

关系数据库向 Neo4j 图数据库转化的应用研究

——以工程科技词系统为例

王 力^{1,2} 韩红旗^{1,2} 高 雄^{1,2} 悦林东^{1,2} 张 琳³ 邱 爽³

(1. 中国科学技术信息研究所, 北京 100038; 2. 富媒体数字出版内容组织与知识服务重点实验室, 北京 100038; 3. 北京建筑大学, 北京 100044)

摘要: 基于关系数据库存储的大规模词系统存在数据查询效率慢、可扩展性差等缺陷, 影响了词系统知识库资源功能的发挥。将词系统转化成图数据库进行存储, 可有效解决上述问题, 有利于信息的检索与发现。以工程科技词系统为例, 针对关系数据库中存在的同名词检索、词间关系和词条属性映射以及分类体系表达问题提出词系统的 Neo4j 图数据库存储方案。该方案可显著提高词条检索效率, 支持最短路径、所有最短路径、路径深度配置的查询及词条信息的快速增删, 解决多领域词表的图数据库转化及分类与词表的映射问题。

关键词: 词系统; 图数据库; 关系数据库; 最短路径; 知识服务

中图分类号: G250

文献标识码: A

DOI: 10.3772/j.issn.1674-1544.2021.05.008

Research on the Application of Relational Database to Neo4j Graph Database

—— Taking the Engineering Science Word System as an Example

WANG Li^{1,2}, HAN Hongqi^{1,2}, GAO Xiong^{1,2}, YUE Lindong^{1,2}, ZHANG Lin³, QIU Shuang³

(1. Institute of Scientific and Technical Information of China, Beijing 100038; 2. Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content, Beijing 100038; 3. Beijing University of Civil Engineering and Architecture, Beijing 100044)

Abstract: The large-scale vocabulary system based on relational database storage has some defects, such as slow data query efficiency and poor scalability, which affect the use of word system knowledge base resources. Transforming the word system into a graph database for storage can effectively solve the above problems and facilitate the retrieval and discovery of information. Taking engineering science and technology vocabulary system as an example, the Neo4j storage scheme of word system was proposed to solve the problems of same-noun retrieval, inter-word relationship, term attribute mapping and classification system expression in relational database. This scheme can significantly improve the efficiency of entry retrieval, support the query of shortest path, all shortest path, path depth configuration and entry information fast addition and deletion.

作者简介: 王力 (1993—), 女, 中国科学技术信息研究所助理研究员, 研究方向为知识管理与知识服务; 韩红旗 (1971—), 男, 中国科学技术信息研究所研究员, 研究方向为数据挖掘(通信作者); 高雄 (1990—), 男, 中国科学技术信息研究所助理研究员, 研究方向为自然语言处理; 悦林东 (1991—), 男, 中国科学技术信息研究所工程师, 研究方向为数据库管理; 张琳 (1975—), 女, 北京建筑大学教授, 研究方向为智能信息处理; 邱爽 (1999—), 男, 北京建筑大学本科生, 研究方向为知识图谱。

基金项目: 中国工程科技知识中心项目“知识组织体系建设”(CKCEST-2021-2-6); 中国科学技术信息研究所重点工作项目“多模态知识图谱构建及应用”(ZD2021-11); 2020年北京高等学校高水平人才交叉培养“实培计划”项目。

收稿时间: 2021年4月6日。

Solves the graph database transformation of multi-domain thesaurus, and the mapping problem between classification and thesaurus.

Keywords: Vocabulary system, graph database, relational database, shortest path, knowledge service

0 引言

科技创新的活动中离不开科技信息资源的支撑。日益增长的信息数量正在对科技资源服务和知识技术提出了更高的要求。知识组织系统是知识技术的核心,可有效提高各类资源的开发利用效率。词表系统是基于知识组织的核心思想,对某一主题法实体的具体体现,包括分类表、叙词表、可检词单、同义词环、术语表、名称规范文档等^[1]。其主要用途是方便科技信息工作者更准确、全面地标引和检索文献。汉语科技词系统是在词表系统的基础上,充分吸收本体思想而构建的专业领域知识服务系统,既能支持传统的叙词表、分类、范畴等,又能支持用户自定义的属性和关系,具有良好的知识组织系统兼容性,本质上是一种知识库资源。词系统能够有效满足科技信息资源深度加工的需要,为高水平科技服务提供了知识组织内容资源^[2]。

工程科技词系统全称是中国工程科技知识中心知识组织系统,由中国工程科技知识中心委托中国科学技术信息研究所在20余分中心提供的领域词表基础上建设的,旨在促进各类资源的汇集、融合、连通和共享,支持知识中心的各项知识服务功能。其发展历经了集成融合、关系和属性扩展以及词系统协同构建3个阶段。该系统以词条(Term)作为基本组织对象,包含基本信息、定义及注释、属性、多维分类、词条之间的关系以及形式化概念描述等知识结构。其中,词间关系可精确定位一个概念;属性可对一条词条进行限定;分类包括3个分类体系,即《中国图书馆分类法》(CLC)、《国际专类分类法》(IPC)和领域相关分类体系(DSC);形式化概念描述有利于词系统自动化分析和计算。通过对这4个要素的不断优化和完善,可以有效提高词系统知识服务功能的发挥,促进行业知识服务产品的推

