doi:10.3772/j.issn.2095-915x.2024.04.001

基于多源数据的疾病知识图谱构建研究

孙丝雨[1] 侯跃芳[2] 丁敬达[1] 梅佳月[2] 孙佳[2]

1. 上海大学文化遗产与信息管理学院 上海 200444；
2. 中国医科大学健康管理学院 沈阳 110122

开放科学（资源服务）标识码（OSID）

摘要：[目的 / 意义] 基于 PubMed、OMIM 等医学数据库中的多源数据设计疾病知识图谱构建方案，为疾病的生物学实验研究及诊断治疗提供参考和依据。[方法 / 过程] 首先利用语义分析工具 SemRep 抽取 SPO 三元组，通过实体对齐、关系映射等数据处理方法进行知识融合，然后利用 Neo4j 图数据库实现知识存储及可视化展示，以多囊卵巢综合征为例进行实证检验和分析，最终获得 61589 个 SPO 三元组、34697 个实体和 27 种语义关系并归纳总结 7 种语义模式。[局限] 数据处理时，涉及人工审查，但由于数据量较大，审查过程中可能存在些许误差。[结果 / 结论] 本研究改进现有的知识融合方法，验证了该疾病知识图谱构建方案的可行性。为后续基于疾病知识图谱进行医学领域知识发现探索奠定基础。

关键词：疾病知识图谱；SPO 三元组；知识融合；语义分析

中图分类号：G302 G35

A Research on the Construction of Disease Knowledge Graph Based on Multi-source Data

SUN Siyu[1] HOU Yuefang[2] DING Jingda[1] MEI Jiayue[2] SUN Jia[2]

1. School of Cultural Heritage and Information Management, Shanghai University, Shanghai 200444, China;
2. School of Health Management, China Medical Univercity, Shenyang 110122, China

Abstract: [Objective/Significance] Based on the multi-source data in PubMed, OMIM and other medical databases, the construction scheme of disease knowledge graph is designed to provide reference and basis for biological experimental research, diagnosis and treatment of diseases. [Methods/Processes] Firstly, SPO triples are extracted by SemRep, and knowledge fusion

基金项目 辽宁省教育厅科学研究经费项目（人文社科类基础研究项目）“基于多源数据网络链接预测的知识发现模型构建”（JCRW2020005）。

作者简介 孙丝雨（2000-），硕士研究生，主要研究方向为文本挖掘与信息计量；侯跃芳（1972-），博士，教授，主要研究方向为文本挖掘与知识发现，E-mail：yfhou@cmu.edu.cn；丁敬达（1978-），博士，教授，主要研究方向为信息计量与科学计量；梅佳月（1998-），硕士研究生，主要研究方向为文本挖掘与知识发现；孙佳（1998-），硕士研究生，主要研究方向为文本挖掘与知识发现。

引用格式 孙丝雨 , 侯跃芳 , 丁敬达 , 等 . 基于多源数据的疾病知识图谱构建研究 [J]. 情报工程 , 2024, 10(4): 3-13.

doi:10.3772/j.issn.2095-915x.2024.04.001

is carried out by data processing methods such as entity alignment and relationship mapping. Then, knowledge storage and visual display are realized by Neo4j graph database. Taking polycystic ovary syndrome as an example, 61589 SPO triples, 34697 entities and 27 semantic relationships are finally obtained, and 7 semantic patterns are summarized. [Limitations] In the process of data processing, manual examination is involved, but due to the large amount of data, there may be some errors in the examination process. [Results/Conclusions] This study improves the existing knowledge fusion method and verifies the feasibility of the disease knowledge graph construction scheme. It lays a foundation for the follow-up exploration of knowledge discovery in medical field based on disease knowledge graph.

Keywords: Disease Knowledge Graph; Subject-Predication-Object; Knowledge Fusion; Semantic Analysis

引言

随着生物医学信息的爆炸式增长，传统的医学知识管理和应用方法已逐渐难以满足当前研究和临床需求。文章创新性地提出一种基于多源医学数据库的疾病知识图谱构建方案，旨在通过先进的语义分析和知识融合技术，为复杂疾病的研究提供一种全新的知识管理和决策支持工具。知识图谱呈现网状的结构，以三元组（头实体，关系，尾实体）的知识构成形式来表达现实世界中的知识或事实，其中图的节点代表实体或者概念，图的边代表实体或概念之间的各种语义关系[1]。多囊卵巢综合征（Polycystic Ovary Syndrome，PCOS）是一种影响全球许多育龄女性的异质性内分泌疾病[2]，通常与雄激素水平过高、胰岛素抵抗等有关[3]，且增加了引起严重并发症的风险，如心血管疾病等[4]。

