一种基于知识图谱的数据检索与可视化方法

李　磊[1],鲁兴河[1],康警予[2],陈　忠[1],朱　峰[1]
(1.中国电子科技集团公司第二十八研究所,江苏 南京 210007;
2.陆军装甲兵学院 演训中心,北京 100093)

摘　要:在现有的数据基础上通过本体构建工具建立起包括组织、人员及设施等本体概念,通过对本体间建立关系,利用本体概念下的参数关系构建实体知识图谱,为信息系统的数据资源利用提供技术保障。同时对需要支持检索的本体模型建立索引,并利用图谱间的关联关系,直观、高效地向用户展现检索结果,满足检索结果的个性化和智能化,从而更好地为决策提供支撑。

关键词:知识图谱;本体模型;数据可视化
中图分类号:TP391　**文献标志码**:A　**文章编号**:1008- 1739(2020)05- 61- 4

A Data Retrieval and Visualization Method Based on Knowledge Graph

LI Lei[1], LU Xinghe[1], KANG Jingyu[2] CHEN Zhong[1], ZHU Feng[1]
(1.The 28th Research Institute of CETC, Nanjing 210007, China;
2. Military Exercise and Training Center, Army Academy of Armored Forces, Beijing 100093, China)

Abstract: Based on the existing data, the ontology concept including organization, personnel and facilities through the ontology construction tool is established. By establishing the relationship among the ontology, the parameter relationship of the ontology concept is used to construct the entity knowledge graph for providing technical support for the data resource utilization of information systems. At the same time, the ontology model that needs to support retrieval is indexed, and the association relationship among the graphs is used to display the retrieval results to the user intuitively and efficiently, so as to satisfy the individualization and intellectualization of the retrieval results, thereby better supporting the decision-making.

Key words: knowledge graph; ontology model; data visualization

0 引言

知识图谱是在传统知识工程的基础上以及语义 Web 的发展中孕育并发展而来的知识表示技术[1],旨在描述客观世界的概念、实体[2-3]、事件及其之间的关系。知识图谱亦可被看作是一张巨大的图,图中的节点表示实体或概念,而图中的边则由属性或关系构成[4]。知识图谱已被用来泛指各种大规模的知识库,知识图谱技术逐步渗透到各个领域[5-6]。同时,随着作战保障和业务处理系统稳步发展,各类数据资源逐渐丰富,各领域军事应用需求的不断增长,作战指挥、作战保障和日常业务处理信息系统建设投入不断加大，各类作战保障和业务处理信息系统规模逐步扩展，积累形成了一批可用、实用的信息资源,成为构建知识图谱的重要支撑。

收稿日期:2019-12-18

1 知识图谱构建方法

基于本体模型构建知识图谱方法流程图如图 1 所示。

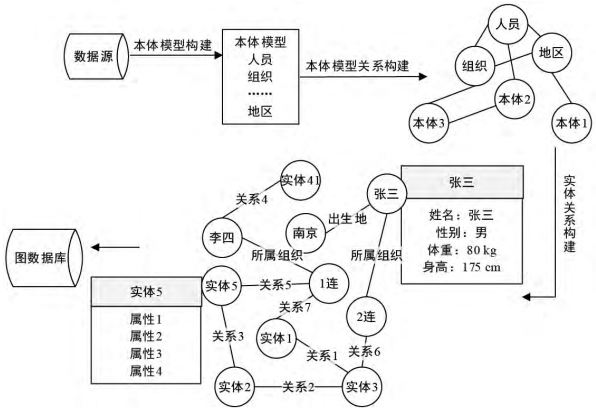

图 1 基于本体模型构建知识图谱方法流程图

首先将存储在关系型数据库内数据构建成多个本体模型,

然后利用原有数据库表内字段之间的关系构建本体模型间关系，完成现有数据架构下的数据关系图谱构建，接着获取本体模型下所有的实体数据，利用本体模型关联参数构建实体数据关系，形成实体数据关系网，最终将本体模型、实体数据和关系按照邻接表的方式存入到图数据库中，并实现基于图结构的索引技术，提高对图数据库中节点和关系的查找速度。

1.1 本体模型构建方法

存在于数据库中的各基础和业务数据通常包含各种本体模型，如人员、设施及地名等，这些本体多以表为单位进行存储，本体之间的关系通过主外键进行关联。本文提供了一种配置化的本体模型构建工具，此工具首先获取数据库用户空间下的所有表结构，用户根据表的存储信息构建本体模型，再通过字段关联将关联信息加入到本体模型中，形成了多个独立的本体模型，使数据库使用人员能够迅速获取到数据库内的数据结构信息，然后根据需求进行数据访问。