广和应用。

现有词系统包含了21个词表,有2 469 796个词条、3 180 610个关系,并以关系数据库的形式存储,广泛应用于知识中心和各下属分中心的知识组织体系及资源建设中。自20世纪80年代以来,关系数据库一直是数据库领域发展的主力并持续至今,具有简单、易操作的优点。但当面临大规模数据作业时,则需要耗费大量的时间和操作成本,并且缺乏连通性^[3]。基于关系数据库构建的词系统,其词条和关系是分开存储的,并利用主键、外键等建立数据库表间的关系。这种存储方式带来的主要问题是:若用户想获取某个关系词条的名称,或检索与某词条具有某种关系的词条集合等,常需通过资源消耗较大的连接(join)操作来实现;基于关系数据库存储的词表缺乏跨知识链接的能力,具体表现为同名概念词汇在多个领域的无关系存储为跨领域的知识检索带来不便;缺乏一套灵活的分类体系存储方案,无法将分类与概念知识进行有效链接。此外,在增加属性或关系等词表扩展上,关系数据库也对数据库设计 and 使用者提出了较高的要求。

和关系数据库相比,图数据库更善于处理大量复杂、低结构化且互连接的数据,支持用户频繁的查询,并且采用去中心化分布式存储,易于扩展应用到多台服务器上,支持Java、Python等多种流行语言调用^[4]。随着互联网技术的发展,传统的关系数据库已不能很好地满足用户信息组织和知识服务方面的需求,限制了词系统功能的发挥。因此,将大规模词系统从关系数据库向图数据库的转化逐渐成为系统升级的首要选择。Neo4j是一种同时支持节点和关系的属性定义的图数据库,可以清晰、灵活地表示节点、关系和属性元素,能实现专业数据库级别的图数据模型的存储;可以灵活添加或更改数据和数据类型,提供更快的事务处理和数据关系处理功能,有利

于缩短开发时间,实现项目的敏捷、迭代开发。因此,本文基于工程科技词系统在发展和应用过程中的问题,提出了词系统的图数据库存储方案,并选择Neo4j作为存储的图数据库实现词系统从关系数据库向图数据库的转化,重点解决词系统中同名词检索、词条属性及词条间关系映射和词系统分类体系表达等问题。

1 图数据库研究现状及分析

近年来,图数据库应用广泛^[5-7],它支持百亿乃至千亿量级规模的巨型图的存储。图数据库是以图论为理论基础,以节点和关系所组成的图作为数据模型的数据库,具有良好的可扩展性和互操作性,可实现高效的图数据利用和图分析^[8]。图分析可以深入探索各种实体(如组织,人员,交易)之间复杂的相互关系。

主要的图数据存储系统包括RDF图模型和属性图模型。前者以RDF(Resource Description Framework)三元组为存储对象。RDF由节点和边组成,节点表示实体/资源或者属性,边表示实体和实体之间的关系以及实体和属性的关系。RDF图模型具有较成熟的标准体系和标准查询语言SparQL,常见的数据库有Jena和Virtuoso等。属性图是目前主流图数据库选择的数据模型,更确切地说是带标签的属性图(Labeled-Property Graph),它的节点和边都可以定义属性。常见的图数据库有Neo4j、GraphDB和FlockDB等^[9]。其中,Neo4j实现了专业级别的图数据模型存储和高性能的图遍历功能,已被数十万家公司和组织采用。Neo4j主要元素包括节点、关系、属性和实体标签。这些不同元素存储在不同的文件中,具有明确的存储职责划分,并以图的形式进行链接,边缘信息则作为属性存储起来(图1)。Neo4j主要采用的是Cypher查询语言。该语言是一种声明式图数据库查询语言,具有丰富的表现力,能高效地查询和更新图数据^[10]。

将关系数据库转化为RDF图模型已有了较多的研究。例如,Angels R^[6]等详细地分析了图数据库的整体框架,并阐述了知识在图数据库中

的存储形式。师波等^[11]提出了一种能自动将关系数据库数据转换为RDF的方法,并通过实验验证了该方法的可行性。张晔等^[12]针对基于关系数据库构建的Acemap系统中存在的多表联合查询和扩展性差的问题,提出了三元组形式存储的图数据库解决方案。随着本体研究的不断深入,OWL本体作为RDF本体的扩展,已得到广泛应用。黄奇等^[13-14]提出了基于图形数据库的OWL本体存储模型,该模型满足模式结构的规范性、稳定性、可理解性,保证了语义的完备性,是一种理论上可行的规范存储模型,并以全球产品分类GPC为例,设计实现了产品分类本体的图数据库存储。这种RDF图虽然学术界研究较多,易于数据的发布、分享,但由于三元组的实体和关系不包含属性,不支持两个同样实体之间的多个同类关系,缺少灵活性,在实际应用中存在诸多不便。因此,近年来很多商业应用开始考虑采用图数据库取代以RDF图为基础的SparQL数据库。

随着基于属性图的图数据库的成熟和发展,已有一些学者针对关系数据库和图数据库之间转化的问题开展了研究。例如,郭林斐等^[15]分析了关系数据库的不足和图数据库的优势,基于Neo4j建立了用于处理不确定性历史数据的通用数学模型,解决了不确定性历史数据的语义框架问题。陈青云等^[16]对叙词表进行加工,以机械领域为例,将传统叙词表转换成SKOS本体,并利用Neo4j加以存储。沈思等^[17]从知识服务的角度出发,结合了图数据库的特点提出了分类表知识组织结构,并给出了分类表的图数据库存储方案。

