文章编号：2095-2163(2022)05-0141-04 中图分类号：TP311.5 文献标志码：A

基于知识图谱的软件工程数据库设计

陈永芳

（茂名职业技术学院 计算机工程系，广东 茂名 525000）

摘 要：知识图谱技术不仅能全面系统地实现对软件工程数据库的汇总和调度，而且对系统知识的整合、软件工程行业技术人员知识响应和提供有着重要作用。本文在此技术的依托下，设计基于知识图谱的软件工程数据库，通过对软件工程知识图谱流程和研究方法探索分析，实现软件工程知识全面性、可扩展性、经济性和多样性功能，同时详细介绍设计原理、术语表示、知识表示（距离模型、单层神经网络）等软件工程数据库知识图谱技术功能，并进行相关功能、性能测试工作，由此实现了软件工程数据库的高效设计，并可为知识图谱软件工程数据库知识解译提供重要技术依据。

关键词：知识图谱；软件工程数据库；开发设计；测试；实现

Design of software engineering database based on knowledge atlas

CHEN Yongfang

（Department of Computer Engineering, Maoming Polytechnic, Maoming Guangdong 525000, China）

【Abstract】Knowledge atlas technology can not only comprehensively and systematically realize the summary and scheduling of software engineering database, but also play an important role in the integration of system knowledge and the knowledge response and provision of technicians in software engineering industry. Based on this technology, this paper designs a software engineering database based on knowledge atlas. Through the exploration and analysis of the process and research methods of software engineering knowledge atlas, the functions of comprehensiveness, scalability, economy and diversity of software engineering knowledge are realized. At the same time, the design principles, terminology representation, knowledge representation (distance model, single-layer neural network) and other software engineering database knowledge graph technology functions are introduced in detail, and related functions and performance testing are conducted. The research realizes the efficient design of software engineering database and knowledge atlas, and provides important technical basis for knowledge interpretation of software engineering database.

【Key words】knowledge atlas; software engineering database; development design; testing; realization

0 引 言

根据国家“十四五”规划工程实施提案规定，中国在 2020 年底将加快工业软件的发展，以实现国家软件安全和工业、制造业的合理转型，这不仅是软件工程领域发展的机遇，也是在实践过程中面临的重大挑战[1-3]。目前，在软件工程领域中，通过数据库构建进行软件管理，而在数据库建立过程中，主要通过构建工程知识图谱进行软件工程数据库的研发、分配、工作推进、故障消除、检索、推送等，一方面增加了软件工程数据库的智能化和管理优质，另一方面显著提升了软件开发的质量和效率，使用户在获取软件知识时更便捷、高效。

知识图谱是指对软件工程的概念和功能进行多层次、全方位的汇总，进而构建类似思维领域的思维导图，再通过图形可视化、应用模型、数学统计和科学统计等手段构建智能化推荐框架，实现对软件工程资源细致化描述，使各软件知识在软件数据库中获得快速响应[4]。当前，构建的知识图谱主要功能是快速检索、获取信息资源深度和信息获取效率提升等，不仅是金融、科学教育、软件工程和培训中的重要技术支持，而且是知识图谱数据源中的重要组成部分之一[5]。

为实现智能化、科技化、合理化且准确化的软件工程数据库构建，本文基于软件工程基础数据，以知识图谱为技术手段构建软件工程数据库，该数据库具备全面性、可扩展性、多样性和经济性等基础特征，并通过术语表示、数学模型构建等方法实现对软件工程数据库的获取、查询和优质管理。随后，通过对软件工程数据库知识图谱数据进行性能、功能双测试，进而优化基于知识图谱的智慧化信息资源调度系统，充分发挥软件工程数据库的综合经济效益，实现软件工程资源信息的可持续发展战略。

作者简介：陈永芳（1974-），女，本科，讲师，主要研究方向：数据库应用开发、人工智能应用、软件开发等。

收稿日期：2021-12-06

1 软件工程数据库知识图谱构建理论及技术

1.1 软件工程数据库体系构成

本文为详细掌握基于知识图谱的软件工程数据库,便于用户和机构高效、多方面地获取软件信息资源,构建软件工程知识图谱体系示意图(见图 1)。通过图 1 可知,研究构建体系主要由数据获取/导入、知识建模和知识融合构成。其中,数据获取/导入主要通过外部关系数据库、半结构数据库和网络数据库导入软件工程信息资源,从而扩展软件工程知识图谱的知识素材,实现数据的多元化。知识建模主要分为关系型数据建模和文本关键要素识别,是知识图谱技术的核心内容,其主要功能是通过术语表示、知识抽取和实体建模等过程进行软件工程资源知识获取、识别等。知识融合则主要分为知识对齐、图谱更新内容,这里的知识对齐具体分为实体对齐、属性对齐和实体关系学习,以此为基础,通过软件知识资源的获取、识别、对齐后实现软件工程资源数据的更新和融合,进而构建基于知识图谱可视化、智能化的软件工程数据库。