1.2 本体模型关系构建方法

人员本体与组织本体模型关系如图 2 所示。

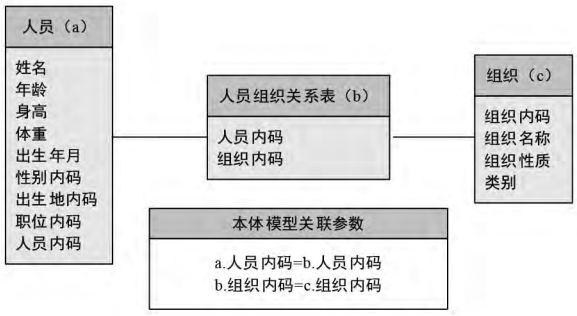

图 2 人员与组织本体模型关系

本体模型构建完成后会形成多个独立的本体，这些本体之间存在诸多关系，这些关系在数据库中一般通过关联表实现，如建立一张人员与组织的关系表，表结构为人员内码和组织内码，一行数据就表示了某个人员的所属组织。本文提出了一种基于知识图谱的本体模型关系构建方法，步骤如下所示：

步骤 1：选择需要建立关系的多个数据模型，模型的数量不定，如 M_1，M_2，M_3，……M_N。

步骤 2：选择每个模型的关联字段，建立字段之间的关系，此关系可以是相等关系，如内码相等，也可以是其他复杂关系，如子字符串、取模计算等。

步骤 3：将本体模型关系存入图数据库中，存入的信息包括本体模型的字段信息、参与关联的模型名称、关联的参数。

1.3 实体关系构建方法

在本体模型关系构建之后，就可以根据关系参数构建实体关系，构建方法如下：

步骤 1：对参与构建关系的每个本体模型通过统一的数据访问接口获取所有的数据。

步骤 2：由数据库表中对于表的注释和对于表中字段的注释，将实体数据由英文属性名转为中文属性名，如组织实体中英文字段“zzmc”转为中文字段名称“组织名称”，使所有数据表现更为直观。

步骤 3：将所有本体模型的实体数据存入图数据库中。

步骤 4：利用本体模型的关联参数构建实体关系，例如对于组织、人员组织关系和人员这 3 个本体，如果某个组织实体的组织内码等于人员组织关系实体的组织内码且此人员组织关系实体的人员内码等于某个人员实体的人员内码，则在此组织实体和人员实体之间构建组织下属人员关系。

步骤 5：重复步骤 1～步骤 4，直至所有的本体模型关系都完成对应实体关系的构建。

2 数据检索与可视化方法

基于知识图谱的数据检索[7-8]是指通过语义检索[9]，对大量数据进行过滤、分析和管理，实现搜索数据的结构化并且提供详细的主题相关信息，有利于建立数据间知识体系，理解各种实体概念以及它们的关联。本文创新之处在于构建了一套完整的从知识图谱构建到检索展现的系统，实现了对结构化数据进行本体构建、实体抽取、索引构建和检索结果展现的全流程可视化操作。

2.1 方法体系架构

数据检索与可视化方法体系架构如图 3 所示，自底向上可分为数据源层、图谱及索引构建层和外部应用层。

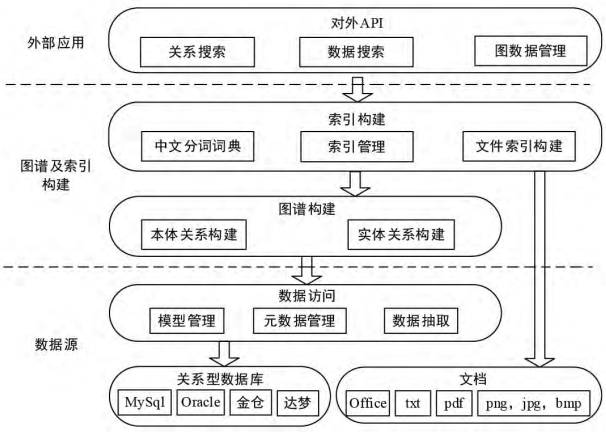

图 3 数据检索与可视化方法体系架构

① 数据源：支持对关系型数据库和文档类数据进行数据检索，通过数据访问模块可实现对包括金仓、达梦等国产化数据库在内的多种数据库系统进行模型管理、元数据管理和数据抽取，为上层知识图谱的构建提供数据基础。

② 图谱及索引构建：首先是创建中文分词词典，需要根据数据特点进行词典的扩充，提高分词的准确率，然后对文档型

数据进行全文索引，将索引数据存入索引库中，最后对需要支持搜索的本体模型相应的字段构建索引。

③ 外部应用：为提高本方法的适用场景，采用了对外 API 提供应用系统调用接口。数据搜索接口可实现按照索引类别对知识图谱数据和文档类数据进行快速检索；关系搜索可实现对实体间关系的查询；图数据管理可实现对知识图谱内数据进行修改和更新的能力。