从当前研究可以发现,尚未形成一套完整的

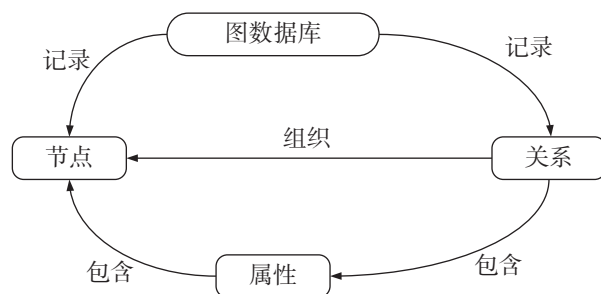


图1 图数据库模型

词表系统向图数据库转化存储方案。现有的部分研究虽探讨了词表和分类的图数据库存储,但并未解决多领域词表的图数据库转化及分类与词表的映射问题。因此,本文在前人研究的基础上,分析了工程科技词系统的知识结构特点,针对词系统现存问题,提出了基于Neo4j图数据库的词系统存储方案。

2 基于Neo4j的词系统存储方案

依据现有词表系统发展过程中产生的实际需求,提出基于Neo4j的词系统存储方案。将现有词系统从关系数据库转化为Neo4j图数据库主要考虑以下3个问题:一是现存词表是用多张关联表的关系数据库格式进行存储,词条范围覆盖多个工程领域,有些相同名称的词条出现在不同领域的词表中,具有不同的词条含义。因此,在关系数据库中索引这类词条时,会耗费大量操作和时间,大大降低了词系统的知识服务功能的发挥。二是当前词系统中词条之间的关系是一种重要的知识存在,通过词条关系可以较为精确地定位一个概念。词条的属性是对词条的另一种限定,是词条另一种重要的描述形式。因此,是否将关系和属性进行良好的表达是词系统能否提高知识服务水平的关键。三是分类表是词系统的一个重要资源,当前词系统支持多维分类,即一个词条可以用不同的分类法类目加以标识。因此,如何将分类体系转化成Neo4j进行存储也是至关重要的。

2.1 同名词处理

同名词是指在不同领域词表中出现的相同名称的词条。例如,名称为A的词条(Term)既存在于叙词表(Thesaurus) I中,又存在于叙词表(Thesaurus) II中。对同名词进行检索可以实现跨领域的知识链接。然而,在关系数据库中,对同名词A的词条(Term) A检索时为顺序检索,这大大降低了检索效率。在转换为图数据库时,一个解决方案是在不同词表的同名词条间建立sameAs关系,然而这会增加关系(边)的存储数量,而且不利于了解哪些概念具有跨领域特征。

例如,若5个词表存在同一个同名词,需要增加10条具有sameAs关系的边,且需通过查找全部sameAs关系并在去重后才能知道哪些词存在于多个领域。因此,本文提出了同名词检索转化方案(图2)。

图2以两个同名词为例,多个同名词的处理与之相同。同名词条(Term) A在叙词表(Thesaurus) I中与词条(Term) B具有关系(relation) α ,在叙词表(Thesaurus) II中与词条(Term) C具有关系(relation) β 。建立一类特殊节点,标签命名为Polysemy。这样同名词条(Term) A就可以表示为Polysemy{Term A},该节点与分布在叙词表(Thesaurus) I和叙词表(Thesaurus) II中的Thesaurus I {Term A}及Thesaurus II {Term A}建立链接。

当用户检索词条(Term) A时,可同时找到在叙词表(Thesaurus) I中与词条(Term) A有关系(relation) α 的词条(Term) B以及在叙词表(Thesaurus) II有关系(relation) β 的词条(Term) C。

2.2 词表关系处理

词间关系是词系统中一种重要知识的存在。词间一级关系有3种基本类型,分别是等同关系、层级关系和相关关系。其中,等同关系是有方向的,涵盖了叙词表或主题词表中的用代关

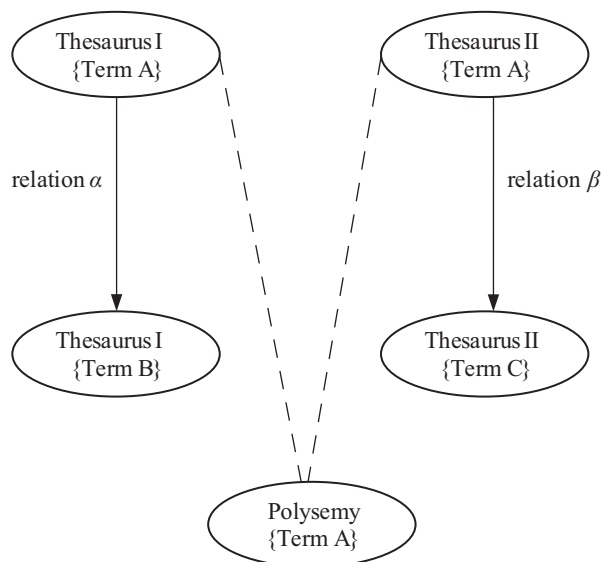


图2 同名词检索转化方案

系；层级关系是有方向的，涵盖了叙词表或主题词表中的属分关系；相关关系是无方向的，对应了叙词表或主题词表中的参关系。和传统叙词表或主题词表相比，词系统中的关系定义支持二级关系，更为复杂和灵活。其中，等同关系的二级类型包括“全称是”“缩略为”等；层次关系的二级类型包括“参与构成”“构成成分”等；相关关系依据领域的不同又分为控制关系、时间关系、空间关系、因果关系等类型。本文将这类关系称为相关关系，二级类型包括“替代”“影响”等（表1）。

为了解决二级关系在Neo4j中的存储，我们将二级关系名称转变为一级关系的属性来实现

词条间关系的描述。在实现方案中，关系数据库中的一级类型映射到Neo4j中作为关系名称出现，而二级类型作为关系的属性类型，即将二级类型作为关系的类型（type）呈现。例如，“等同关系—全称是”“层次关系—参与构成”“相关关系—影响”可采用图3所示的方式进行转化和存储。Neo4j中的关系一般是有方向的，因此在转化的过程中，对于无方向的关系需要创建双向的关系。

