摘要

心血管疾病知识图谱的自动构建与更新研究

心血管疾病是一类包括冠心病、心肌梗死、心绞痛、高血压、心律失常、心力衰竭等多种疾病的疾病群。随着人们生活方式的改变，心血管疾病的发病率越来越高，成为全球范围内的重要公共卫生问题。借助人工智能技术和知识图谱技术，可以构建医学知识图谱和药物知识图谱，以帮助医生更好地理解和掌握医学知识，同时推动新诊疗防范和新药研发。因此，心血管疾病知识图谱应用具有广泛的应用前景，可以为预防、诊断和治疗心血管疾病提供有力的支持。然而，随着心血管疾病研究的广泛开展和研究成果的不断涌现，传统的知识图谱构建需要耗费大量的人力，进行数据清洗、命名实体识别、关系抽取、冲突发现等工作。因此，实现心血管病知识图谱的自动构建与更新可以提高知识图谱的准确性和完整性，实现知识的及时更新，支持个性化的医疗决策。

为此，本篇工作提出了一种自动构建和更新的心血管疾病知识图谱的框架 ASK（**A**uto-construction and **S**elf-reflection Framework for Biomedical **K**nowledge Graph），基于生物医学文本内容，融合命名实体和关系抽取、链接预测和冲突消解等技术，实现了高效、准确、可靠的知识图谱 7×24 小时自动构建与更新。ASK 框架由自动构建、图谱更新和自省模块三部分组成。在构建模块中实现了定时从指定数据来源处获取医疗数据文本并实现抽取工作，采用 BioBERT 为基础的自然语言处理模型提取实体和关系，该模型使得构建知识图谱的自动化程度更高，减少了人工收集语料与构建图谱的工作量；在知识图谱更新模块中，将从语料中自动抽取的实体和关系融入到已有知识图谱中，引入了冲突消解算法，检测和解决新知识可能带来的冲突问题，提高知识图谱融合的准确性和质量。通过自省模块，我们实现了预测图谱中的隐含链接并且排除隐含链接可能导致的冲突，丰富知识图谱的内容，并针对这些冲突进行相应的调整和优化。在冲突消解过程中，我们可以检测和解决知识图谱中存在的矛盾、错误、重复等问题，从而实时提高知识图谱预测知识的准确性和完整性。

在我们的实验阶段中，我们进行了一系列的模型与算法的比较，以说明框架中每个模块选择模型或算法的原因，同时给出了 ASK 框架中不同模块中知识图谱的演化。

关键词：

知识图谱构建，知识图谱嵌入，链接预测，冲突消解，知识工程

Abstract

Study on Automatic Construction and Updating of Cardiovascular Disease Knowledge Graph

Cardiovascular disease is a group of diseases including coronary heart disease, myocardial infarction, angina, hypertension, arrhythmia, heart failure, and other related illnesses. With changes in people's lifestyles, the incidence of cardiovascular disease is increasing, becoming an important public health problem worldwide. With the help of artificial intelligence and knowledge graph technology, medical and drug knowledge graphs can be constructed to help doctors understand and master medical knowledge better, and to promote new diagnosis, treatment, prevention, and drug development. Therefore, the application of cardiovascular disease knowledge graphs has broad prospects and can provide powerful support for the prevention, diagnosis, and treatment of cardiovascular disease. However, with the extensive research on cardiovascular diseases and the continuous emergence of research results, traditional knowledge graph construction requires a lot of human effort to carry out tasks such as data cleaning, named entity recognition, relation extraction, conflict discovery, and so on. Therefore, achieving automatic construction and updating of cardiovascular disease knowledge graphs can improve the accuracy and completeness of knowledge graphs, realize timely knowledge updates, and support personalized medical decision-making.

To this end, this work proposes an Auto-construction and Self-reflection Framework for Biomedical Knowledge Graph (ASK) for automatic construction and updating of a cardiovascular disease knowledge graph. The framework integrates named entity and relation extraction, link prediction, and conflict resolution techniques based on biomedical text content to achieve efficient, accurate, and reliable 24/7 automatic construction and updating of the knowledge graph. The ASK framework consists of three modules: automatic construction, graph update, and self-reflection. The construction module periodically retrieves medical data text from specified sources

and extracts entities and relationships using a natural language processing model based on BioBERT. This model increases the automation of knowledge graph construction and reduces the workload of manual data collection and graph construction. In the knowledge graph update module, entities and relationships automatically extracted from the text corpus are integrated into the existing knowledge graph. A conflict resolution algorithm is introduced to detect and resolve conflicts that may arise from new knowledge, improving the accuracy and quality of knowledge graph fusion. Through the self-reflection module, we predict implicit links in the knowledge graph and exclude the possible conflicts caused by implicit links, enriching the content of the knowledge graph, and adjusting and optimizing it accordingly. During conflict resolution, we can detect and resolve issues such as contradictions, errors, and duplicates in the knowledge graph, thus improving the accuracy and completeness of the predicted knowledge in real-time.

During our experimental stage, we conducted a series of model and algorithm comparisons to demonstrate the reasons for selecting specific models or algorithms in each module of the framework. We also provided the evolution of the knowledge graph in different modules of the ASK framework.

