中实体、实体类和关系的定义，基于数据源之间的数据语义关系构建通用的分层数据模型。该模型包括实体数据图和实体模型图，分别用来捕捉实体之间的关系和实体类之间的关系，为数据语义查询和分析提供了基础。

2008 年，Franklin 等[13]指出，未来数据空间的研究包括 5 个方面：模式映射和匹配、索引融合、数据空间剖析、数据血缘和溯源、信息抽取和关键字查找。

2.2 数据资产相关概念

数据资源是指具有使用价值的数据集合，可以被组织和利用。数据资源不仅包括原始数据，还包括经过处理和分析后得到的有价值的信息和洞察力。数据资产是指企业过去交易或者事项形成的、由企业合法拥有或者控制的，能够为企业带来未来经济利益的、以物理或者电子方式记录的数据资源[14]。数据资产的构成主要涵盖数据资源和数据产品两大类型。数据产品的形态包括数据集、数据服务(如数据 API)和数据应用(如 SaaS 平台或数据看板)。数据资产具有可复制性、非独占性、依存性(需依托特定载体或系统存在)、价值易变性、时效性、可加工性属性。数据要素化包括数据资源化、数据资产化与数据资本化 3 个递进的过程。

数据资产可入表的确认条件：1)由企业过去的交易或在生产经营过程中积累产生；2)企业拥有或者控制数据资源，企

业享有其所有权，或虽无所有权但能被企业控制；3）预期会给企业带来经济利益（增加收入、减少成本等）；4）成本能够可靠地计量，价值（市场价值、经济利益等）可合理量化评估。

数据资本化以及数据资产入表的完成依赖于数据资产的价值评估。目前数据资产价值评估方法有 3 类：1）传统的无形资产价值评估方法，采用成本法、收益法与市场法来评估数据资产的价值；2）利用层次分析法、博弈论进行数据资产估值的方法；3）数据资产的非货币价值评估方法。

2.3 检索增强生成 RAG 技术

近年来，检索增强生成（RAG）技术在降低大模型出现幻觉的概率上已经取得了显著的进步。它通过检索阶段将用户的问题转换为向量，从外部知识库或私有文档中快速检索相关片段并在生成阶段将检索到的信息输入大模型，从而生成结合上下文的具体问答。这种方式可以控制大模型输出尽可能相关并且语义相通的内容，让大模型的回答更加准确可信，从而提高 LLM 在不同专业领域的实际应用潜力。

典型 RAG 系统通常包含文本切块、向量转换、数据存储、信息检索、二次排序及内容生成等核心模块，具备高度的灵活性与可扩展性。在这一框架下，涌现出诸多具有代表性的实现方式。RAGFlow 是一款基于深度文档理解构建的开源 RAG 引擎，它结合 LLM，针对用户各类不同的复杂格式数据，提供可靠的问答以及有理有据的引用。QAnything 是致力于支持任意格式文件或数据库的本地知识库问答系统，通过引入重排序（Rerank）机制，进一步提升了检索效果。Dify 是一个开源的大型语言模型应用开发平台，具备高度可配置性，为用户提供了更为灵活多样的选择。RAG-GPT 利用 LLM 和 RAG 技术，从用户自定义的知识库中进行学习，为广泛的查询提供上下文相关的答案。

微软研究团队率先提出 GraphRAG[15]，将知识图谱技术与 RAG 有机结合，利用大型语言模型从非结构化文本中提取实体、关系和摘要，构建知识图谱，并结合图和向量索引技术进行高效的检索和问答，提高了对复杂用户查询的检索和响应质量。香港大学提出的 LightRAG[16]，通过将图结构纳入文本索引以建立相关实体之间的复杂关系，从而提升系统的上下文理解能力。LightRAG 采用双层检索框架，通过向量化的关键词匹配，实现局部和全局信息的高效整合，降低了计算和存储开销。

3 基于数据空间的企业数据资源组织与检索

3.1 数据资源的组织方法

现有企业存在众多的多源异构信息系统，PLM 系统管理产品研发数据，MES 系统管理各种生产制造数据，质量管理系统管理质量检测数据与质量报告，知识档案系统管理归档知识文档。这些多源异构信息系统难以集中管理和融合产品研发工艺与生产制造质量的数据与知识文档。

为了实现企业多源异构数据资源（包括数据文件和知识文档）的数据检索与大模型智能问答，设计了企业数据空间的分层数据组织结构，按照数据资源目录、数据资源模型、数据索引 3 个层次进行组织与管理，如图 1 所示。其中数据空间是与企业主体相关的所有数据和数据间关系的集合，数据资源目录是从多维多角度对数据空间中的数据进行分类和组织的一种树型目录结构[4]，也是企业中数据资源的分类标准。其按照产品 BOM 结构树、项目管理 WBS 任务树组织管理产品及其零部件，或者项目的产品 CAD/CAM/CAE 文件以及工艺文件数据。非结构化生产数据与质量数据/文档通过所属产品或项目号属性进行关联。