2.3 词表属性处理

词系统中属性描述模式和关系相似，存在二级属性（表2）。

表1 关系数据库中的关系表

序号	一级类型	二级类型
1	等同关系	全称是
2		缩略为
3		学名是
4		俗称为
5		基本等同
6	层次关系	参与构成
7		构成成分
8		参与组成
9		组成部件
10		类属
11	相关关系	替代
12		影响
13		类比
14		相似
15		表示

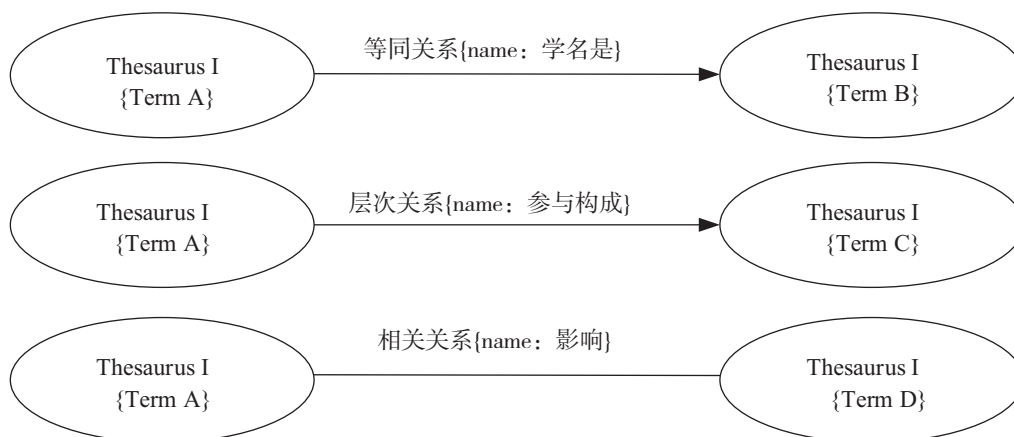


图3 关系转化示意

在将关系数据库存储的词系统转化为Neo4j的过程中,属性表的处理较为复杂,主要原因在于属性表中一级类型和二级类型的表示问题。如果只把二级类型保留作为词条间关系,则丢失了一级类型属性信息。同样地,若只把一级类型作为词条间关系名称,则失去了更为详细的二级类型属性信息。为了保留词条的二级属性信息,提出属性转化解决方案(图4)。

在该方案中,将一级类型的属性抽象为一类标签(采用PropertyNode表示),将一级属性名称用name属性来描述;将二级类型的属性抽象为一类关系,将二级属性的名称作为属性关系类型的type属性。这样在实例化时,词条的属性值作为PropertyNode节点的一个属性值,就可以在保留词条的一级和二级属性信息的同时,解决词条属性值的存储问题。

2.4 分类表处理方案

工程科技词系统的分类体系包括《中国图书馆分类法》(CLC)、《国际专类分类法》(IPC)和领域相关分类体系(DSC)。在关系数据库中,这3个分类体系分别作为3张表进行存储。将分类表转化成Neo4j存储的主要目的是构建一个可

以灵活使用的分类体系。本文以《中国图书馆分类法》(以下简称“《中图法》”)为例,提出词表转化的方案,IPC和DSC可以采用类似的方法实现。

《中图法》是以科学分类和知识分类为基础、结合文献内容特点及形式特征进行逻辑划分和系统排列的类目表,是类分文献、组织文献分类排架、编制分类检索系统的工具。《中图法》共分5个基本部类、22个大类,大类下面又分为多层次子类。采用汉语拼音字母与阿拉伯数字相结合的混合号码,用一个字母代表一个大类,以字母顺序反映大类的次序,在字母后用数字作标记。

在《中图法》分类表转化的过程中,为了表示《中图法》的一个类目,我们定义了一类“CLC”标签。该类标签具有两个基本的属性:一个属性名是CID,用来存储类目的编码,如“TD1”;另一个属性名是name,用来存储类目的名称,例如“矿山地质与测量”。为了表示类目之间的层级关系,我们定义了关系类型“subClass”作为上下级类目之间的链接。该关系类型具有一个基本属性“level”,用来表示关系的级别(图5)。

表2 关系数据库中的属性表

序号	一级类型	二级类型
1	特征	优点
2		缺点
3		数值
4		特点

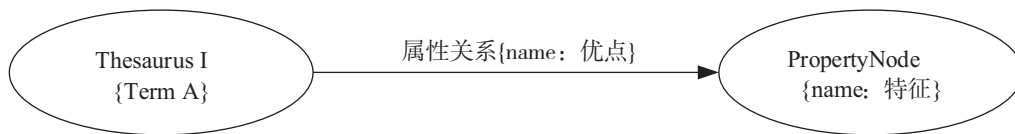


图4 属性转化方案示意

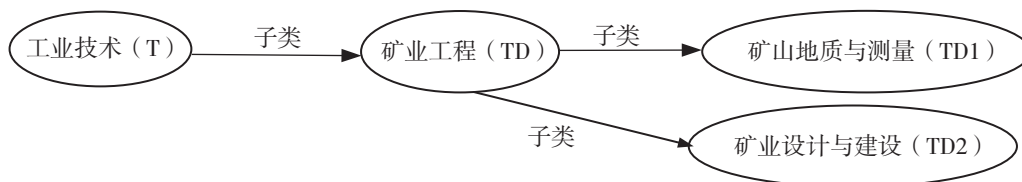


图5 属性转化方案示意

词系统中建立了词条与分类类目之间的链接关系。一般来说，词条和《中图法》类目之间存在着多对多的关系，即一个词条可能与多个《中图法》类目存在关联关系，同时一个《中图法》类目与多个词条间存在关联关系（图6）。

