Q/ZXZQAI-002-2023

中 信 证 券 股 份 有 限 公 司 企 业 标 准

Q/ZXZQAI-002-2023

中信证券知识图谱构建标准

Construction Standard for Knowledge Graph in CITICS

2023-10-18 发布　　2023-10-25 实施

中信证券股份有限公司　　发布

目　次

前　言

本标准依据 GB/T 1.1-2020《标准化工作导则 第一部分：标准化文件的结构和起草规则》给出的规则起草。

本标准由中信证券股份有限公司提出。

本标准由中信证券股份有限公司归口。

本标准起草部门：中信证券股份有限公司信息技术中心。

本标准主要起草人：方兴，岳丰，刘殿兴，陈辉华，苑博文，张天骁，孙少卿，余子安，陈昭铭。

引　言

知识图谱（Knowledge Graph），用以描述真实世界的各类实体以及它们之间的关联关系，也可以简单地把知识图谱理解成多关系图 （Multi-relational Graph）。

知识图谱具有关系的表达能力强，结构友好等特点，广泛运用于知识管理，智能风控和个性化推荐等应用场景，近年来在金融领域得到了大量的应用。知识图谱的构建是实现知识图谱应用的关键环节之一，对于提高金融企业的竞争力和创新能力具有重要意义。

为了规范和统一中信证券内部知识图谱应用的开发流程，确保知识图谱构建的质量和可靠性，提高开发效率，降低开发成本，特制定本《中信证券知识图谱构建标准》。本标准旨在为公司内部知识图谱应用开发人员提供一套统一的构建标准和最佳实践，以确保知识的可靠性、可维护性和可扩展性。

本标准适用于公司内部所有知识图谱的构建工作，包括知识收集、知识建模、知识获取、知识融合、知识评估、知识推理、知识存储等环节。通过遵循本标准，能够更加高效地开发出高质量的知识图谱应用，为企业的业务发展提供有力支持。

中信证券知识图谱构建标准

1 范围

本标准规定了中信证券股份有限公司知识图谱构建步骤的知识收集、知识建模、知识抽取、知识融合、知识存储，知识查询和图谱可视化以及知识计算和推理的规范。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 42131-2022 人工智能 知识图谱技术框架

GB/T 5271.17-2010 信息技术 词汇 第17部分:数据库

GB/T 22239-2019 信息安全技术 网络安全等级保护基本要求

GB/T 35273-2020 信息安全技术 个人信息安全规范

YD/T 4044-2022 基于人工智能的知识图谱构建技术要求

3 术语和定义

3.1 知识图谱 Knowledge Graph

一种以结构化的形式描述客观世界中概念、实体及其关系的方式。

3.2 资源描述框架 Resource Description Framework

一种能够对结构化的元数据进行编码、交换和再利用的基础架构，简称为 RDF。

3.3 三元组 Triplet

用来表示实体、属性和实体之间关系的数据结构。它包括三个部分：头实体、关系和尾实体。

3.4 模式 Schema

是数据库的逻辑结构。具体到图数据库，包括知识图谱的实体、关系、属性、约束等元素的定义和描述。

3.5 实体 Entity

存在或者可能存在的任何具体或抽象的事务,包括这些事物间的关联。

3.6 关系 Relation

具有相同属性的各实体值的集合以及这些属性。

3.7 属性图 Property Graph

是实际业务中使用最广泛的图数据建模方式，主要包括三种元素：点，边和属性，用点和边表达拓扑关系，在点和边上附着属性来存储数据。

3.8 网络本体语言 Ontology Web Language

是一种用于定义和实例化 Web 本体的语言，简称为 OWL。OWL 本体包括对类、属性及其实例的描述，允许逻辑推理。

4 知识图谱构建步骤

4.1 知识收集

知识收集明确收集知识图谱的数据来源，为后续知识图谱构建做数据准备。知识收集包括知识采集和知识导入。

知识采集明确业务场景所需的数据来源，如业务条线存放数据库中的结构化数据，日志文件，JSON 文件等半结构化数据，以及文档、图像、语音、视频等非结构化数据等；运用多种技术方法进行知识采集，包括但不限于 HDFS，HBase，JDBC，ODPS 连接，支持 ES、本地文件（csv、json 数据格式）、MySQL、DB2、Oracle、Postgresql、Sqlserver 和 Hive 等知识抽取。