因此，为了促进多囊卵巢综合征医学知识的共享、传播及利用，文章将一体化医学语言系统作为桥梁，实现了 PubMed、OMIM 等不同医学数据库中数据的标准化，解决了传统方法中数据异构性带来的挑战。采用最新的自然语言处理工具 SemRep，通过定制化的语义分析，精确抽取医学文献中的 SPO 三元组，为构建知识图谱提供了高质量的原始数据。在知识融合方面，提出一套包含实体对齐、关系映射和属性整合的系统化方法，显著提高了知识图谱中数据的准确性和一致性。利用高性能的 Neo4j 图数据库，不仅实现了医学知识的高效存储，还通过直观的可视化技术，极大地提升了知识图谱的可访问性和用户的交互体验。

1 相关研究

随着医学信息的日益增加和知识融合技术的不断完善，越来越多的学者利用不同的数据源进行疾病知识图谱的构建研究。蔡妙芝等[5]利用 SemRep 在糖尿病专题文献中抽取 SPO 三元组，利用 Neo4j 对其进行可视化展示并进行糖尿病知识发现。刘勇等[6]将电子病历、医学指南等不同数据源中的医学数据进行语义整合，构建了糖尿病医疗知识图谱。Fang 等[7]针对异构数据源之间的知识重复问题，提出了一种基于首尾实体融合模型的实体融合方法。聂莉莉等[8]利用权威的医学文献和书籍中的医学知识，运用自然语言处理技术，构建以“疾病—症候—特征”三层结构模型为基础的呼吸系统医疗诊

断知识图谱。翟东升等[9]从 IncoPat 专利数据库和 TCMSP、OMIM 等数据库中获取中医药相关数据，利用深度学习、字符串匹配等方式进行知识抽取、数据规范及实体对齐，构建了基于多源数据的中医药知识图谱。付洋等[10]以高品质的百科数据及医学文献为基础构建心脏病本体，以自顶向下和自底向上相结合的方法，半自动化地构建了心脏病中文知识图谱。Zhu 等[11]整合了 34 个不同的生物医学数据集中的数据资源，在 Neo4j 中生成了罕见病综合知识图谱，并通过四个案例进行研究分析。

综上所述，近几年医学知识图谱的构建研究取得了一些新的进展。首先，用于构建医学知识图谱的数据来源越来越丰富，包括科研文献、各种医学知识库、电子病历、医疗百科等；其次，医学知识图谱所涵盖的疾病类型越来越多，并逐步从常见疾病发展到罕见疾病。目前，针对多囊卵巢综合征，还鲜见有相关知识图谱的构建研究。此外，虽然知识融合方法和技术得到了快速发展，但目前尚未形成通用性较强的统一融合框架，对多源、异构、大规模医学数据进行知识融合仍存在困难。有鉴于此，文章基于多源数据设计一套疾病知识图谱构建方案，重点在知识融合方法上进行改进，以一体化医学语言系统（Unified Medical Language System, UMLS）[12]为标准，实现不同数据源中实体概念对齐，并将不同知识库中的关系类型进行规范化映射。针对各知识库的特点对其审查状态进行分类，再利用该关系属性实现多来源的同一三元组的知识融合，在确保三元组唯一性、消除歧义的同时，提升知识图谱的质量与可信度。为验证该疾病知识图谱构建方案的可行性，文章选取多囊卵巢综合征进行实证检验并深入分析，最后归纳出七种语义模式，以期为后续基于疾病知识图谱进行医学领域知识发现、潜在关联预测等方面的研究提供新的方向与思路。