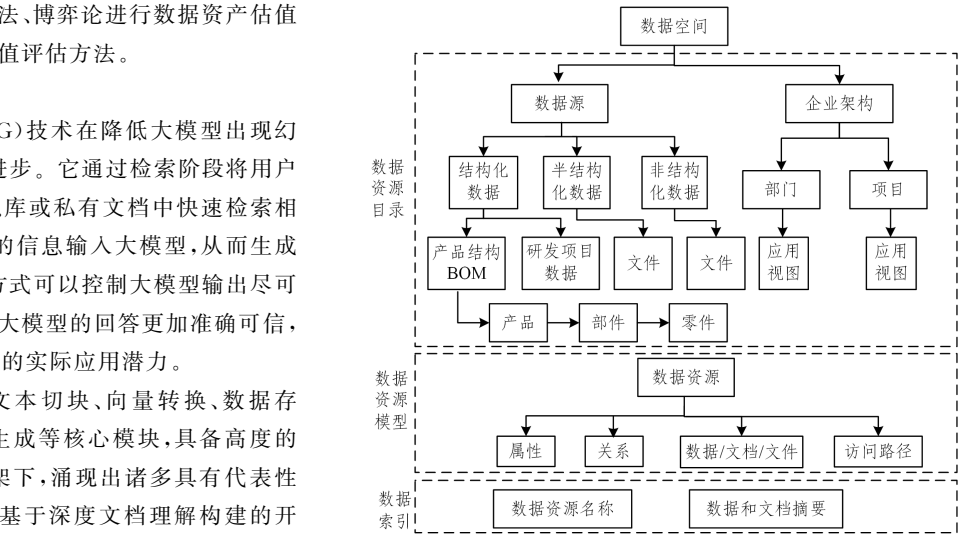

图 1 企业数据空间的分层组织结构

Fig. 1 Hierarchical structure of the enterprise dataspace

数据资源目录的每一个叶子节点都有对应的数据资源，数据资源可以是实际的数据源，如具体的数据集、数据库系统、知识文档文件等，也可以是虚拟的数据源，如数据视图。在数据空间中，数据资源目录的结构是灵活的、动态的，一个数据资源可以属于多个目录节点[4]。由于大型企业大多已构建了产品全生命周期 PLM 系统，因此基于产品结构 BOM 构建产品-部件-零件多层树型结构以及项目管理 WBS 的文档结构，利用产品 BOM 树型结构或项目管理结构建立产品研发设计文件、文档、工艺文件数据之间的上下级关联关系，将 MES 系统中的生产制造数据和 QMS 质量检测数据、质量报告等文档与 PLM 系统研发工艺数据按照归属的产品、零部件编号关联链接。

数据资源模型代表着不同数据资源的数据结构，包括数据资源的属性（数据名称、编码、所属项目、数据类型等）、所属的产品或零部件物料编码（数据资源与产品零部件关系定义）、物理存储的数据/文档文件、数据资源访问路径。如图 2 所示，企业数据空间利用属性图模型将所有数据描述并关联起来[4]，节点（Node）是属性图模型中的一个基本元素，用来表示各种类型的数据，可以是数据源、数据资源目录分类节点、具体的数据资源等。节点的标签表示数据的类型或模式信息，属性集描述节点的具体信息。节点可以拥有多个属性和一个标签，关系是任意两个节点间可能存在的关联关系，包括部件-零件上下级关系、零件与文档/文件所属关系等。关系将节点关联起来构成图，也就是图论中的边，关系是有方向的。属性图模型的数据结构，可以形式化定义为一个三元组 $G=(E,R,S)$，E 为知识库中的实体集合，R 为关系集合，S 代表知识库中的三元组集合（Headnodes，Relationship，Tailnode）。

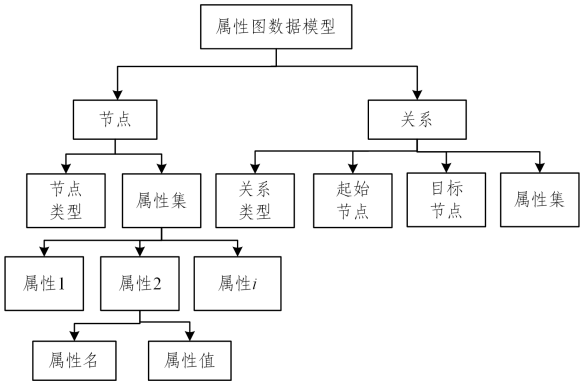

图 2 属性图数据资源模型

Fig. 2 Data resource model of attribute graph

3.2 基于产品 BOM 的数据文档图结构

3.2.1 基于 BOM 构建数据资源图谱的意义

BOM(Bill of Materials)结构的知识图谱化是将传统线性表结构升级为语义网络的过程,其核心意义在于通过节点属性与关系的显式标注实现工程逻辑的机器可理解表达,同时以图论邻接矩阵理论建模多级嵌套结构(产品→部件→零件→文档),支持数学建模与算法分析。在实践层面,知识图谱化显著提升了数据管理效率:通过统一实体标识符消除跨系统数据孤岛,时间戳属性实现动态版本控制,唯一性约束降低存储冗余;工程决策智能化方面,结合图算法与语义推理,可精准识别关键路径、优化供应链配置、缩短设计协同周期;查询与推理能力方面,Cypher 路径匹配与 OWL 本体规则验证显著增强复杂关系处理能力。相较于传统 BOM 管理,知识图谱在数据结构(图结构支持复杂关系)、查询效率、语义表达、动态扩展及可视化能力等方面具有显著优势。

3.2.2 基于 BOM 结构构建知识图谱

以简化的火箭发动机为例,构建了包含层级结构、属性信息及文档关联关系的 BOM 知识图谱。该知识图谱的核心实体与关系首先通过 JSON 文件进行结构化定义。

图 3 展示了一个典型的火箭发动机的数据资源 JSON 描述文件,其内容包括 BOM 层级结构、属性信息和文档关联。

1)BOM 层级结构:覆盖从顶层产品(如火箭发动机)到零部件(如燃烧室、喷嘴等)的多级嵌套关系;使用产品工程物料清单 BOM(EBOM)和制造物料清单 BOM(MBOM)数据。

