

[数据资产]

基于电力数据中台的资产地图 自动化构建技术研究

文/盛 爽 黄佩卓 刘 洋 陈佳琦
国家电网有限公司大数据中心

研究基于电力数据中台的资产地图自动化构建技术实现方法，通过该方法实现对不同用户提供一体化的应用数据入口，采用知识图谱结合元仓的方式为构建方案核心，结合AI算法能力，通过深度理解用户需求，从数据元仓和数据图谱中检索出相关内容结果返回给用户。

电力企业数据中台建设现状

电力数据具有以下三方面鲜明特点：一是覆盖范围广，包括发电运行数据、电网运行数据及用户用电数据等；二是价值密度高，数据真实性高，且贯穿电力系统“发输变配用”各个环节，能够全面真实反映宏观经济运行情况、各产业发展状况、居民生活情况和消费结构等；三是实时准确性强，电力行业自动化水平较高，数据实时性和真实性高，具有很强的独占性和不可替代性。

为充分挖掘数据价值，发挥数据作为生产要素的价值，引入数据中台技术体系。数据中台基于统一的数据接入、存储计算、数据分析、数据服务、数据资产管理及数据运营管理等能力，面向业务和用户，提供数据、平台及工具三方面服务，持续推动数据业务化，加快数据赋能业务。

构建数据资产地图

1. 数据资产地图的作用和价值

随着数据中台的全量归集，支撑上层应用的增加，数据作为生产要素的价值正在得以发挥和释放，数据正在加速资产化的过程，同时业务人员、研发人员及管理人员对于数据的需求也不一样，更加关注对于数据内容的萃取，采用简单的目录结构或树形结构

来支持数据应用已经不能满足用户需求。

因此，为了满足“数据是什么？数据在哪里？数据如何取及数据怎么用？”的关键问题，构建数据资产地图，数据资产地图是随着数据中台建设、运营持续沉淀业务逻辑，实现数据与业务融合的载体，面对不同类型用户提供一体化的应用数据窗口，作为建设和应用的桥梁作用和价值越来越重要。关键问题见表1。

表1 关键问题汇总表

需求	内容
数据是什么？	需要知道数据形式是怎么样的？比如结构化或非结构化；数据定义是什么？比如技术、业务及管理；数据的使用特性如何？比如频度、评价、质量及安全级别等；数据包含什么信息？比如时间周期、业务分类及信息密度
数据在哪里？	数据存放的位置在哪里？与业务相关的数据还有过去曾经用过且值得参考的数据有哪些？与数据使用相关的辅助性工具在哪里？等等
数据如何取？	提供一系列工具及辅助信息，帮助实现数据采样、数据查询、数据下载及数据安全沙箱建立各种使用数据的方式
数据怎么用？	帮助使用者更快速便捷地获取数据内在信息，例如数据摘要信息、数据快速分析、数据报表工具、数据分析模型及知识帮助指导等

2. 构建数据资产地图的难点

构建数据资产地图不同于简单的资源目录，不仅是将表名、基础字段信息等基础信息进行展示，

还要萃取数据内容，如全链路血缘、增加业务信息、访问次数及应用成效等，数据资产地图的应用价值更大，同样建设难度也随着增长。

数据资产地图建设面临三大难点：一是数据如何采集？虽然数据汇聚到数据中台，但是关于数据内容还是分散在数据中台的各个组件中或者前端应用中，如组件运行数据、调度任务数据及前端调用数据等，如何进行有序采集，不影响正常应用；二是获取到的数据如何组织？因为用户对于数据内容的需求具有不确定性，如何组织能够保障对于数据内容需求的全覆盖及关联查询等；三是数据资产地图对于数据时效性要求较高，解决构建自动化采集、加工和展示链路等关键问题。

数据资产地图自动化构建技术方案

1. 设计思路

数据资产地图业界基本依托于数据元仓构建，但是仅依托元仓构建会导致两个根本性能问题：一是数据关联困难，所有关联关系都是事先梳理的，对于复杂的网状关联关系难以处理，性能无法满足；二是扩展困难，数据元仓本身还是设计出来