在将《中图法》转化为Neo4j存储后，可以方便地表示词条在《中图法》之间存在的多对多关系。为此，我们定义了“TermOfCLC”关系，从一个词条指向《中图法》一个类目（图7）。

3 实验结果及分析

本实验的硬件环境为64位的Windows 7操作系统；8 GB内存、AMD A10CPU。软件环境为1.8版本的JDK 8.0的Mysql和3.4.17版本的Neo4j。采用的数据来自于中国工程科技知识中心的知识组织系统，共包含21个词表。采用提出的存储方案后实现了将词系统从MySQL关系数

据库向Neo4j图数据库的转化。图8是转换后的一个片段的可视化展示。在图8中，“polysemy”代表新增加的同名词类的节点，“数据合并”和“数据处理”这两个词分别在“地质学主题词表”“信息中心主题词表”和“环境保护主题词表”这3个表中出现过，具体关系如图8中左边部分所示。图8的下面分别展示了关系和属性的具体描述信息。

为了验证转化后词系统的检索性能，开展了检索时间对比实验和关联路径查询实验。检索时间对比实验是为了验证转化后词表知识资源利用的高效性；关联路径查询实验是为了展示图数据库在检索的词表知识路径的便捷性。

3.1 检索时间对比实验

选择“有色金属”“黑色金属”两个领域的词表进行对比实验。数据信息共包括词条27 373条，词条关系70 785条，词条属性16 673条，

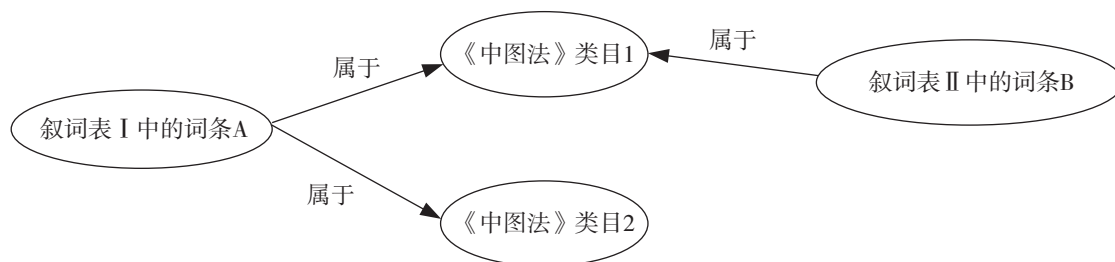


图6 分类关系描述示意

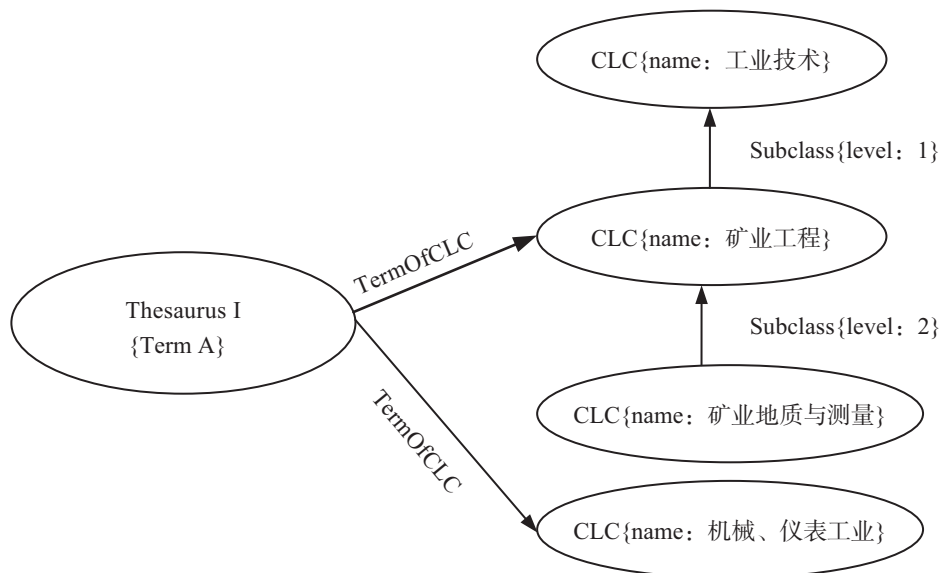


图7 分类体系转化方案示意

词条分类 11 837 条;《中国图书馆分类法》《国际专利分类法》两个分类法, 共有类目 109 441 条。

随机选择一个分类类目特种结构材料 (TB383), 在 Mysql 和 Neo4j 中分别检索其下所有的词条信息。检索结果见表 3。

和关系数据库的平均检索时间 0.950 秒相比, 基于 Neo4j 存储的工程科技词系统在检索词条时只需 0.129 秒, 大大降低了检索时间, 有效地提高了检索效率。

3.2 词条关系路径查询

两个词条关系路径查询有助于了解两个概念词汇之间的关联关系, 即了解相关知识之间深层次的关系。以词系统中的“球墨铸铁”和“脱碳相”两个词条为例 (图 9), 这两个词条之间并没有任何直接关系。如果要在关系数据库中查找它们的路径, 需要找到“球墨铸铁”的关系词条, 然后通过关系词条查找其直接关联的词条, 再从关联的每一个词条出发, 重复延伸关系, 一直找到“脱碳相”为止。假设“球墨铸铁”到“脱碳

相”之间有 N 个节点, 每个节点的关系数量为 T。在关系数据库中, 若要得到“球墨铸铁”与“脱碳相”之间的关系, 只有遍历最大 $N \times T$ 次, 才能得到。