2)属性信息:涵盖零部件的物料属性(如型号、材质)、工艺参数及技术规范。

3)文档关联:明确产品、部件及零件与设计文件、工艺文件、质量检测文件(如测试报告)之间的映射关系。

通过这种结构化描述,可完整呈现产品全生命周期中的技术数据链,为后续知识图谱的构建与应用奠定基础。

基于火箭发动机 BOM 的 JSON 数据,构建一个结构化、语义化的知识图谱,实现以下目标。

1)层级关系可视化:清晰表达产品-部件-零件-文件的多级嵌套结构。

2)属性关联:整合技术参数(如型号、制造商)、物理属性(如文件大小、创建时间)及工程规范(如冷却方式)。

3)动态扩展性:支持新增部件、零件或文件节点,并维护关系一致性。

4)智能查询:通过图数据库(如 Neo4j)实现复杂路径查询(如“某部件的所有文件”或“某零件的上层产品”)。

首先定义知识图谱的核心实体与关系,如表 1 所列。

```
{
  “product_info”:{
    “product_name”:“火箭发动机”,
    “engine_model”:“HCKZ001”
  },
  “file_list”:[
    {
      “file_code”:“3GB010”,
      “file_type”:“PDF”,
      “file_name”:“火箭发动机设计与原理文件”
    },
    {
    “file_code”:“3GB012”,
      “file_type”:“Word”,
      “file_name”:“火箭发动机性能参数与测试文件”
    }
  ],
    “bom”:
      {       “bom_line_number”:“0021”,
      “model”:“7XM010”,
      “manufacturer”:“100903”,
      “material_description”:“TURBINE_1500R”,
      “related_files”:[“燃烧室冲压模具设计图纸”]
    },
    ]{
      “bom_line_number”:“0022”,
      “model”:“7XM011”,
    “manufacturer”:“100904”,
      “material description”:“VALVE_300S”,
      “related_files”:[“燃烧室质检数据”]
    }
  ],
  “components”:[

    {
      “component_name”:“燃烧室壳体”,
      “related_files”:[
        “火箭发动机燃烧室壳体成形工艺设计”,
        “燃烧室壳体及燃气发生器”
      ]
    },
    {
      “component_name”:“绝热层”,
      “related_files”:[“待补充”]
    },
    {
      “component_name”:“推进剂”,
      “related files”:[“待补充”]
    }
  ]
}
```

图 3 火箭发动机 BOM 结构的 JSON 文件

Fig. 3 JSON file of BOM structure for rocket engine

表 1 知识图谱核心实体与关系定义

Table 1 Knowledge graph core entity and relationship definitions

实体类型	属性
产品	名称、型号、制造商
部件	名称、型号、描述
零件	名称、材质、工艺要求
文件	编码、类型、大小、创建时间
关系类型	**方向**
CONTAINS_PART	产品→部件→零件
HAS_FILE	部件/零件→文件

使用 Neo4j 图数据库构建 BOM 图谱，利用其原生 Cypher 查询语言高效处理层级关系。核心步骤如下。

步骤 1 节点创建：采用 MERGE 语句确保产品、部件、零件及文件节点的唯一性（基于 name+engine_model 等属性组合判重）。

步骤 2 关系构建：逐层遍历 JSON 数据，动态建立：CONTAINS层级关系与：HAS_FILE 文件关联。

步骤 3 数据一致性：通过 WITH 子句维护上下文，实现跨层级关系的准确连接。

图 4 为构建出的知识图谱，可以清晰看到各种不同层级的节点以及节点之间的不同关系，可以使用 Cypher 语法实现路径查询、属性过滤以及多跳推理等功能。

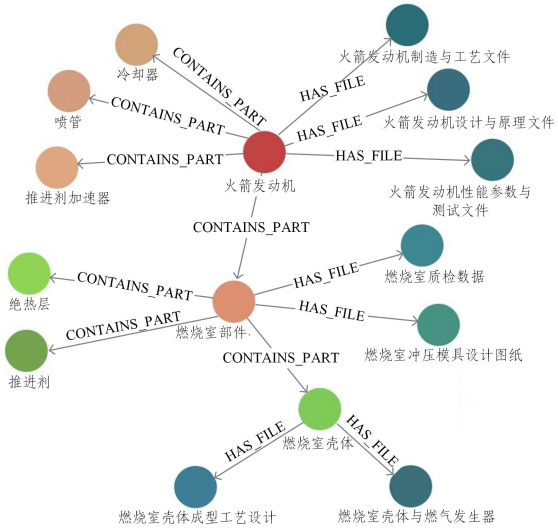

图 4 火箭发动机数据资源的知识图谱

Fig. 4 Knowledge graph of data resources for rocket engine

将 BOM 结构进行知识图谱化有诸多优势。相比传统表格，知识图谱能更直观地表达层级关系和隐含逻辑（如“某零件仅用于某型号发动机”），能够支持自然语言接口（如“哪些文件与燃烧室部件相关?”），且易于集成其他数据源（如供应链信息、测试数据）。

3.2.3 基于分层树的知识检索优化

传统知识图谱检索方法在处理企业 BOM 目录树结构化知识时存在显著局限性。企业 BOM 数据具有典型的多层级特性，通常包含产品（Product）、部件（Component）、零件（Part）和文件文档（Document）四级结构。例如，一个产品可能由多个部件组成，每个部件又由若干零件构成，而零件可能关联技术文档（如工艺文件、检验报告等）。然而，现有检索方法（如基于全文匹配或简单图遍历）难以高效处理这种多粒度、强关联的分层结构，导致以下问题。

1）抽象与具体问题的平衡不足：用户可能需要从全局（如产品级）到局部（如零件级）的多尺度检索，而传统方法无法动态调整检索粒度。

2）Token 消耗高：为覆盖所有层级信息，需在单次检索中加载大量上下文，导致模型资源浪费和效率下降。

HNSW（Hierarchical Navigable Small World Graphs）[17]是一种高效的近似最近邻搜索（Approximate Nearest Neighbor，ANN）算法，专门用于处理高维空间中的大规模数据检索。其核心思想是通过分层图结构和贪心搜索策略，在保证较高精度的同时显著提升搜索效率。基于 HNSW，本文提出了 C-HNSW（Community-based HNSW），利用知识图谱中的链接和属性来检测能够组织成层次树状结构的社群，也称为属性社群。属性社群是一组相同类型的实体，这些实体不仅密集相连，而且共享相似的主题，提供用于回答特定问题的详细信息，而摘要则可以为回答抽象问题提供一个浓缩的观点。此外，在层次树状结构中，较低层次的社群和实体包含来自知识图谱的详细知识，而较高层级的社群提供全局摘要，这使得系统能够处理不同层次的查询。