2 疾病知识图谱构建方案

如图 1 所示，文章设计的基于多源数据的疾病知识图谱构建方案主要包括五个部分：知识图谱架构、数据收集与知识抽取、数据处理与清洗、知识融合、知识存储及可视化。

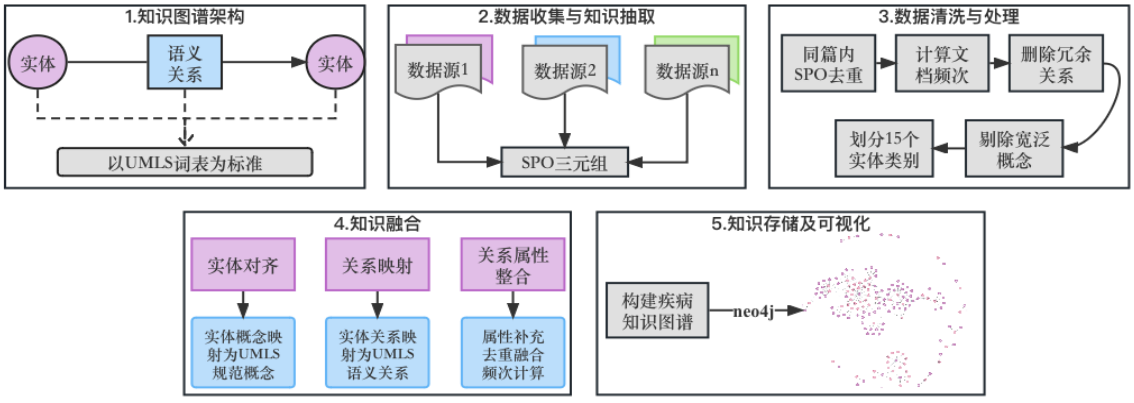

图 1　疾病知识图谱构建方案

doi:10.3772/j.issn.2095-915x.2024.04.001

SemRep[13] 由美国国立医学图书馆开发，是一个从生物医学文本中提取语义关系的 NLP 系统。它是一个基于规则的系统，以 UMLS 的超级叙词表、语义网络、专家词典和词汇工具为基础，对生物医学文献进行分词、语义分析、短语映射、语义谓词归一化、语义约束等的处理 [12]，自动提取相关文献中的术语概念和语义关系，形成主语—谓语—宾语（Subject-Predication-Object，SPO）三元组。其中主语和宾语是具有特定语义类型的 UMLS 词表概念，谓词是 UMLS 语义网络 [14] 扩展版本中的关系类型。目前，SemRep 被广泛应用于生物医学领域的知识图谱构建和知识发现。故文章利用 SemRep 从 PubMed 文献的标题和摘要中抽取 SPO 三元组，并以 UMLS 词表为标准对知识库中实体和语义关系进行规范化处理。

2.1 疾病知识图谱架构

知识图谱架构是构建知识图谱的基础，直接决定了所构建知识图谱的性能与质量。为有效融合医学文献数据和生物知识库数据，以 UMLS 词表为基础设计知识图谱架构，具体包括 SPO 三元组的实体、关系命名及其属性定义（见图 2）。实体指具有可区别性且独立存在的某种事物，医学领域实体指疾病、药物等；关系指两个或多个实体间的语义联系，医学领域实体间的关系包括疾病—疾病、疾病—药物等 [15]。文章定义的实体属性与实体间关系属性如图 2 所示，其中，实体类别 [7] 为依据 UMLS 词表提供的语义类型划分的 15 个语义组。

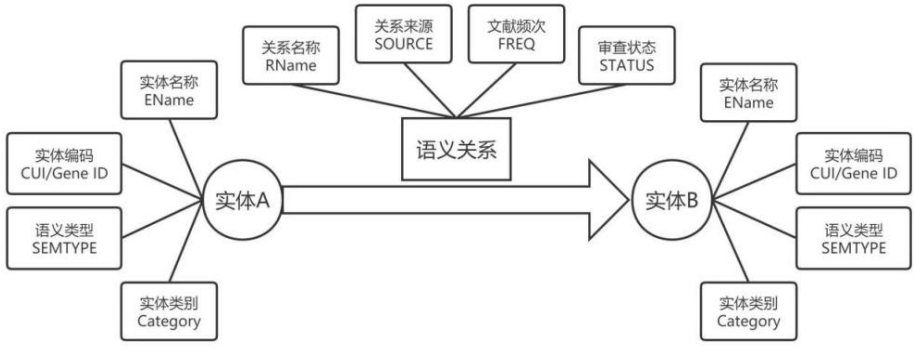

图 2 SPO 三元组属性定义

2.2 数据收集与知识抽取

相关数据收集是构建知识图谱的首要步骤，数据质量将直接影响知识图谱的准确性及可用性。知识抽取是指从各种数据源中抽取出实体（概念）、属性以及实体间的相互关系，在此基础上形成标准化的知识表达。文章选定 PubMed 文献库和 OMIM、CTD、DisGeNET、MalaCards 四种知识库作为数据来源。首先，在 PubMed 文献库中检索多囊卵巢综合征相关文献，检索式为：”Polycystic Ovary Syndrome”[Mesh]，截止到 2022 年 12 月 31 日，共得到 17481 篇医学文献。利用 SemRep 对这些文献的标题和摘要进行处理，得到 SPO 三元组数据表格，其中包含实体名称、实体 CUI 号（基

因为 Gene ID）、语义类型、语义关系等。

然后，分别从 OMIM、CTD、DisGeNET、MalaCards 四个知识库中搜集多囊卵巢综合征相关数据，具体过程如下：

（1）由于 OMIM 数据库中与多囊卵巢综合征相关的数据为文本类型，故对该数据库的数据处理方法与 PubMed 文献库类似，并对处理结果进行人工审核。

（2）在 CTD 数据库中搜索“Polycystic Ovary Syndrome”，下载化学物质—基因相互作用、疾病相关化学物质、相关基因三部分数据。CTD 对化学物质—基因相互作用提供了三种关系描述，即 increases（增加）、decreases（减少）、affects（影响，未指明方向）。CTD 对化学物质—疾病也提供了三种关系描述，即 marker/mechanism（与疾病相关的化学物质或可能在疾病的病因中起作用的化学物质）、therapeutic（对疾病具有已知或潜在治疗作用的化学物质）和 inferred（推断关系）。推断关系是通过 CTD 审核的化学物质—基因相互作用建立的，如化学物质 A 直接与基因 B 相互作用，基因 B 与疾病 C 相关，则推断化学物质 A 与疾病 C 有关系（通过基因 B）。CTD 对基因—疾病同样提供了以上三种关系。