而在 Neo4j 中, 可以使用查询语句直接获取到“球墨铸铁”与“脱碳相”之间的关系。查询语句为: MATCH (a:`黑色金属`{中文名称:"球墨铸铁"}), (b:`黑色金属`{中文名称:"脱碳相"}) return (a)-[*]->(b); 也可直接使用 allshortestPaths 获取所有最短路径, 即 MATCH n=allshortestPaths((a:`黑色金属`{中文名称:"球墨铸铁"})-[*]-(b:`黑色金属`{中文名称:"脱碳相"})) return n。查询结果见图 10。

可见, 在词系统中获取没有直接关系且相距较远的两个词条之间的关联时, Mysql 数据库需要极其复杂的操作过程才能实现, 普通人员几乎无法完成, 而在 Neo4j 中却可轻松实现。Neo4j 不仅支持所有路径的查询, 而且提供了最短路径、所有最短路径、路径深度配置的查询, 为关联词

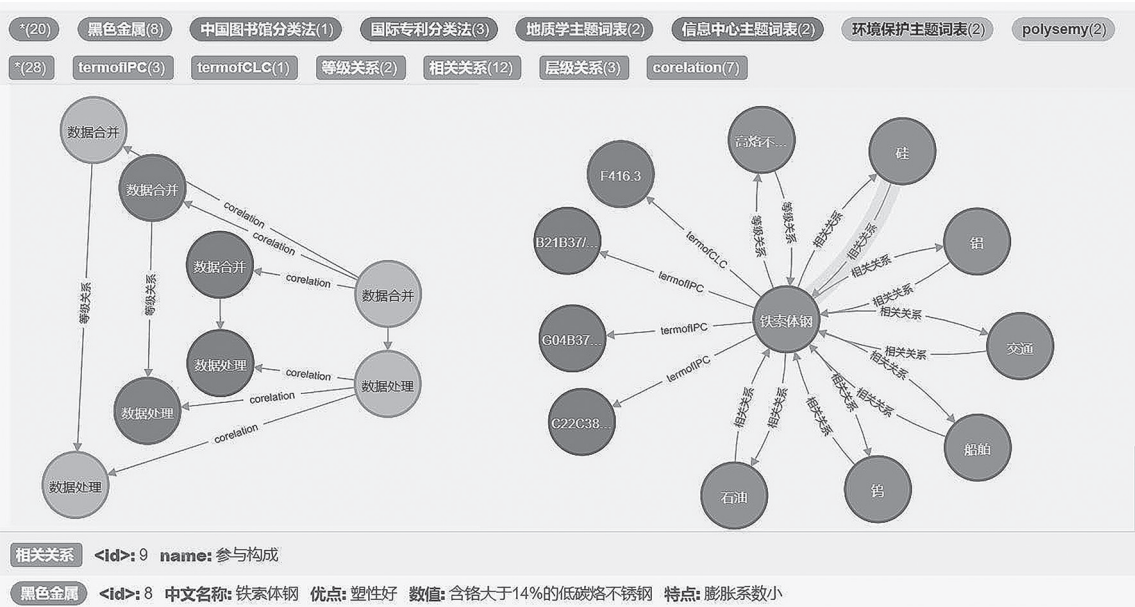


图 8 Neo4j 存储示意

表 3 关系数据库与图数据库检索时间

单位: 秒											
数据库	1	2	3	4	5	6	7	8	9	10	平均
Mysql	0.942	0.813	1.149	0.911	0.971	0.970	0.964	0.939	0.898	0.943	0.950
Neo4j	0.115	0.129	0.112	0.120	0.196	0.105	0.140	0.160	0.106	0.108	0.129

条的查询提供了便捷性。

3.3 词条信息动态更新

因为关系数据库要考虑规范化的问题，一个词条完整的信息常常保存在若干个表中，而图数据库是属性图，不存在这个问题，在词条信息动态更新时更加简便。因此，在增删改查方面，Neo4j的操作效率和可维护性一般要远高于关系数据库。下面以为一个词条增加一个属性来说明。具体来说，操作是在词条属性中添加一个新属性（字段），并为其添加一个词条属性的值，描述当前词条属性的来源。

在关系数据库中，描述词条属性共有3个表，分别为词条信息表、属性类型表、词条属性信息表。在关系数据库中的词条“生铁”，在词

条属性表中描述信息如图11所示，该图是由3个表关联查询得到的结果。

若对“生铁”添加新的属性“应用场景：轨道”，添加来源字段为“汉语科技词系统”。在关系数据库中需先添加字段：

```
alter table login_user add source varchar(255)
DEFAULT NULL;
```

然后执行插入操作：

```
INSERT INTO fhcb_tbl_term_attribute(termId,
attributeId, content, source)
```

```
SELECT t2.id as termId,
```

```
(SELECT t3.id from fhcb_tbl_attribute t3
where t3.name= '应用场景') as attributeId, '轨
道' as content, '汉语科技词系统' as source
```