1）离线索引阶段

离线索引分为 3 个主要步骤，即知识图谱构建、分层聚类以及 C-HNSW 索引构建，如图 5 所示。

（1）知识图谱构建

除了第 3.2.2 节中所述手动构建方法，设计了基于大语言模型的自动化知识图谱构建流程，该流程包含以下关键步骤。

①提示词设计与优化

为了让大语言模型准确理解 BOM 文档结构和提取实体关系，设计了三阶段渐进式提示策略。

任务定义提示：明确告知模型对 BOM 文档进行实体与关系提取任务。

结构化输出提示：要求模型以特定 JSON 格式输出，确保提取结果的一致性。

示例引导提示：通过少样本学习（Few-shot Learning），提供产品-部件-零件-文档的示例三元组。

提示词模板示例如下：

请分析以下 BOM 文档，识别其中的产品、部件、零件和相关文档。

将识别结果以 JSON 格式输出，格式为：

```
{
  “entities”:[
    {“id”:“…”,“type”:“产品|部件|零件|文档”,“name”:“…”,“属性 1”:“值 1”,…},
    …
  ],
  “relationships”:[
    {“source”:“entity_id”,“target”:“entity_id”,“type”:“CONTAINS|HAS_FILE”,“属性”:“值”},
    …
  ]
}
```

示例 1 产品“火箭发动机”包含部件“燃烧室”，燃烧室关联文档“燃烧室设计图”

②实体命名标准化与消歧

在 HireRAG 中，处理实体消歧主要通过以下方式。

利用 LLM 进行判断：图提取的最后一步是解析代表相同现实世界但名称不同的实体，这是通过 LLM 完成的。会将一系列实体提供给 LLM，要求其确定哪些实体应该合并，然后将这些实体合并为一个实体，并更新它们的关系。

社区检测辅助：在图增强阶段使用层次化的 leiden 算法生成实体属性社区的多层次结构，通过这种社区聚类可以发现紧密相关的实体集群，从而为实体消歧提供一定的依据和参考，有助于更好地确定不同实体之间的关联和归属。

③大模型输出质量控制

为保证知识图谱构建质量，设计了如下自动化质量控制流程。

关系一致性验证：检查提取的关系是否符合领域约束（如“零件不能包含产品”）。

置信度阈值过滤：大模型输出每个三元组的置信度分数，低于阈值的会被标记审核。

增量学习机制：根据专家反馈调整提示词，优化提取准确率。

基于现有研究和工业实践的分析，这种基于大语言模型的知识图谱构建方法有望显著提高图谱构建效率，优化的提示策略和质量控制机制可以使实体识别准确率达到 90% 以上，关系提取准确率超过 85%。本文方法的设计融合了最新的自然语言处理技术和图结构分析方法，能够进一步提升知识图谱构建的准确性和效率。

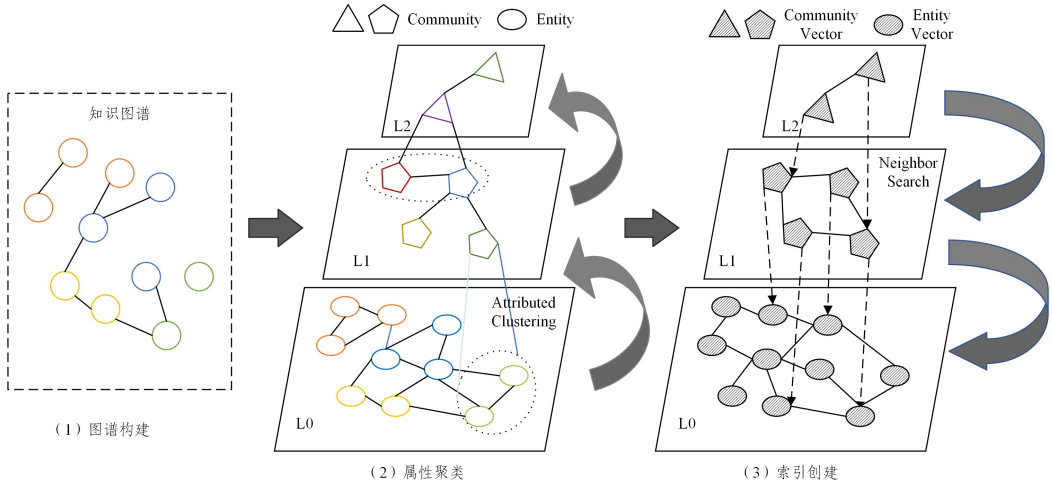

图 5 离线索引创建

Fig. 5 Knowledge offline index creation

(2)BOM 结构动态更新机制

针对 BOM 结构变更（零件增删改）导致的知识图谱更新问题，设计了高效的增量更新策略。

变更检测与传播：当检测到 BOM 节点（产品/部件/零件）的增删改时，确定受影响的子图范围（如被删除零件及其关联的所有文档和关系）。

图谱增量更新：节点/关系的增删直接在 Neo4j 中执行相应的“CREATE”“DELETE”“MERGE”“SET”操作。对于属性的修改，更新节点或关系的属性值（“SET”操作）；对于删除操作，级联删除其独有的属性和关系（需谨慎处理共享关系），对于修改，评估是否影响其向量表示（如关键属性改变）。