（3）在 DisGeNET 数据库中搜索“Polycystic Ovary Syndrome”，下载该数据库中全部基因—疾病关联数据，同时下载该数据库中部分疾病—疾病关联数据，限定依据为经人工审查的数据库推断的疾病间关联。由于 DisGeNET 数据库中的知识关联来源于不同的数据库，因此，文章提供关系来源属性供研究人员进行追根溯源，其中 17 种来源数据库归为四种类型，即人工审查数据库（CURATED）、动物模型数据库（ANIMAL MODELS）、推断型数据库（INFERRED）和文献数据库（LITERATURE）。

（4）在 MalaCards 数据库中搜索“Polycystic Ovary Syndrome”，下载该数据库中相关疾病、药物和治疗、基因三部分数据。最后，整理以上各个数据库中的数据，以进行后续数据清洗及知识融合。

2.3 数据处理与清洗

数据处理与清洗是知识图谱构建过程中极为重要的步骤，筛选并保留高质量的 SPO 三元组，才能确保知识图谱的可信度。因此，文章首先对来源于 PubMed 文献库的 SPO 三元组进行处理与清洗，主要包括：同篇 SPO 三元组去重；计算同一 SPO 在不同文献中出现的频次；剔除否定谓词如“NEG_AFFECTS”等；人工剔除无意义的宽泛词，如“Human”等；依据语义类型划分 15 个实体类别。然后，依据各个知识库的数据特点，保留实体名称、文献频次等，补充实体类别、关系名称等，并将冗余数据删除，为后续知识融合阶段打下基础。

2.4 知识融合

知识融合是指通过映射和匹配使不同来源的知识在同一框架规范下进行整合、消歧和加工的过程。知识融合对提升知识图谱的质量、知识复用以及实现异构数据源之间的语义互通都具有重要意义[16]。因此，文章提出了一套系统化的知识融合方法，包括实体对齐、关系映射和关系属性整合，对不同来源的异构数据进行融合。

doi:10.3772/j.issn.2095-915x.2024.04.001

（1）实体对齐

为确保同一实体概念标识的唯一性，发现并匹配不同数据源中相同的实体概念[17]，文章利用 Postman 软件和 UMLS API 将从各个知识库中抽取的实体概念映射到 UMLS 词表中规范化的术语概念，完成实体对齐。但由于 UMLS 词表并不能包含所有的医学实体概念，针对那些未映射成功的实体概念，需要进行人工审查，确定其是否被 UMLS 词表收录。若已收录，则保留该实体概念的规范化名称及 CUI 号，若未收录，则保留该实体概念的原名称并将其 CUI 号赋予“NULL”。若出现多个映射结果，也需要进行人工审查，以确保映射结果的唯一性。

（2）关系映射

如上文所述，CTD 数据库为实体间关系提供了多种描述，需要进行映射处理，DisGeNET、MalaCards 中的实体间关系也需要定义和补充。文章以 UMLS 语义网络中的语义关系为标准，对知识库中的关系类型进行规范化处理，确保不同来源的 SPO 三元组中的实体间关系表达完整且统一，详情如表 1 所示。

表 1　知识库中的关系映射

数据来源	实体间关系	原关系名称	映射后关系名称
CTD 数据库	化学物质—疾病	marker/mechanism	CAUSES
		therapeutic	TREATS
		inferred	ASSOCIATED_WITH
	基因—疾病	marker/mechanism	CAUSES
		therapeutic	TREATS
		inferred	ASSOCIATED_WITH
	化学物质—基因	increases	STIMULATES
		decreases	INHIBITS
		affects	AFFECTS
DisGeNET	基因—疾病	无	ASSOCIATED_WITH
	疾病—疾病	无	ASSOCIATED_WITH
MalaCards	药物—疾病	无	TREATS
	基因—疾病	无	ASSOCIATED_WITH
	疾病—疾病	无	ASSOCIATED_WITH

（3）关系属性整合

经过实体对齐，已经较好地融合了实体的属性，实现了 SPO 三元组概念表达的一致性。对于关系的属性，由于各个数据库的审查状态不尽相同，文章将审查状态分为两种类型：人工审查（CURATED）、非人工审查（UNCURATED），如表 2 所示。

若出现同一三元组来源于不同数据库的情况（如某一 SPO 三元组既来源于 PubMed 文献库又来源于 CTD 知识库），为保证 SPO 三元组的唯一性，基于审查状态属性将多来源的同一三元组进行去重融合，将关系来源属性和审