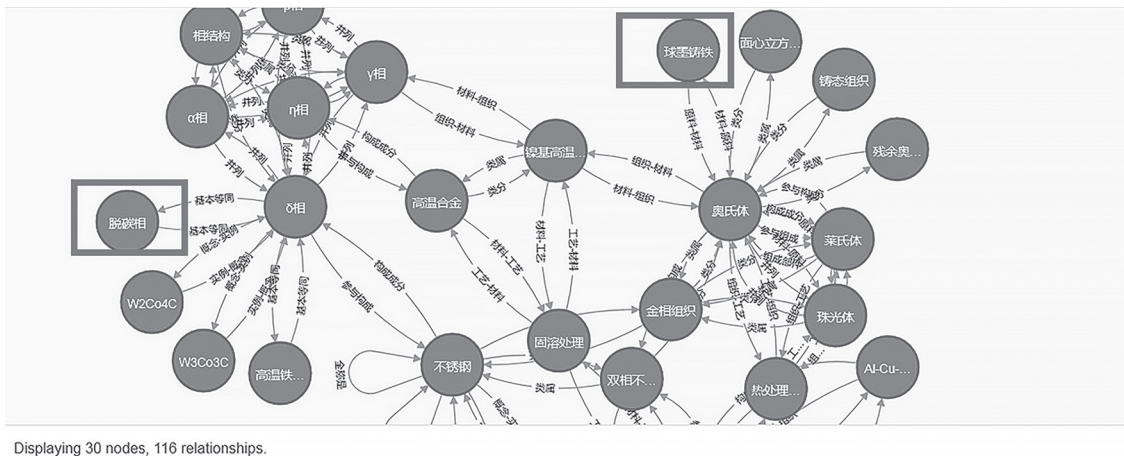


图9 图数据库示意

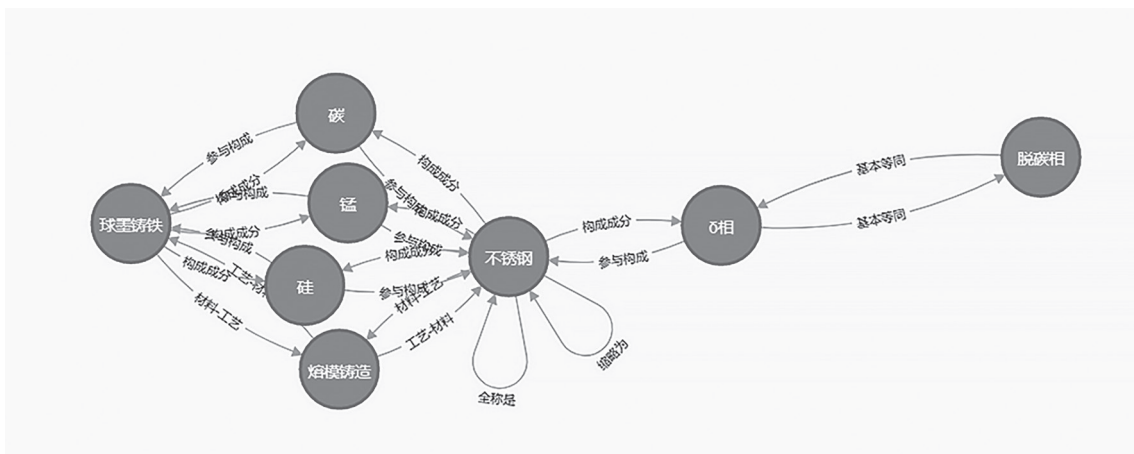


图10 关系路径查询示意

id	name	name1	content
70855	生铁	数值	碳2%以上
71777	生铁	应用场景	齿轮
72101	生铁	缺点	杂质较高
72676	生铁	应用场景	铁管
73643	生铁	应用场景	曲轴
73688	生铁	优点	强度高
74356	生铁	抗拉强度	490MPa以下
74697	生铁	缺点	韧性较差
74837	生铁	优点	耐磨性好
74941	生铁	优点	铸造性好
75214	生铁	应用场景	炼钢
76548	生铁	应用场景	活塞
76891	生铁	应用场景	机床床座
77223	生铁	缺点	不能塑性加工
78548	生铁	现状	工业生铁含砂
79106	生铁	应用场景	铸造

图 11 关系数据库中词条属性信息

from fhcb_tbl_term t2 where t2. 'name' = '生铁'

在执行插入操作的过程中,需要分别在词条所在的词表和属性类型表中找到其对应的id,在获取id后再进行插入。此外,对于使用关系数据库为存储的应用系统中,在开发时,还需要对数据库表修改实体类文件的字段,相应的接口层都需要进行修改,涉及大量的代码编程,操作较繁琐。

在Neo4j图数据库中,可直接插入词条属性,操作简单,无需对整个结构进行修改。其执行语句为:

```
MATCH (n{'中文名称': '生铁'}) SET
n.source = '汉语科技词系统', n.`应用场景`='轨道' RETURN n;
```

4 结论与展望

搜索和存储数据所付出的成本浪费是当前大数据研究普遍存在的问题^[18-19]。本文基于工程科技领域词系统知识资源有效利用的需求,针对同名词检索、词间关系和属性以及词系统中分类体系表达这3个问题,提出了由关系数据库向Neo4j存储的转化方案。该转化方案可以有效地解决同名词检索、词条属性及词条间关系映射和

词系统分类体系表达。词表系统转化为图数据库后,不仅提高了知识检索的速度,而且可以便捷地实现知识关联地查询,可以更加方便地实现词条的动态更新。

图数据库为知识组织系统发展带来了新的机遇和挑战。它不仅能够解决现有词表系统的存储和数据更新维护等问题,而且以其良好的特性帮助用户更好地理解知识,实现词表知识的高效利用。本文提出的词表系统向图数据库转化方案,可以解决多领域词表系统的图数据库转化,以及分类与词表的映射这两个问题,为未来知识组织系统的有效利用提供了基础。

本文的研究对象是工程科技领域词系统,和叙词表、主题词表或一般类型词表相比,词系统从规模和结构上更加庞大和复杂。因此,本文提出的方案可以广泛适用于其他类型词表的图数据库转化上。但在实际的应用过程中,还需在本方案基础上结合各个类型词表自身特征进行修改。