C-HNSW 索引增量更新：原则上需要避免全局重建，优先进行局部更新。①对于新增的节点，需要计算新节点的向量表示，以及根据其属性社区归属（由分层聚类结果决定），将其插入到对应底层社区的 HNSW 索引中（使用 HNSW 的增量插入接口）。若新节点属于新社区（或导致社区分裂），需在 C-HNSW 的相应层级创建新社区节点并建立层间链接（影响范围限于该层级及上层路径）。②对于删除的节点，将其从所属底层社区的 HNSW 索引中标记为删除（逻辑删除）或物理移除（需重建该社区的小索引），移除其在高层社区摘要中的贡献（可能触发高层摘要重新计算）。③对于修改节点（向量变化显著时），将其视为在旧位置删除＋在新位置插入。变化不显著时，可仅更新索引中的向量值（如果索引库支持）。

更新代价分析：包括时间复杂度和空间复杂度分析。①时间复杂度：主要开销在受变更影响社区的局部索引更新上。设变更影响 c 个底层社区，每个社区平均大小为 s，则插入/删除单节点的代价接近 $O(\log s)$。社区结构调整（分裂/合并）的代价为 $O(s\log s)$，但发生频率远低于单点变更。全局重建代价为 $O(N\log N)$（N 为总节点数）。②空间复杂度：增量更新本身不显著增加额外存储。逻辑删除会暂时占用空间，需定期清理。

(3)基于大语言模型的分层聚类

设计了一个迭代的基于大语言模型的分层聚类框架。具体来说，通过链接属性相似度高于阈值的实体来增强知识图谱，然后将每对链接的实体与表示其属性相似度值的权重关联起来。接下来，使用任何给定的图聚类算法（如加权莱顿[18]、加权谱聚类和 SCAN[19]）生成属性社区，并为每个社区生成摘要。

算法 1 给出了上述迭代聚类过程。给定一种图增强方法和聚类算法，在每次迭代中执行以下步骤。增强图并计算边权重（第 3 行）；对增强后的图进行聚类（第 4 行）；使用大型语言模型为每个社区生成摘要（第 5 行）；构建一个新的带属性图，其中每个节点表示一个属性社区，如果两个社区的成员是相连的，则这两个节点被链接（第 7 行）。重复迭代，直到满足

停止条件（如节点不足或达到指定的层级限制）。由于每次迭代对应一层，因此所有分配的社区 HC 可以被组织成一个多层层级树状结构，记作 Δ，其中一层的每个社区包括下一层的多个社区。

算法 1 基于大型语言模型的层次化聚类

输入：知识图谱 G(V,E)、增广函数 Aug、图聚类算法 GCluster 以及终止条件 T

输出：层次化图

1. T←False，HC←Ø；
2. 重复
3. 对于每个 $e'(u,u)\in E'$，执行 $G'(V,E')\leftarrow Aug(G(V,E))$ 更新边 e' 的权重为 $1-\cos(zu,zv)$；
4. C←GC 聚类（$G'(V,E')$）；
5. 对于每个 $c\in C$，通过 LLM 生成 c 的摘要；
6. HC←HC∪C；
7. G(V,E)←使用 C 和 E′构建一个新图；// 根据 G(V,E) 更新 T；
8. 直到 T= True；
9. 返回 HC.

C-HNSW 索引的构建采用自上而下的方法，利用查询过程迭代插入节点并建立链接。首先，从顶层开始插入节点，并在同一层内形成层内链接。然后，将每个节点与其下一层的最近邻节点相连，建立层间链接。这个过程确保了索引的高效构建，同时考虑了知识图谱中的语义信息和社区结构。

在构建过程中，C-HNSW 利用知识图谱中的属性和关系来增强节点之间的连接。例如，节点之间的层间链接不仅基于向量相似性，还考虑节点在知识图谱中的共同属性和关系。这种设计使得 C-HNSW 能够更好地捕捉知识图谱中的语义信息，从而提高检索的准确性和效率。

2）在线检索

在线检索阶段，对于用户输入的问题（如“查找适用于高温环境的发动机缸体文档”），系统首先对 C-HNSW 索引进行分层搜索。查询过程是 C-HNSW 的核心，可以通过迭代过程实现：从最高层的一个随机节点开始，逐步向下层遍历，直到找到查询点的 k 个最近邻。此过程中采用贪婪遍历策略，通过维护一个候选扩展队列 Q 和一个动态最近邻集合 R 来实现快速搜索。

3.3 性能对比实验

为验证所提方法的有效性，在企业 BOM 数据集上进行了系统性实验，与多种主流索引与检索方法进行对比。实验环境：Ubuntu 18.04，128 GB RAM，4 × NVIDIA GeForce RTX GPU。

3.3.1 数据集

本文构建了一个多层级工业 BOM 知识图谱数据集，该数据集具体信息如下。

节点：23 467 个实体（147 个产品、1 253 个部件、9 876 个零件、12 191 个文档）。

关系：57 892 条边（CONTAINS_PART：19 427 条；HAS_FILE：38 465 条）。

属性：平均每个节点 12.4 个属性。

3.3.2 对比方法

仅推理方法：使用大语言模型直接回答问题，而不使用任何检索数据，即零样本和 CoT[20]。

仅检索方法：检索模型从所有文档中提取相关片段，并将它们作为大模型的提示。本文选择了强大且广泛使用的检索模型 BM25[21] 和 vanilla RAG。

基于图的 RAG：这些方法在检索过程中利用图形数据。本文选择了 RAPTOR[22]，GraphRAG，LightRAG。其中，GraphRAG 有全局搜索模式和局部搜索模式；对于 LightRAG，本实验直接选择了最好的混合模式 LightRAG-Hybrid。

3.3.3 评价指标

对于特定的问答任务，使用 Recall@k 来评估前 k 个检索结果中相关项占所有相关项的比例，使用 F1 分数评估在端到端的问答表现，使用检索时间（完成一次检索的平均时间）来评估检索的效率。