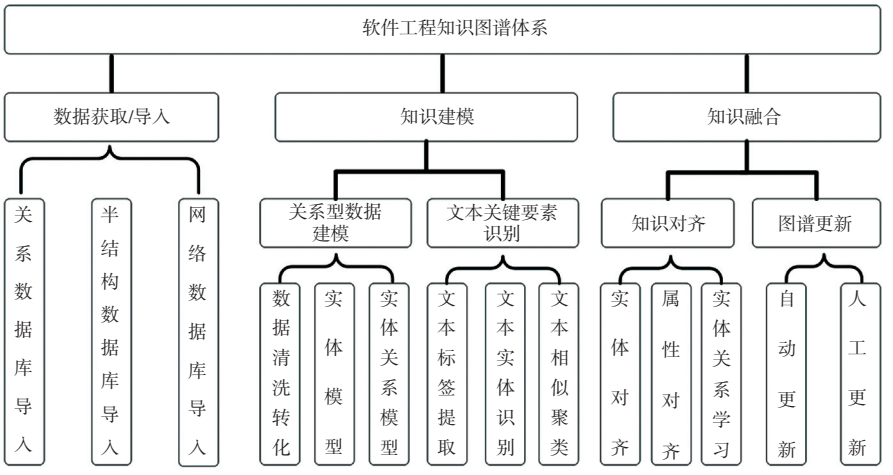

图 1 软件工程数据库知识图谱示意图

Fig. 1 Schematic diagram of knowledge map of software engineering database

1.2 知识图谱架构研究方法

本文通过知识图谱方法对软件工程数据库进行研究。研究中在构建知识图谱时,主要通过词频分析法、关联词分析法和社会网络分析法等研究方法实现知识图谱架构。其中,词频分析法,通过软件工程中的关键词,如软件、科学工程、信息技术等进行索引,将具备相关软件知识资源的内容汇聚在一起,同时通过关键词频获悉词频关注度,进而分析关键词探析软件工程结构、研究热点等内容,实现知识汇总。关联词分析法,主要通过同类中的相关性,揭示研究对象与对象间的特征关系,进而将有关词频加入知识图谱信息库,实现软件工程数据库的建立。社会网络分析法,将软件信息资源扩展到社会各领域中,通过社会分析法,揭示相关领域间的关系和发展状态,以量化研究构建社会网络个体关系模型,通过网络属性,探析不同成员间网络结构特征和社会属性特征,用于实现软件工程知识的索引获取,将知识资源通过知识图谱技术展示给用户,使用户获取的信息精度和信息量皆为最优,实现知识图谱架构体系的优质性。

2 软件工程数据库知识图谱设计

2.1 设计原理

2.1.1 数据库全面性

本文基于知识图谱技术构建软件工程数据库,应用图像(可视化)和资源数据(数据表示)理论,通过定性、定量化分析应用对关系型数据、半结构型数据和网络型数据进行汇总,并以软件工程学科知识和相近管理学科知识作为数据源,实现软件工程数据的全方位、多层次定位分析,以保障数据信息的安全性和全面性,达到用户高效获取信息资源的目的。

2.1.2 数据库可扩展性

在软件工程知识图谱体系构建后,软件工程专业知识的内容将更为透彻,但知识在学习中是不断挖掘和拓展的过程,在不同用户获取软件工程知识、学习知识的过程中,对知识的解译程度不一,因此,

新的软件知识将不断涌现，并不断被知识图谱体系收录，以此实现了软件工程数据库的扩展，知识存储量也随即得以提高。

2.1.3　数据库多样性和经济性

本文基于知识图谱理论和技术所构建的软件工程数据库，主要包括软件工程基础知识、热点知识、前言理论知识、前言知识图谱知识、多层次软件知识、全方位软件工程理论等，实现了软件工程数据库的多样化，同时，在一定程度上也为软件工程理论研究提供了技术支持。在经济性方面，一方面降低了软件工程学科和数据库在知识获取方面的花销和周期，另一方面通过相关联系知识，提高了用户的使用概率，发挥了知识图谱工具的优势，经济发展能力得到提升。