参考文献

- [1] 赵洁.受控词表互操作质量分析[D].太原:山西大学,2016.
- [2] 张运良,徐硕,朱礼军,等.汉语科技词系统:一种可用于科技信息资源深度内容分析的语义资源[J].图书情报工作,2011,55(4):100-105.
- [3] SANTOSH P, ERIKA J, MANISH M, et al. A research on architectural and performance comparison of relational database, NoSQL and graph database (Neo4j) [EB/OL].[2021-04-23]. https://www.researchgate.net/publication/327981497_A_research_on_Architectural_and_Performance_Comparison_of_Relational_Database_NoSQL_and_Graph_database_Neo4j.
- [4] 郭蕴颖.基于知识图谱的电网信息搜索引擎的设计与实现[D].沈阳:中国科学院大学(中国科学院沈阳计算技术研究所),2020.
- [5] ANGLES R, GUTIERREZ C. Querying RDF Data from a Graph Database Perspective[C]// Proceeding of ESWC. Heraklion:Springer-Verlag, 2005:346-360.
- [6] RENZO A, CLAUDIO G. Survey of graph database models[J]. ACM computing surveys (CSUR), 2008, 40(1): 1-39.
- [7] FLESCA S, GRECO S. Querying Graph Databases[C]//

- Proceedings of EDBT. Konstanz: Springer-Verlag, 2000: 510–524.
- [8] MARTINA E, MARJAN H, TATJANA D, et al. Applying k-vertex cardinality constraints on a Neo4j graph database[J]. Future generation computer systems, 2021, 115: 459–474.
- [9] GEORGIOS D, PANAGIOTIS G, REAS K. Graph-communities in Neo4j: four algorithms at work[J]. Evolving systems, 2018, 11(6): 1–20.
- [10] NASSIR A. Study to use Neo4j to analysis and detection SIM-BOX fraud[J]. Journal of pure & applied sciences, 2018, 4(17): 31–35.
- [11] 师波, 王凤利, 安建成. 从关系数据库到RDF的转换[J]. 电脑开发与应用, 2015, 28(1): 5–7.
- [12] 张晔, 贾雨葶, 傅洛伊, 等. AceMap学术地图与 AceKG学术知识图谱: 学术数据可视化[J]. 上海交通大学学报, 2018, 52(10): 1357–1362.
- [13] 黄奇, 钱韵洁, 袁勤俭, 等. 基于图形数据库的OWL本体存储模型研究[J]. 情报学报, 2019, 38(3): 310–321.
- [14] 黄奇, 钱韵洁, 袁勤俭, 等. 基于图形数据库的产品分类本体存储研究[J]. 情报学报, 2019, 38(8): 849–859.
- [15] 郭林斐, 刘广钟. 基于Neo4j不确定性数据处理技术的研究[J]. 计算机技术与发展, 2020, 30(1): 25–31.
- [16] 陈青云, 曹建飞, 陈荣祯. 从叙词表到知识图谱的构建研究与实践[J]. 农业图书情报, 2019, 31(1): 44–53.
- [17] 沈思, 苏新宁. 知识服务环境下分类表的知识组织探究[J]. 图书情报工作, 2014, 58(7): 113–118.
- [18] 潘刚, 张运良, 钟庆虹. 工程科技领域知识服务的思考与实践[J]. 情报工程, 2018, 4(5): 4–12.
- [19] 邢玉艳, 刘耀, 刘茹. 基于 schema 的信息安全标准资源解析研究[J]. 情报工程, 2019, 5(5): 108–118.

“国家科学数据中心”联合专刊征稿启事

《大数据》《数据与计算发展前沿》《中国科学数据(中英文网络版)》《中国科技资源导刊》

为系统呈现科学数据中心最新研究和应用成果,促进科学数据资源的有效管理、共享共用和分析挖掘,提高科学数据应用服务能力,《大数据》《数据与计算发展前沿》《中国科学数据(中英文网络版)》《中国科技资源导刊》4个期刊将联合首批20个国家科学数据中心,合作出版“国家科学数据中心”联合专刊,分别从政策、技术、数据、开放共享最佳实践等不同层面深入报道最新成果,共同推进科学数据领域的科学传播和学术交流。

“国家科学数据中心”联合专刊分为4个专题方向,相应论文分别发表于4个期刊。**专题方向一:**国家科学数据管理政策和标准规范、开放共享、数据分级(收录于《大数据》期刊);**专题方向二:**科学大数据在采集、汇交、保存、安全、分析、挖掘、呈现等方面的理论与前沿技术(收录于《数据与计算发展前沿》期刊);**专题方向三:**数据标准规范、优质数据集出版(数据论文)(收录于《中国科学数据(中英文网络版)》期刊);**专题方向四:**科学数据资源管理研究、精选数据集案例、优秀服务案例、开放共享最佳实践(收录于《中国科技资源导刊》期刊)。

《中国科技资源导刊》征稿范围

围绕“国家科学数据中心”联合专刊专题方向四,主要征集以下内容。征集对象不局限于各国家科学数据中心。

- (1) 科学数据资源管理研究与分析
- (2) 精选数据集及其科学应用成果
- (3) 面向国家重大战略需求和科技创新前沿的优秀服务案例
- (4) 数据开放共享的最佳实践案例
- (5) 国际各领域有代表性的平台和经验

专刊投稿说明

作者可根据专题方向进行投稿,稿件字数不设上限,建议不少于5000字,8000字左右为宜,具体投稿指南参见各期刊官网。来稿请注明“国家科学数据中心”联合专刊名称。投稿截止时间为2021年9月30日,刊出时间为2022年初。欢迎投稿!