3.3.4 实验结果

本文主要使用 deepseek-r1：70b 作为默认的大型语言模型（LLM），并使用 nomic-embed-text[23] 作为嵌入模型来对所有文本进行编码。此外，使用 KNN 进行图增强，以及加权莱顿算法进行社区检测。对于每个检索项 k，在每一层搜索相同数量的项目，默认 $k=5$。结果如表 2 所列。

表 2 不同检索方法性能对比（$k=5$）

Table 2 Performance comparison of different retrieval methods ($k=5$)

基线类型	方法	*Recall*@5/%	*F1-Score*/%	检索耗时/s
仅推理	零样本	23.6	47.7	6.31
	思维链	28.7	54.5	8.15
仅检索	BM25	73.6	69.4	7.79
	Vanilla RAG	80.6	75.6	9.35
基于图的 RAG	GraphRAG_low	82.1	78.3	15.45
	GraphRAG_High	88.0	87.2	17.92
	LightRAG	84.3	80.9	12.83
	RAPTOR	97.5	92.5	19.66
本文方法	HierRAG	**98.6**	**95.8**	**14.87**

表 2 中的结果清晰展示了本文 HierRAG 方法在多个评估维度上的卓越表现。通过对比不同类型的基线方法，可以得出以下关键发现。

首先，从性能指标看，HierRAG 方法以 98.6% 的 Recall@5和 95.8%的 F1 分数显著优于所有对比方法，尤其相比于仅使用推理的方法（零样本和思维链），其召回率提升了 3～4 倍，与传统检索方法（BM25 和 Vanilla RAG）相比也有 18%～30%的性能提升。特别是与当前先进的基于图的 RAG 方法（如 RAPTOR）相比，HierRAG 仍然有 1.1%的召回率提升和 3.3%的 F1 分数优势，同时检索耗时减少了约 32%。

从方法分类角度看，上述结果展示了明显的性能梯度。纯推理方法表现最差（召回率小于 30%），纯检索方法表现中等（召回率为 70%～80%），而基于图的 RAG 方法表现较优（召回率大于 80%）。这证实了语义结构信息对于精确检索的重要性。在基于图的方法中，GraphRAG_High（召回率为 88.0%）优于 GraphRAG_low（召回率为 82.1%），说明高层语义表示对检索质量具有显著影响。

在效率方面，HierRAG 在提供顶级性能的同时保持了合

理的时间开销(14.87 s),仅比 LightRAG 多约 2 s,却比 RAPTOR 快约 4.8 s。这表明本文方法在性能和效率之间取得了较好的平衡,特别适合对检索质量和响应速度均有较高要求的应用场景。

这些结果充分验证了 HierRAG 方法融合层次化知识结构与检索增强生成技术的有效性,尤其在处理复杂多层级知识体系时,能够实现更精确、更高质量的信息检索与生成,为行业内相关应用树立了新的性能标杆。

3.4 数据资源组织与数据资产管理方法

大型制造企业使用产品生命周期管理 PLM 系统管理产品研发数据、产品 BOM 数据、工艺文件与图文档管理,MES 系统管理生产制造数据、质检数据以及各类型知识文档,如图 6 所示。

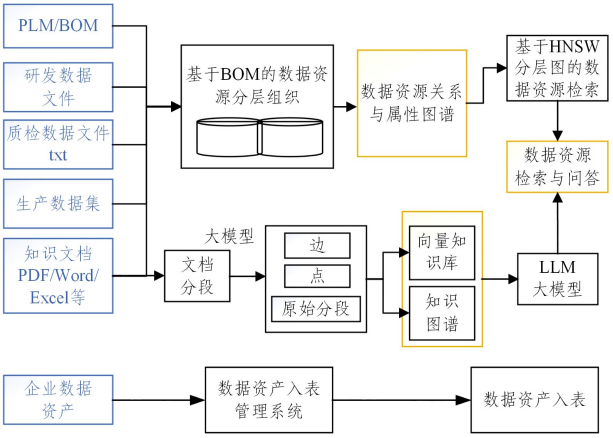

图 6 数据资源与数据资产管理方法

Fig. 6 Data resource and data asset management method

通过 BOM 结构分层组织管理企业数据资源,构建数据资源关系与属性图谱。基于 C-HNSW 建立知识图谱分层树索引,采用 DeepSeek 大模型提供数据资源的检索与知识问答。针对企业较常查询的知识文档,采用 LightRAG(检索增强生成)进行文档分段、分块文档的点和边构建,将创建的文档文本向量化存储至向量知识库和知识图谱,供大模型检索与问答。

企业数据资源如果满足数据资产入表的条件要求,则需要建立数据资产入表系统,系统提供数据资产目录和数据资产登记、数据资产合规性审查、数据资产质量评估、数据资产成本归集与价值评估等功能,帮助完成数据资产入财务报表。

4 数据资产入表管理系统

企业数据资产入表主要流程包括企业数据资源治理与数据资源盘点、数据资产确权登记、数据资产质量评估、数据资产合规性审查、数据资产成本归集与分摊、数据资产价值评估、数据资产的报表与披露。为了对企业数据资产进行有效的数字化管理,设计了企业数据资产入表管理系统,分为数据资源治理与数据资产入表管理两部分。数据资源治理提供企业数据资源(数据文件、数据库等数据集)管理、数据预处理、数据安全权限管理、数据对象与数据模型管理功能,一般企业的数据中台提供了数据资源治理相关功能。数据资产入表管理系统提供数据资源盘点、数据资产登记、数据资产合规审查、数据资产质量评价、数据资产计量、数据资产价值评估和报表披露等业务管理模块,如图 7 所示。