2.2　设计思路

本文通过对软件工程数据库知识图谱进行开发设计，以利于便捷信息资源的管理和获取。首先将其主要分为分层次、步骤和模块三个方面。在分层次中，每层间具有本身的特征属性，但每层之间又存在着相互管理，通过分析数据资源原理、分层设计后有助于实现软件工程数据库知识图谱的安全、科学管理过程。其次，分步骤中，通过软件工程知识主体、研究热点、前沿知识汇总、存储、特征关联和获取等，将数据库设计为科学、高效、有管理制度的发展趋势，不仅有利于知识图谱软件工程数据库的良性发展，而且确定了知识的集中性和研究热点原则。在模块方面，通过构建知识实体获取、表示模块、数据库构建实现和数据库功能测试等模块，不同模块间相互关联，遵循数据资源共享、共建原则，以此，有利于知识图谱技术的可持续发展。

2.3　设计构架

图 2 为知识图谱设计框架示意图。通过图 2 可知，软件工程数据库知识图谱设计构架在图 1 软件工程数据库总体体系上，将其进一步细化为 5 部分，包括数据库、数据库整合、知识表示、图谱构建和应用服务等。其中，知识表示是主体研究内容，其实体对齐和质量评估是知识图谱建成的关键点，其功能一方面保证了数据获取过程中的准确性和有效性，另一方面软件工程数据经对齐特征，将关联数据汇总进数据库，实现了不同层面、不同方位上的知识解读，知识图谱研究框架的构建，有利于软件工程数据库的可视化数据展示，进一步有效避免了在数据分析、知识资源分析和知识属性特征认证中的错误性和重复性问题。

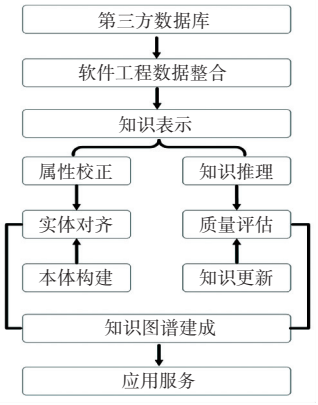

图 2　知识图谱设计构架

Fig. 2　Design framework of knowledge map

2.3.1　术语表示

在知识图谱架构体系中，术语表示是极为关键的，这不仅是知识图谱中学习概念、实体构建、数据资源和属性特征语言上的表现形式，而且是相关术语或者数据库存储标记的集合体。在构架体系中，术语常有单个字或者多个字、词组成，在特定的环境、特定的背景和领域中表达各不同的含义，对同种特征关系的知识图谱软件工程数据库的数据资源解译是极为重要的。术语表示中，术语抽取是核心内容，主要通过语言规则和统计学方法进行抽取表示，在统计方法中，统计基准值主要为 TF－IDF（Term Frequency－Inverse Document Frequency）、卡方分布和互信息分布等，不仅有效解译了术语表示信息，而且提高了用户对软件工程专业知识的认知效率。其中，术语表示统计学表达式可分别写为：

$$tfidf(w) = tf(w) \times \log\left(\frac{N}{df(w)}\right) \tag{1}$$

$$\chi^2 \sum \frac{(obs - \exp)}{\exp} \tag{2}$$

$$mi(x,y) = \frac{p(x,y)}{p(x)p(y)} \tag{3}$$

其中，$tfidf$ 表示最常用的属于分布特征权重基准；$tf(w)$ 表示词频，即词汇 w 在文档中的总次数；$df(w)$ 表示文档频率，即词汇 w 在文档中的数目；N 表示总文档数目。

2.3.2　知识表示

在知识图谱构架体系中，知识表示是指将收集、存储、转化和解译的信息直观清晰地展示在用户界面上。在软件工程数据库设计中，通过编码知识、行为、目标、偏好等，给出多个维度评价知识图谱知识表示，实现知识表示过程中具有足够多的细节知识，以及易理解、易传输和易提取等。同时为更好地实

现知识图谱数据资源知识表示,本文通过距离模型和单层神经网络进行数据库信息知识表示,对此拟做研究分述如下。

(1)距离模型。结构表示方法将头实体 h 和尾实体 t 通过关系 r 的 2 个矩阵投影到同一空间,投影向量之间的距离体现了 2 个实体在关系 r 下的语义相关度。对于每个三元组(h,r,t),损失函数为:

$$f_r(h,t) = \left|\boldsymbol{M}_{r,1}\boldsymbol{I}_h - \boldsymbol{M}_{r,2}\boldsymbol{I}_t\right| \quad (4)$$