针对不同的数据来源应制定不同的采集策略：

1）对于内部数据，应对业务数据进行脱敏转换后进行使用；

2）对于互联网公开数据，应利用爬虫工具或自行开发爬虫程序爬取数据，并经过数据解析、数据清洗后进行使用；

3）对于供应商数据，应通过 API 接口或者数据文件方式进行数据采集。

知识导入将采集到的数据导入到图数据库，关系型数据库或内存数据库等数据存储介质中，以用于后续的分析。

4.2 知识建模

知识建模是指基于行业的应用属性和业务需求，将业务知识转化成图谱形式表达，进行业务抽象，业务建模和模式定义。知识构建包括实体定义、关系定义、属性定义等。根据实际情况，可以选择知识图谱构建的两种方式之一：自上而下的图谱构建和自下而上的知识构建。自上而下的构建方法是从结构化的数据源中提取本体和模式层，然后将数据和模式层映射到本体和模式层；自下而上的构建是从非结构化数据中抽取 RDF 三元组直接加入到知识库。

知识建模过程首先定义实体，配置该实体集的概念标签、图标类型、ID 列等；其次，定义实体之间的关系，关系可以是自环的，单向和双向的。关系上可以配置属性，也可以不配置任何属性。再次，进行属性的定义，确保在冗余度最低条件下满足业务应用和可视化展示。最后，可以引用其他领域已有的关系、实体定义，通过已有的配置模板和公共模板等进行知识建模，提高建模效率。

4.3 知识抽取

知识抽取是从原始数据中抽取图谱知识，包括实体抽取、关系抽取、事件抽取、属性抽取等，并将结果更新或连接到知识图谱中。

实体抽取，也称为命名实体识别，是指抽取原始数据中的信息元素，如人名、组织名、地理位置、金额值等；关系抽取，是指抽取实体和实体之间的某种语义关系；事件抽取，是指抽取有用的信息事件；属性抽取是指抽取出与该实体相关的属性及其属性值。

知识抽取主要有三类方法，根据具体情况具体使用：

1）基于模板的方法：通过预先定义或者学习得到的模板进行抽取和判别；

2）基于统计机器学习的方法：利用机器学习或深度学习技术进行抽取；

3）面向开放域的方法：主要针对海量数据进行抽取。

4.4 知识融合

知识融合是将多源异构的知识进行正确性判断，并进行融合，去除冗余和消除冲突等，形成统一的知识图谱。知识融合包括实体对齐，属性对齐，指代消解等。

1）实体对齐，是指数据融合时，保证同一实体在知识图谱中唯一，主要方法有基于相似性计算、基于关系推理以及基于知识表示学习的实体对齐方法等；

2）属性对齐：是指数据融合时，将不同数据源的属性进行匹配，确保具有一致的语义，主要方法有基于相似性计算，基于语义信息以及基于规则的属性对齐方法等；

3）指代消解：是指将代表同一实体的不同指称划分到一个等价集合的过程，主要方法有基于句法和基于语料库的指代消解方法等。

4.5 知识存储

知识存储是将知识图谱的概念层和数据层进行有效的物理存储，实现以图数据库为核心，多种存储介质共存的多样底层存储，支持对大规模图数据的有效管理和并行计算。

图数据库存放属性图、RDF 等图数据模型对应的数据，支持增删查改的功能和事务能力。

不同的知识图谱应用统一存放在分布式图数据库集群中，提供不同的图计算入口，实现基于角色的访问控制、加密、多租户、高可用性、备份和还原的功能。

4.6 知识查询和图谱可视化

知识查询是通过形式化的查询语言为用户提供检索和查询知识的接口。关系型数据库的标准查询语言是 SQL，图数据库的标准查询语言是 GSQL，具体而言是基于 openCypher 图查询语言。系统通过图谱可视化的方式展示给管理者和使用者查看，拥有实体查询、关系查询以及语句查询等功能，可以自主查看不同知识构建的知识图谱。支持 Grid，Dagre，Circle，Concentric，BFS，Cose，Klay，Spread 和 Cola 等多种图谱布局。

4.7 知识计算和推理

知识计算和推理是指基于 OWL 进行本体推理和基于图计算关联性推理。系统内置丰富的图算法，包括深度图算法、分布式图计算算法、图表示算法等，方便用户实现知识计算和推理，进行属性补全和关系预测等任务。

参考文献

[1] GB/T 42131-2022 《人工智能 知识图谱技术框架》

[2] GB/T 5271.17-2010《信息技术 词汇》

[3] GB/T 22239-2019《信息安全技术 网络安全等级保护基本要求》

[4] GB/T 35273-2020《信息安全技术 个人信息安全规范》

[5] YD/T 4044-2022《基于人工智能的知识图谱构建技术要求》