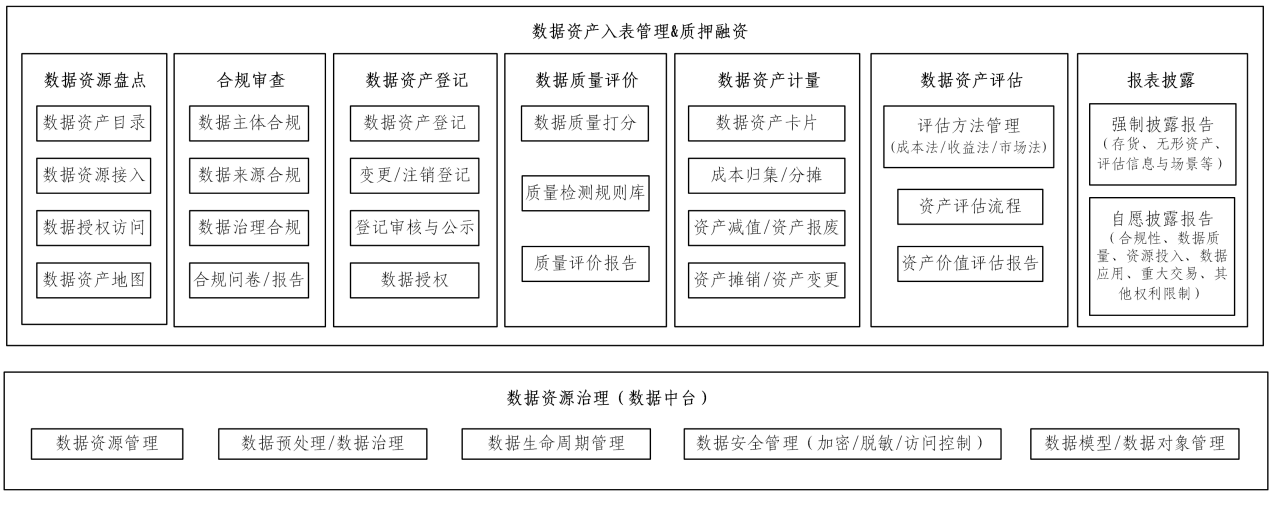

图 7 数据资产入表管理系统

Fig. 7 Data asset management system

数据资源盘点:从全局视角查看企业数据资产,包括数据资产目录、数据资源接入、数据源管理、数据资源申请与授权访问、数据资产地图功能。系统构建和管理单个企业或多个企业的数据资产目录、子目录,通过数据资源接入,接入和同步数据集的数据库表或数据文件至系统服务器管理。系统可视化展示所有的数据资源和数据资产的类别、数据集存储规模统计地图。

数据资产登记:包括数据资产登记、登记变更与注销、数据资产登记查询、资产登记审核与登记公示、证书与核验、数据资产申请授权与授权等功能。数据资产分为数据资源类和数据产品类两种类型进行区分登记。数据资产以卡片登记形式对数据资产进行信息登记管理,并根据数据资产类型绑定数据表或文件,生成的数据资产卡片及其绑定的数据表可以在数据目录下展示和查看。

数据资产质量评价:主要从数据质量的规范性、完整性、准确性、一致性、时效性和可访问性 6 个维度,建立了图 8 所

示的 17 个数据质量评价指标体系。通过设置数据质量检测规则，由检测规则的程序代码对数据资产的数据元素和数据记录进行质量评价指标计算。计算得分 X（见式(1)），得到数据检查问题的统计得分数据，生成数据资产质量评价总体得分和 6 个评价维度得分，从而生成数据资产质量评估报告。

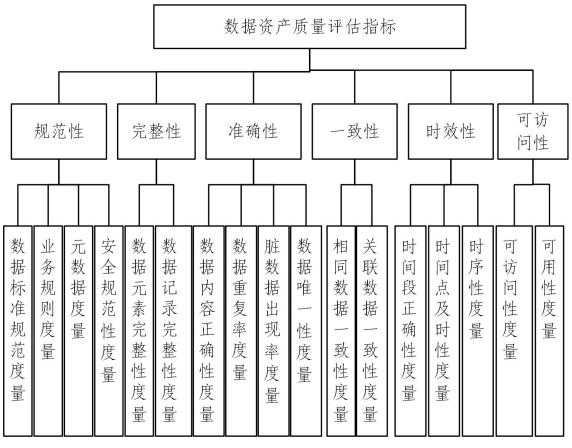

图 8　数据资产质量评价指标

Fig. 8　Evaluation indicators for data asset quality

数据元素或数据记录完整率：

$$X=\frac{A}{B} \tag{1}$$

其中，A 为满足数据质量评价指标（包括数据标准规范、值域规范、重复率度量、内容正确率等）要求的数据集中元素（或数据记录）个数，B 为被评价数据集的所有元素（或数据记录）个数。

数据元素或数据记录非空值率：

$$X=1-\frac{A}{B} \tag{2}$$

这里的 A 为字段必填项为空值的数量，B 为字段必填项的数据元素（或数据记录）总数。

数据时效性(UTs)=

$$\frac{\text{某单位时间内满足及时性需求的记录数}(Ts)}{\text{某单位时间内记录总数}} \tag{3}$$

数据资产合规审查：合规审查主要由外部律师事务所和数据合规评估服务商进行数据资源合规事实与尽调问答，出具数据合规评估报告，一般有效期为半年到 1 年，过期需要重新进行评估。为了方便数据合规审查，设计了系统问卷答题方式辅助评估，从企业经营主体合规、数据来源合规、数据采集与数据治理合规性、数据管理制度与数据安全性、数据资产信息披露等方面进行合规性评测。合规审查问卷题目包括数据产权证书的合规审查、数据治理的合规审查、数据资产范围的合规审查、数据资产归类的合规审查以及披露义务的合规审查 5 个方面共计上百道问卷自测题，以生成数据资产合规报告。