的，一旦事先有没有考虑到的内容，如事先并未规划基层赋能的描述字段，后续扩展就难以进行。

因此，为了弥补数据元仓的不足，采用知识图谱结合元仓的方式构建自动化技术方案，元仓用于解决基础描述数据的加工和采集问题，知识图谱用于非结构化数据内容的提取，构建不同数据实体的关联关系，支持对于数据内容的萃取和加工等，从而构建起从数据采集、加工到展示的全过程自动化处理过程。

2. 技术路线

数据资产地图共分为数据层、自动化处理层和展示应用层三层。数据层主要是由数据中台存储的结构化、非结构化数据；自动化处理层包含元数据管理和数据图谱加工两大核心功能，元数据管理主要负责对于结构化元数据的采集及加工，数据图谱加工主要负责对于非结构化数据的知识提取、数据图谱的加工和计算；展示应用层主要依托元仓及数据图谱提供面向用户的应用，主要包含数据推荐、语义搜索、关联分析及数据词条等核心功能，形成具备处理结构化数据、非结构化数据以数据内容为核心的新一代数据资产地图。数据资产地图如图1所示。

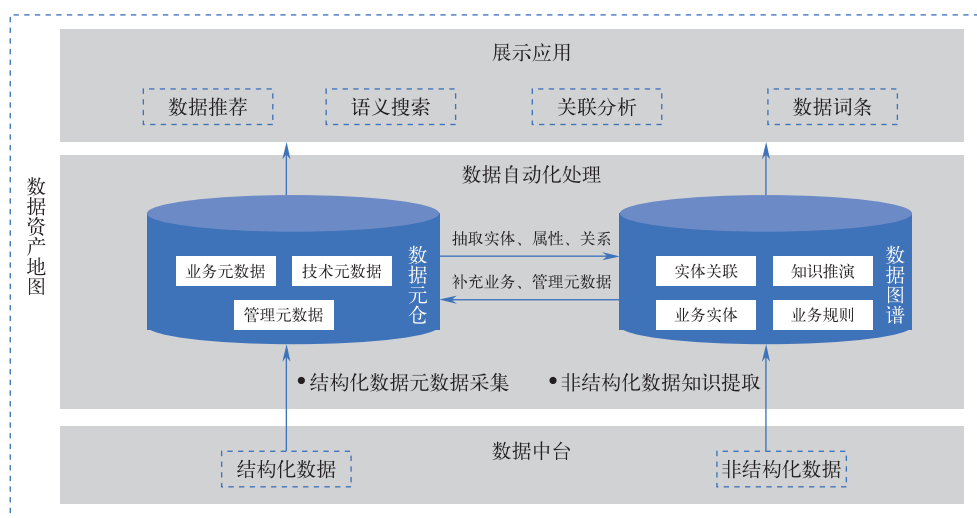


图1 数据资产地图

3. 关键技术实现

(1) 数据采集

针对结构化数据采集，主要通过连接数据中台元仓的方式进行采集，重点在于非结构化数据的采集，包含链接数据、表格数据及文本数据等，数据采集到数据图谱进行存储。数据采集如图2所示。

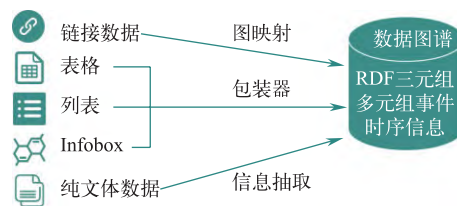


图2 数据采集

采用不同方法将不同结构化形式的数据转换为三元组。结构化数据处理采用D2R技术，把关系型数据库中的数据转换为RDF数据（linked data），半结构化数据采用包装器的方式进行处理，并结合规则处理及深度学习方法还原为结构化的数据；非结构化数据采取基于BERT和强化学习的关系分类算法，通过训练好的实体识别模型、实体链接模型和关系抽取模型，实现对海量的数据识别链接和抽取。抽取得到的知识不断完善动态知识图谱数据库。

(2) 数据组织

为了实现对于用户数据应用需求的全面覆盖，对于采集的数据利用数据元仓和数据图谱进行存储，数据元仓按照采集的对象进行组织，如物理表、指标、报表及标签等保障对于相关基础技术元数据的全面采集；数据图谱按照业务专题、业务对象及业务活动进行组织，通过业务视角对采集的数据进行组织，开展对于数据内容的萃取，从而实现对于数据内容的直接应用。数据组织如图3所示。