表 2　审查状态属性定义

数据来源	关系来源情况	审查状态
PubMed	经 SemRep 抽取得到	UNCURATED
CTD	CTD 数据库提供确定的关系（如 therapeutic）	CURATED
	实体间的关系是推断关系	UNCURATED
DisGeNET	来源于经人工审查的数据库或动物模型数据库	CURATED
	来源于推断型数据库或文献数据库	UNCURATED
MalaCards	已批准的药物、具有明确证据的基因	CURATED
	由数据库推断得到的相关疾病和基因、除已批准以外的其余药物	UNCURATED
OMIM	SemRep 抽取后进行人工审核	CURATED

查状态属性进行整合更新（如 SOURCE:CTD, PubMed，STATUS:CURATED），并对文献频次取较高频次，生成最终的 SPO 三元组语义网，避免同一 SPO 三元组因来源不同在知识图谱中多次出现的情况，确保后续展示时的美观度及查询时的高效性。

2.5　知识存储及可视化

选择合理高效的知识存储方式，能够直接提高知识图谱的查询效率。Neo4j 图数据库因其具有高性能、实用性强、轻量级等优点，成为目前最常用的图数据库之一 [18]。基于此，文章利用 Neo4j 图数据库对 SPO 三元组进行知识存储及可视化展示，完成多囊卵巢综合征知识图谱的构建。

2.6　语义模式归纳

SPO 三元组语义模式是用于描述实体之间关系和属性的一种表示方式，文章定义的 SPO 三元组语义模式表示为“语义类型 A —语义关系—语义类型 B”，如 orch-TREATS-dsyn，orch 即 Organic Chemical（有机化学药品），dsyn 即 Disease or Syndrome（疾病或综合征），该语义模式表示有机化学药品治疗疾病或综合征。参考蔡妙芝等 [5] 总结的“诊断治疗”等 5 种语义模式，文章进一步归纳总结 7 种语义模式并进行详解。

3　结果与分析

3.1　PCOS知识图谱

文章从 PubMed 文献库中共获得 29721 个 SPO 三元组，加入四个知识库中收集的数据，并进行知识融合处理后，总计得到 61589 个 SPO 三元组，34697 个实体和 27 种语义关系。在 Neo4j 图数据库中的部分展示结果如图 3 所示。在图 3 中，以“Polycystic Ovary Syndrome”为中心，周围节点是与该疾病有明确语义关系的相关疾病（红色）、基因（绿色）、药物（黄色）等，例如胰岛素抵抗、FTO 基因、维生素 D 等。右侧图示中展示了该节点的四个属性，即 {EName：“Polycystic Ovary Syndrome”, ID:“C0032460”, SEMTYPE:“dysn”, CATEGORY：“Disorders”}。