数据资产计量：数据资产计量包括数据资产卡片、成本归集与分摊、资产减值/增值变动管理、资产报废。数据资产价值变动包括资产减值单、资产增值单、数据资产报废。

数据资产评估：数据资产主要评估方法包括成本法、收益法与市场法。系统提供数据资产入表和质押融资的流程管理以及数据资产评估报告管理。

结束语　本文针对军工科研单位因保密分级分块管理导致的数据资源孤岛问题，提出了一种基于企业数据空间的数据资源组织与数据资产管理方案。通过企业数据空间属性图模型，实现多源异构数据资源之间的图节点关联映射，构建了融合产品 BOM 树型结构的数据资源知识图谱，有效整合了产品研发工艺数据、生产制造数据与质检数据及相关文档；提出了 hireRAG、基于 C-HNSW（社群化层次导航小世界图）索引方法、融合图增强聚类算法，能更好地捕捉知识图谱中的语义信息。实验表明，该方法在准确率和检索效率上，相比当前表现优异的图 RAG-RAPTOR 有所提升。此外，本文构建了完善的数据资产入表管理系统。

本研究虽然取得了显著进展，但仍存在以下局限性有待进一步研究。知识图谱构建的准确性受限于大语言模型提取能力，对于特殊领域术语和复杂关系的准确标识仍有提升空间。当前的 C-HNSW 索引构建对于超大规模数据（如上亿节点的知识图谱）的高效索引仍面临计算复杂度高的挑战。多源数据的语义融合与跨层级推理能力有待增强，尤其是在处理细粒度语义差异和多模态数据上，还需更多探索。

未来工作将重点探索跨层级推理能力，融合多模态数据分析技术，进一步提升复杂 BOM 结构的数据资源智能检索性能。

参 考 文 献

[1] FRANKLIN M，HALEVY A，MAIER D. From databased to dataspaces：A new abstraction for information management [J]. SIGMOD Record，2005，34(4)：27-33.

[2] FAN S H，HOU M S. Dataspace：A New Data Organization and Management Model [J]. Computer Science，2023，50(5)：115-127.

[3] DITTRICH J P，SALLES M A V. iDM：A unified and versatile data model for personal dataspace management [C]// International Conference on Very Large Data Bases(VLDB'06). 2006：367-378.

[4] WEN B L，JIAO S J，GUO J. Research on Data Organization Method Based on Enterprise DataSpace [J]. Computer Technology and Development，2020，30(12)：56-60.

[5] LI Y K，MENG X F，ZHANG X Y. Research on data space technology [J]. Journal of Software，2008，19(8)：2018-2031.

[6] YAN X F，YU P S，HAN J W. Graph Indexing：A frequent structure-based approach [C]// Proceedings of the 24th International Conference on Management of Data(SIGMOD 2004). New York：ACM，2004：335-346.

[7] HE H，SINGH A K. Closure-Tree：An index structure for graph queries [C]// Proceedings of the 22nd International Conference on Data Engineering(ICDE 2006). Dallas：IEEE Computer Society，2006.

[8] HOLDER L，COOK D，DJOKO S. Substructure discovery in the subdue system [C]// Proceedings of the AAAI Workshop of Conference on Knowledge Discovery in Databases. Menlo Park：AAAI，1994：169-180.

[9] JIANG H L，WANG H X，YU P S，et al. GString：A novel approach for efficient search in graph databases [C]// Proceedings

of the 23rd International Conference on Data Engineering(ICDE 2007). Dallas: IEEE Computer Society, 2007: 566-575.

[10] VAN BRUGGEN R, BATON J. Learning Neo4j 3. x-second edition [M]// Birmingham: Packt Publishing, 2017: 33-36.

[11] ELSAYED I, MUSLIMOVIC A, BREZANY P. Intelligent data spaces for science C WSEAS [C]// International Conference on Computational Intelligence, Man-Machine Systems and Cybernetics. 2008: 94-100.

[12] YANG D, SHEN D, NIE T, et al. Layered graph data model for data management of dataspace support platform [C]// International Conference on Web-Age Information Management. Berlin: Springer, 2011: 353-365.

[13] FRANKLIN M, HALEVY A, MAIER D. A first tutorial on dataspaces [C]// Proceedings of the VLDB Endowment. 2008: 1516-1517.

[14] XUE Q, ZENG X Q, OUYANG Z Y. Discussion on Key Issues of Data Asset Inclusion the Background of Digitalization and Intelligence [J]. The Chinese Certified Public Accountant, 2024, 10: 105-113.

[15] EDGE D, TRINH H, CHENG N, et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization [J]. arXiv: 2404. 16130, 2024.

[16] GUO Z R, XIA L H, YU Y H, et al. Lightrag: Simple and fast retrieval-augmented generation [J]. arXiv: 2410. 05779, 2024.

[17] GROVER A, LESKOVEC J. node2vec: Scalable feature learning for networks [C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 855-864.

[18] XU X, YURUK N, FENG Z, et al. Scan: A structural clustering algorithm for networks [C]// Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2007: 824-833.

[19] WANG C, WANG F, ONEGA T. Network optimization approach to delineating health care service areas: Spatially constrained Louvain and Leiden algorithms [J]. Transactions in GIS, 2021, 25(2): 1065-1081.

[20] KOJIMA T, GU S S, REID M, et al. Large language models are zero-shot reasoners [J]. Advances in Neural Information Processing Systems, 2022, 35: 22199-22213.

[21] ROBERTSON S E, WALKER S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval [C]// Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval(SIGIR'94). London: Springer, 1994: 232-241.

[22] SARTHI P, ABDULLAH S, TULI A, et al. Raptor: Recursive abstractive processing for tree-organized retrieval [C]// The Twelfth International Conference on Learning Representations. 2024.

[23] NUSSBAUM Z, MORRIS J X, DUDERSTADT B, et al. Nomic embed: Training a reproducible long context text embedder [J]. arXiv: 2402. 01613, 2024.

LI Minbo, born in 1970, Ph.D, associate professor, is a member of CCF (No. 53905M). His main research interests include industrial big data and industrial AI.

WANG Shaohua, born in 1984, senior engineer. His main research interests include industrial Internet, AI and supply chain collaboration and data services.

(责任编辑:何杨)