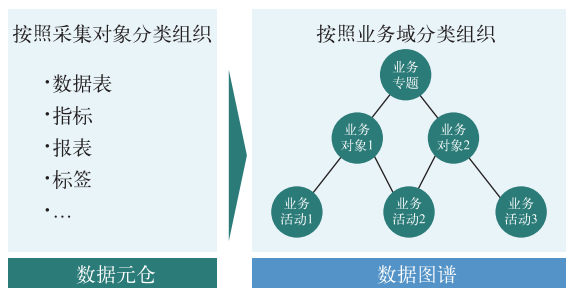


图3 数据组织

(3) 数据自动化更新

数据自动化更新是保障数据资产地图的生命力所在，结合电力企业业务特性和人工智能算法，通过语义推断、数据特征分析及多表关联实现从数据库表到知识图谱的构建，从而实现根据数据中台建设持续对数据资产地图进行更新、完善，实现数据资产地图的动态成长。

1) 将表字段映射到本体库语义类型（含实体及关系谓词），如图4所示。

2) 发现不同数据库中相似数据集合，如图5所示。

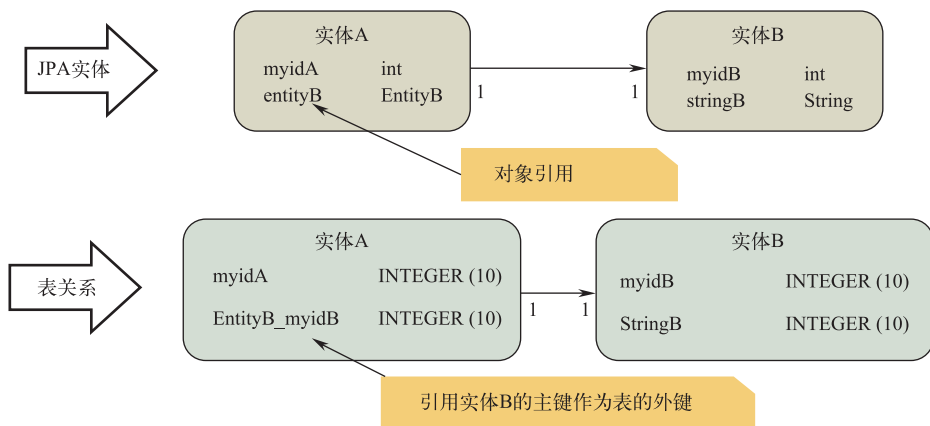


图4 表字段映射

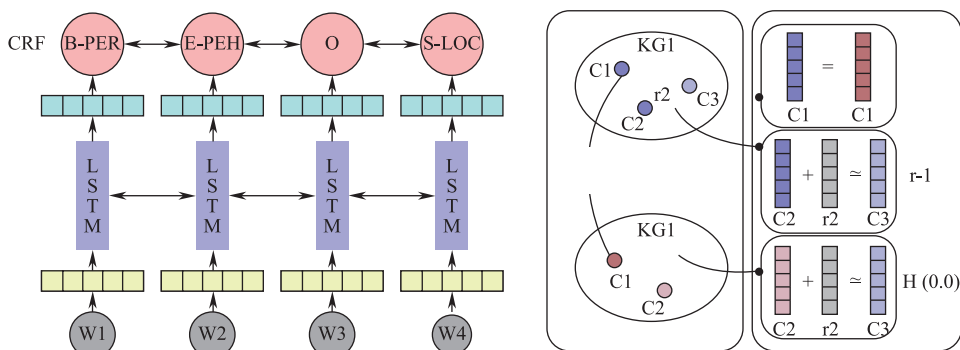


图5 相似数据集合

数据的更新包含新实体数据的发现和已有实体数据的更新,可以通过基于特征词库增量学习的领域知识更新技术确定更新的数据集合。即通过对增量更新数据集的特征词库与初始数据集特征词库的相似度分析,确定不同数据中相似数据集合数据作