图 4 为知识图谱中 SPO 三元组的实体及关系属性示例，节点（实体）集合 E 为 {E1、

doi:10.3772/j.issn.2095-915x.2024.04.001

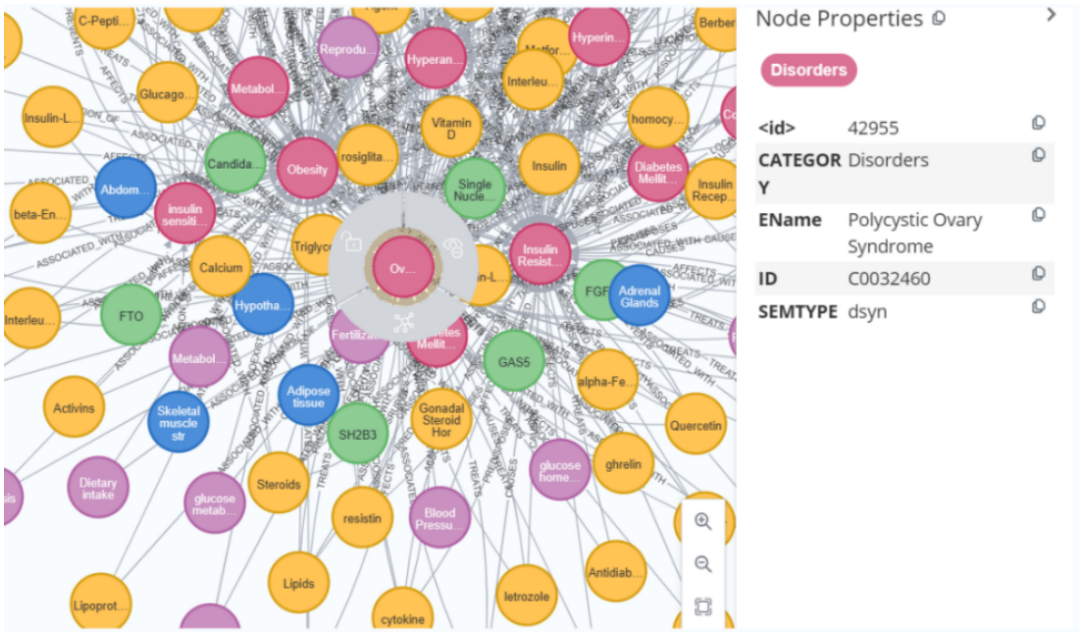

图 3　PCOS 知识图谱局部展示（实体）

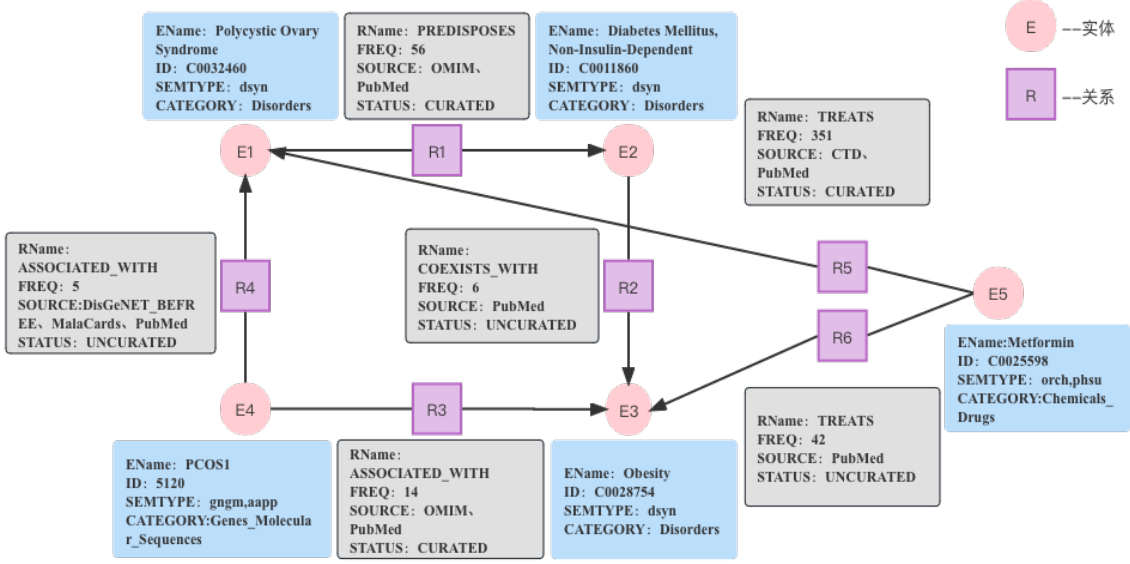

图 4　PCOS 知识图谱属性图示例

E2、E3、E4、E5}，边（关系）集合 R 为 {R1、R2、R3、R4、R5，R6}。由图可知，多囊卵巢综合征容易诱发非胰岛素依赖型糖尿病，非胰岛素依赖型糖尿病与肥胖并发，二甲双胍可用于治疗多囊卵巢综合征和肥胖，PCOS1 基因与多囊卵巢综合征和肥胖有关。