其中,$\boldsymbol{M}_{r,1}$、$\boldsymbol{M}_{r,2}$ 是关系 r 对于头实体和尾实体投影矩阵。

(2)单层神经网络。此处涉及的数学公式可写为:

$$f_r(h,t) = \boldsymbol{\mu}_r^{\mathrm{T}} g(\boldsymbol{M}_{r,1}\boldsymbol{I}_h - \boldsymbol{M}_{r,2}\boldsymbol{I}_t) \quad (5)$$

其中,$\boldsymbol{\mu}_r^{\mathrm{T}}$ 为关系 r 的向量表示,$g(\cdot)$ 是 tanh 函数。

单层神经网络模型是结构表示的改进版本,利用神经网络的非线性减轻结构表示协同性差的问题。

3 知识图谱构建系统实现与测试

3.1 数据采集模式库实现

由于知识图谱是属于结构化的词义知识网络库,选取知识概念和相互关系用于知识抽取,进行大量知识的积累,因此,通过知识图谱可实现数据资源的快速响应。目前,在软件工程领域中,通过可视化将数据采集的项目、风险等级预测、质量因素等相关信息资源清晰地展示出来,对问题发现、数据汇总和高效查阅提供了优质手段。其次,知识图谱技术将相同特征信息进行聚类分析处理,很大程度上提高了用户在海量信息中获取关键术语、概念的精准效率。目前,在软件工程领域中(常见的百度搜索、天眼系统等),因行业区域知识相差较大,为符合实际应用需求,需构建知识图谱网状体系进行软件工程数据库的设计,从多方面、多层次内为用户资料获取提供强力支持。

3.2 系统功能模块测试

本文为保证知识图谱系统的各个模块能进行正常运转,通过不同测试方法对用户登录管理模块、知识表示模块、模型模块等方面进行测试,以探寻可能出现的系统问题,期待预期结果和测试方法见表 1。

表 1 系统功能测试内容

Tab. 1 System functional test content

测试名称	测试方法	实际预期结果
用户权限验证	输入登录帐号信息和密匙	进入界面
用户管理界面	对帐号密码的增加、删减等	成功登录帐号信息
知识融合	知识融合、模型融合	系统调用相应面模型完成融合
实体对齐	补全界面知识、补全知识图谱	系统调用模型完成实体对齐
模型测试	按照参数模型进行调整	系统参照模型参数运行
知识检索	进入知识检索界面输入要检索的知识	知识图谱中与用户输入相匹配的实体
知识推荐	点击所推荐的知识菜单	系统成功反馈相应知识的详细信息

4 结束语

本文通过知识图谱技术的优点特征设计了一类软件工程数据库,通过掌握知识图谱的构架过程及词频分析法、关联词分析法和社会网络分析法等研究方法,首先,深入分析了知识图谱数据库的全面性、可扩展性、经济性等优点,其特点是通过软件工程数据库实现用户的信息资源的精准获取目的。其次,通过对设计构架的知识表示、术语表示等的探讨来深入解析知识图谱原理,以便于软件工程数据库的设计、开发和应用。最后,通过知识图谱软件工程数据库的实现和测评,验证其资源获取性能和效率。本文通过知识图谱软件工程数据库的构建,极大提升了用户项目开发、软件工程信息资源调查时的工作效率。

参考文献

[1] 赵亮, 许娜, 张维. 我国数字孪生研究的进展、热点和前沿—基于中国知网核心期刊数据库的知识图谱分析[J]. 实验技术与管理, 2021, 38(11): 96-104.

[2] 李家瑞, 李华昱, 闫阳. 面向多源异质数据源的学科知识图谱构建方法[J]. 计算机系统应用, 2021, 30(10): 59-67.

[3] 刘香一, 甘少杰. 新世纪以来我国托育服务的研究热点与发展态势—基于 CNKI 数据库期刊的 Citespace 可视化知识图谱分析[J]. 大庆师范学院学报, 2021, 41(05): 104-112.

[4] 陈雅茜, 邢雪枫. 基于本体建模的动态知识图谱构建技术研究[J]. 西南民族大学学报(自然科学版), 2021, 47(03): 310-316.

[5] 杨艳丽, 席致宁. 国际新能源发展态势研究—基于科学知识图谱实证分析[J]. 科技创新与应用, 2020(23): 1-5.