为增量更新数据,其他数据作为新的图谱知识进行处理。

3) 自动发现多表关联关系,建立跨表实体对齐,如图6所示。

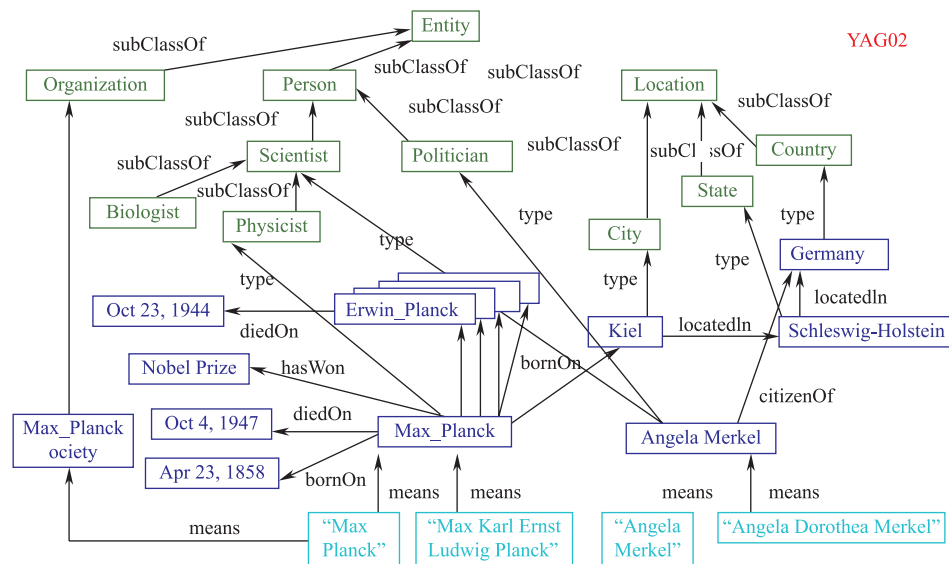


图6 关联关系

实体对齐方法一般通过属性相似度匹配的方式实现,利用有监督学习的机器学习模型,如决策树、支持向量机及集成学习等。依赖实体的属性信息,通过属性相似度,进行跨平台实体对齐关系的推断。南京大学 Websoft 组在 ISWC 2022 中提出了一种面向动态增长知识图谱的持续实体对齐方法。在图谱数据自动化更新场景下,利用图神经网络对知识图谱的实体与三元组进行编码,并根据当前时刻的图谱的变化来调整对齐结果的方式,解决多表关联关系及跨表实体对齐。

结束语

采用数据元仓结合数据谱图的方式作为构建数据资产地图的核心能力,充分借鉴了数据图谱逐步演进便于横向扩展的优势,同时结合了人工智能算法的能力,从而实现了针对数据内容的应用,包含针对数据内容的语义搜索和推荐,以知识卡片的形式输出数据,实现对于不同数据内容之间的关联分析等高级功能,从而将数据与业务不断深化融合,拉近了数据与业务的距离。

将知识图谱技术引入到数据资产的管理是未来

发展的趋势,充分应用知识图谱对于知识抽取、知识加工及知识推演的能力,助力数据中台对于汇聚的全量数据进行进一步的知识挖掘和价值发挥。EA

(收稿日期:2022-10-24)

参考文献

- [1] 刘祥. 领域知识学习中的马尔可夫逻辑网应用研究[D]. 哈尔滨: 哈尔滨工业大学, 2015.
- [2] 白海燕, 梁冰. 利用D2R 实现关系数据库与关联数据的语义模式映射[J]. 现代图书情报技术, 2011: 7-8.
- [3] 陈优敏. 知识图谱构建方法探究[J]. 中国新通信, 2019(5).
- [4] 刘烨宸, 李华昱. 领域知识图谱研究综述[J]. 计算机系统应用, 2020(6).
- [5] 邵领. 基于知识图谱的搜索引擎技术研究与应用[D]. 成都: 电子科技大学, 2016.
- [6] 杨帆, 李靖, 王坚敏, 等. 基于自然语言句法分析的电力变压器运行规范知识图谱构建[J]. 电气应用, 2020, 39(11): 18-23.