3.2　语义模式归纳与分析

文章将语义关系总结归纳为相关疾病、诊断治疗、影响因素、药理作用、疾病特征、功能相关、比较关系七种语义模式，具体结果如表 3 所示。

表 3　PCOS 知识图谱的三元组语义模式

类型	语义关系	语义模式示例	SPO 三元组示例
相关疾病	COEXISTS_WITH	dsyn-COEXISTS_WITH-patf	Spontaneous abortion-COEXISTS_WITH-Insulin Resistance
	COMPLICATES	dsyn-COMPLICATES-dsyn	Hyperinsulinism-COMPLICATES-Disorder of endocrine ovary
	PRECEDES	dsyn-PRECEDES-dsyn	Amenorrhea-PRECEDES-Polycystic Ovary Syndrome
诊断治疗	ADMINISTERED_TO	horm-ADMINISTERED_TO-mamm	estradiol valerate-ADMINISTERED_TO-Rattus norvegicus
	DIAGNOSES	lbpr-DIAGNOSES-dsyn	Follicle stimulating hormone measurement-DIAGNOSES-Polycystic Ovary Syndrome
	PREVENTS	orch-PREVENTS-comd	Metformin-PREVENTS-Oxidative Stress
	TREATS	orch-TREATS-dsyn	Metformin-TREATS-Hirsutism
关联因素	AFFECTS	comd-AFFECTS-dsyn	Oxidative Stress-AFFECTS-Polycystic Ovary Syndrome
	ASSOCIATED_WITH	phsu-ASSOCIATED_WITH-dysn	Vitamin D-ASSOCIATED_WITH-Obesity
	CAUSES	dsyn-CAUSES-dsyn	Hyperandrogenism-CAUSES-Infertility
	PREDISPOSES	dsyn-PREDISPOSES-mobd	Polycystic Ovary Syndrome -PREDISPOSES-Depressive disorder
	PRODUCES	bpoc-PRODUCES-horm	Ovarian Follicle-PRODUCES-Estrogens
药理作用	AUGMENTS	carb-AUGMENTS-patf	Inositol-AUGMENTS-insulin sensitivity
	DISRUPTS	strd-DISRUPTS-dsyn	Steroids-DISRUPTS-Polycystic Ovary Syndrome
	INHIBITS	horm-INHIBITS-horm	Testosterone-INHIBITS-Progesterone
	INTERACTS_WITH	phsu-INTERACTS_WITH-horm	Aromatase-INTERACTS_WITH-Estradiol
	STIMULATES	horm-STIMULATES-aapp	Gonadorelin-STIMULATES-Luteinizing Hormone
疾病特征	ISA	dsyn-ISA-dsyn	Polycystic Ovary Syndrome-ISA-Endocrine System Diseases
	LOCATION_OF	bpoc-LOCATION_OF-neop	Endometrium-LOCATION_OF-Adenocarcinoma
功能相关	MANIFESTATION_OF	dsyn-MANIFESTATION_OF-dsyn	Acne-MANIFESTATION_OF-Hyperandrogenism
	OCCURS_IN	dsyn-OCCURS_IN-humn	Polycystic Ovary Syndrome-OCCURS_IN-Adolescents, Female
	PROCESS_OF	moft-PROCESS_OF-cell	aromatase activity-PROCESS_OF-granulosa cell
	USES	topp-USES-phsu	Oestrogen therapy-USES-Progestins
比较关系	COMPARED_WITH	phsu-COMPARED_WITH-phsu	Acetylcysteine-COMPARED_WITH-Metformin
	HIGHER_THAN	orch-HIGHER_THAN-phsu	Clomiphene-HIGHER_THAN-Metformin
	LOWER_THAN	phsu-LOWER_THAN-phsu	Metformin-LOWER_THAN-Clomiphene Citrate
	SAME_AS	phsu-SAME_AS-phsu	Letrozole-SAME_AS-Clomiphene Citrate

doi:10.3772/j.issn.2095-915x.2024.04.001

在“相关疾病”语义模式中，COEXISTS_WITH 表示疾病的并发症；COMPLICATES 表示使其变得更加严重或复杂，或导致不利影响；PRECEDES 表示在时间上发生得更早。

在“诊断治疗”语义模式中，ADMINISTERED_TO 表示某种药物、治疗或其他医疗干预措施被施用于特定患者或生物体上；DIAGNOSES 表示区分或识别疾病的性质或特征，即诊断 PCOS 的手段或指标；PREVENTS 表示阻碍或消除病情的药物或手段；TREATS 表示以治愈或控制病情为目的治疗方法。

在“关联因素”语义模式中，AFFECTS 表示产生直接影响；ASSOCIATED_WITH 表示疾病、基因和药物之间的关联关系；CAUSES 表示导致某种状况或结果，主要揭示了疾病的致病因素；PREDISPOSES 表示对某种疾病、病理或状况存在风险，主要揭示了疾病的诱发因素；PRODUCES 表示产生或创造，这包括分泌、生物合成、释放等。

在“药理作用”语义模式中，AUGMENTS 表示扩展或刺激一个过程，STIMULATES 表示增加或促进物质间相互作用，二者都反映了正向的关系。DISRUPTS 表示改变或影响已经存在的条件、状态或情况，产生负面影响，INHIBITS 表示减少、限制或阻断物质间相互作用，二者都反映了负向的关系。INTERACTS_WITH 表示物质间相互作用。

在“疾病特征”语义模式中，ISA 表示实体间的从属关系；LOCATION_OF 表示一个实体的位置、部位或区域或一个进程的部位。

在“功能相关”语义模式中，MANIFESTATION_OF 表示一种现象的一部分可以直接观察到，或明显地表现出来，或为潜在的过程提供证据；OCCURS_IN 表示在一个组或群体中有发病率；PROCESS_OF 表示疾病发生在高等有机体中；USES 表示用于执行某些活动。

在“比较关系”语义模式中，COMPARED_WITH、HIGHER_THAN、LOWER_THAN、SAME_AS 都是比较谓词，分别表示相较于、高于、低于、相同，如奥利司他治疗多囊卵巢综合征的疗效与二甲双胍相同（orlistat-SAME_AS-Metformin）。

4 总结与展望

文章首先设计了疾病知识图谱的构建方案，以 PubMed 文献库和 OMIM 等四种知识库为数据来源，以 UMLS 词表为基础定义了实体和关系的命名及其多种属性，并提出了一套系统化的知识融合处理流程，以多囊卵巢综合征为例验证了其可行性，可为其他疾病知识图谱的构建提供一定参考。虽然文章中的知识融合涉及一定程度的人工审核，但所采用的方法，如实体对齐、关系映射和关系属性整合，可以为其他研究者在知识融合自动化方面提供思路和框架。随着自然语言处理（NLP）和机器学习技术的发展，这些人工环节有望被自动化算法所替代，从而提高效率并减少误差。