2.2 索引构建

结合自身数据存储格式的特点，通过建立全图索引和顶点内索引，有效支撑了图的遍历查询，实现了对关系和数据的快速查询。大部分对图的查询都是基于某个属性查询符合条件的顶点和边的特点，首先对图中的顶点和边的属性进行全图索引。一般查询条件分为 2 种：一种为确定性查询，如判断字符串和数值的相等、大于或小于；另一种是范围性查询。为提高建立索引的速度，针对这 2 种情况分别提供复合索引和混合索引 2 种不同的索引建立方法，同时提供建立联合索引的方法，即在多个属性上建立关联索引，通过对属性值的联合查询能够快速定位到符合条件的边和顶点。

在对文本建立索引时，必须指定建立的是全文索引还是字符串索引。全文索引在建立时会将字符串值进行标签化，标签化的方法用户可以自己指定，默认情况下会将字符串以非字符文本进行切割，然后去除长度小于 2 的标签。全文索引支持某个指定的子串、以某个子串开头和结尾及符合某个指定的正则表达式 3 种查询方式。字符串索引不对文本进行标签化，以整个文本的值建立索引，其支持的查询方式有文本相等、文本不等、文本以某个给定字符开头和结尾、文本匹配给定的正则表达式 4 种。通过对文本建立不同的索引方式，可以大大减少建立索引的开销，同时也会加快对文本的查询。

顶点内索引是针对每个顶点的数据进行索引，在顶点的存储模型中，每个顶点都存储了它所有的相邻边，相同标签的边会存储在一起，通过指定相同标签表的排序属性，可以按照某个属性值对边进行降序或者增序排序，参与排序的属性值也可以有多个，此时会在第 1 个属性值相等的条件下按第 2 个属性值进行排序，以此类推，这样在根据属性值查找复合条件的边时可以通过二分查找、递归查找等算法进行加速。

2.3 数据可视化展现

根据所获取的各种数据特点，按照数据分析和展现的需要，对数据进行整合后通过键值对、表格、图片及知识图谱等多种方式实现数据的智能可视化展现，流程如图 4 所示。

设计方法如下：

步骤 1：用户输入查询词和查询条件，查询模块将查询内容发送到后端，交由图数据库根据遍历策略遍历出所有相关的节点，获取所有查询到的数据。

步骤 2：根据所获取的节点所属的索引类别，按照数据自身特点以及向用户高效直观展现检索结果的需要，选择合适的展示方式。

步骤 3：将获取的数据在对应的展示组件进行展示，以搜索装备为例、装备属性以键值对展示、装备性能以表格形式展示、装备照片以图片展示及装备间关系以知识图谱的方式进行展示。

步骤 4：点击知识图谱中展现的节点，可以链接到搜索节点的子节点或父节点页面，利用知识图谱的方式获取更多搜索结果。

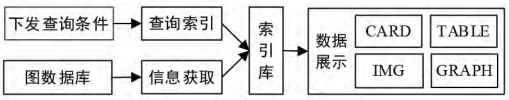

图 4 数据可视化模块设计

3 结束语

通过表间字段关系构建数据库内本体模型关系，并依据本体模型关系构建实体数据关系，同时采用邻接表的方式存储图数据中的顶点和边信息，有效地节省了存储空间，最终提供统一的查询入口客户端，利用键值对、图、表和图谱等多种方式实现数据的展示，支持不同场景下作战数据的可视化，满足检索结果的个性化、智能化需求。本方法能够实现对任一关系型数据库内数据本体和实体构建关系，快速构建知识图谱，辅助业务人员掌握数据关系，充分实现对现有数据的利用。

参考文献

[1] 徐增林,盛泳潘,贺丽荣,等.知识图谱技术综述[J].电子科技大学学报,2016,45(4):589-606.

[2] 李涛,王次臣,李华康.知识图谱的发展与构建[J].南京理工大学学报,2017,41(1):22-34.

[3] 张香玲,陈跃国,马登豪,等.实体搜索综述[J].软件学报,2017,28(6):1584-1605.

[4] 刘峤,李杨,杨段宏,等.知识图谱构建技术综述[J].计算机研究与发展,2016,53(3): 582-600.

[5] 曹倩,赵一鸣.知识图谱的技术实现流程及相关应用[J].情报理论与实践,2015,38(12):127-132.

[6] 杜方,陈跃国,杜小勇.RDF 数据查询处理技术综述[J].软件学报,2013,24(6):1222-1242.

[7] CHEN Y G, GAO L X, SHI S M, et al. Improving Context and Category Matching for Entity Search[C]// In: Proc. of

the 28th Conf. on Artificial Intelligence. Palo Alto: AAAI, 2014:1622.

[8] 李慧颖,瞿裕忠.基于关键词的 RDF 数据查询方法[J].东南大学学报(自然科学版),2010,40(2):270-274.

[9] LI Huiying,QU Yuzhong.KREAG:Keyword Query Approach over RDF Data Based on Entity-triple Association Graph[J]. Chinese Journal of Computers,2011, 34(5):825-835.