通过利用 Neo4j 图数据库，文章实现了知识图谱的高效存储和直观可视化展示，这不仅提高了知识图谱的查询效率，也使得医学专业人员能够更容易地理解和利用知识图谱。在深入分析后，文章归纳总结了 SPO 三元组的七种语义模式，这有助于机器更好地理解、处理和

使用知识图谱中的数据，从而实现更高效、智能的语义分析和查询。后续可以尝试开展 PCOS 知识图谱的网络展示及查询，供生物医学相关人员使用，为多囊卵巢综合征的诊断治疗提供相应的参考和支持；同时，基于该知识图谱，尝试进行多囊卵巢综合征知识发现探索，为科研人员提供潜在知识关联与研究思路；该知识图谱提供了规范化的、机器可理解的与多囊卵巢综合征疾病相关的知识，也可尝试将其应用于智慧问答系统的开发，推动智能医疗的发展。

参 考 文 献

[1] 董文波 , 孙仕亮 , 殷敏智 . 医学知识推理研究现状与发展 [J]. 计算机科学与探索 , 2022, 16(6): 1193-1213.

[2] ROTHENBERG S S, BEVERLEY R, BARNARD E, et al. Polycystic ovary syndrome in adolescents[J]. Best Practice & Research. Clinical Obstetrics & Gynaecology, 2018, 48: 103-114.

[3] WITCHEL S F, OBERFIELD S E, PEÑA A S. Polycystic Ovary Syndrome: Pathophysiology, Presentation, and Treatment with Emphasis on Adolescent Girls[J]. Journal of the Endocrine Society, 2019, 3(8): 1545-1573.

[4] GLUECK C J, GOLDENBERG N. Characteristics of obesity in polycystic ovary syndrome: Etiology, treatment, and genetics[J]. Metabolism: Clinical and Experimental, 2019, 92: 108-120.

[5] 蔡妙芝 , 李晓瑛 , 赵嘉玮 , 等 . 基于 SPO 语义三元组的疾病知识发现 [J]. 数据分析与知识发现 , 2022, 6(1): 134-144.

[6] 刘勇 , 齐梦霁 . 基于糖尿病防治的医学知识图谱构建的研究 [J]. 医学信息 , 2020, 33(18): 11-14.

[7] FANG A, LOU P, HU J, et al. Head and Tail Entity Fusion Model in Medical Knowledge Graph Construction: Case Study for Pituitary Adenoma[J]. JMIR Medical Informatics, 2021, 9(7): e28218.

[8] 聂莉莉 , 李传富 , 许晓倩 , 等 . 人工智能在医学诊断知识图谱构建中的应用研究 [J]. 医学信息学杂志 , 2018, 39(6): 7-12.

[9] 翟东升 , 娄莹 , 阚慧敏 , 等 . 基于多源异构数据的中医药知识图谱构建与应用研究 [J]. 数据分析与知识发现 , 2023, 7(9): 146-158.

[10] 付洋 , 刘茂福 , 乔瑞 . 心脏病中文知识图谱的构建 [J]. 武汉大学学报 (理学版), 2020, 66(3): 261-267.

[11] ZHU Q, NGUYEN D T, GRISHAGIN I, et al. An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD)[J]. Journal of Biomedical Semantics, 2020, 11(1): 13.

[12] 李晓瑛 , 李军莲 , 李丹亚 . 一体化医学语言系统及其在知识发现中的应用研究 [J]. 数字图书馆论坛 , 2019(9): 24-29.

[13] RINDFLESCH T C, FISZMAN M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text[J]. Journal of Biomedical Informatics, 2003, 36(6): 462-477.

[14] KILICOGLU H, ROSEMBLAT G, FISZMAN M, et al. Constructing a semantic predication gold standard from the biomedical literature[J]. BMC bioinformatics, 2011, 12: 486.

[15] 刘悦悦 , 李燕 . 医学知识图谱研究综述 [J]. 软件导刊 , 2023, 22(5): 241-247.

[16] 范媛媛 , 李忠民 . 中文医学知识图谱研究及应用进展 [J]. 计算机科学与探索 , 2022, 16(10): 2219-2233.

[17] 张晗 , 安欣宇 , 刘春鹤 . 基于多源语义知识图谱的药物知识发现：以药物重定位为实证 [J]. 数据分析与知识发现 , 2022, 6(7): 87-98.

[18] 孙敏敏 , 毛雪岷 . 基于 Neo4j 的肺部疾病知识图谱构建 [C]// 中国管理现代化研究会 , 复旦管理学奖励基金会 . 第十五届（2020）中国管理学年会论文集 , 2020: 6.