Keywords:

Knowledge graph construction, Knowledge graph embedding, Link prediction, Conflict resolution，Knowledge engineering

目录

第 1 章 绪论

1.1 研究背景和意义

1.1.1 心血管病的重要性和相关研究情况

心血管疾病是指影响心脏和血管的疾病，包括冠心病、心肌梗死、高血压、心力衰竭、心律失常等。这些疾病通常与心脏和血管的结构或功能异常有关，可能导致心脏无法有效地泵血或血管无法顺畅地输送氧和营养物质，最终可能导致心脏和其他器官的损伤和功能障碍。心血管疾病是全球范围内的主要健康问题之一，是导致死亡和残疾的主要原因之一[1]。常见的危险因素包括高血压、高血脂、糖尿病、肥胖、吸烟、缺乏体育锻炼、不健康的饮食、精神压力等[2]。心血管疾病可能导致严重的并发症和残疾，甚至危及生命。早期发现和治疗心血管疾病非常重要，同时预防心血管疾病的危险因素和保持健康的生活方式也是预防和控制心血管疾病的关键[3]。

尽管心血管疾病的研究已经取得了很大进展，但仍然存在一些问题和挑战。以下是其中的一些：1）样本大小：许多心血管疾病研究的样本大小较小，这可能会影响其结果的可靠性和适用性[4]。2）代表性样本：研究人员需要确保他们的研究样本具有代表性，以确保其研究结果的适用性[5]。3）数据来源：许多心血管疾病研究依赖于医疗记录和医疗保险索赔数据等次生数据源。这些数据源的质量可能有所不同，并且可能不完全反映真实的疾病情况[6]。4）研究方法：尽管心血管疾病研究中使用的研究方法已经得到了很大改进，但仍然存在一些挑战，如如何评估干预措施的效果、如何对干预措施进行随访等[7]。5）数据共享：由于心血管疾病研究涉及大量数据，因此数据共享是一个重要问题。研究人员需要确定如何处理、存储和共享这些数据，以便将来的研究人员可以使用它们[8]。6）资金和支持：心血管疾病研究需要大量的资金和支持，包括政府、私人和慈善机构的支持。然而，这些资源有时可能不足，这可能会限制研究人

员的能力和动力。总之，虽然关于心血管疾病研究已经取得了很大进展，但仍然需要克服一些问题和挑战，以便更好地理解这些疾病的本质，并为其预防和治疗提供更好的方案[9]。

1.1.2 知识图谱与其在医学领域中的相关应用

知识图谱是一种以图形形式表示实体之间关系的知识库。它是一种人工智能技术，用于捕捉现实世界中的知识，并在计算机系统中表示和组织这些知识[10]。下面知识图谱相关的一些知识：1）实体：知识图谱中的实体是现实世界中具有独立存在和特定属性的事物，例如人、地点、组织和事件等。2）属性：实体具有各种属性。3）关系：实体之间的关系可以用边来表示，例如人与人之间可以有亲属关系、工作关系和社交关系等。4）三元组：知识图谱中的基本元素是三元组，它由主体、谓语和客体组成，表示实体之间的关系。

知识图谱已经在各个领域得到了广泛的应用，其优势在于能够提高数据的质量和准确性，帮助人们更好地理解和利用大数据。1）搜索引擎：知识图谱可以帮助搜索引擎更好地理解用户的搜索意图，提供更加准确的搜索结果。2）语音助手：知识图谱可以帮助语音助手更好地理解用户的语音指令，并提供更加智能的回答和建议。3）智能客服：知识图谱可以帮助智能客服更加准确地识别用户的问题，并提供更加个性化的解决方案。4）智能问答系统：知识图谱可以帮助智能问答系统更加准确地回答用户的问题，甚至可以自动推断用户未明确表达的问题。5）智能推荐系统：知识图谱可以帮助智能推荐系统更加准确地理解用户的兴趣和需求，并提供更加个性化的推荐结果[11]。6）智能医疗：知识图谱可以帮助医疗行业更好地管理和利用医疗数据，提供更加精准的医疗服务和诊疗方案[12]。

生物医学知识图谱是基于生物医学领域知识的知识图谱，它将各种生物医学信息（如疾病、基因、药物、生物通路等）以及它们之间的关系构建成一张图谱，提供了一种方便有效的方式来组织、存储和查询生物医学知识[13]。生物医学知识图谱可以用于许多方面，如新药开发、临床诊断、治疗方案设计等。

举例来说，当医生需要为患者制定治疗计划时，可以通过生物医学知识图谱查询相关的基因、疾病、药物等信息，以便更好地理解疾病发生的机制和选择最合适的治疗方案[14]。近年来，随着大量生物医学数据的积累和技术的进步，生物医学知识图谱的构建和应用也受到了越来越多的关注。一些知名机构和公司，如美国国立卫生研究院（NIH）、欧洲生物信息研究所（EMBL-EBI）、谷歌等，都在生物医学知识图谱的研究和应用方面进行了深入的探索和实践。

1.2 国内外研究现状

1.2.1 生物医学知识图谱相关研究

生物医学知识图谱是一个重要的研究领域，涉及多个方面的技术和应用。数据源和收集方法：生物医学知识图谱的构建需要大量的生物医学信息，数据源包括医学文献、专利、疾病数据库、药物数据库、基因数据库等。数据收集方法主要包括人工标注、自然语言处理、机器学习等[15]。构建生物医学知识图谱的重要一步是实体识别和关系抽取。实体识别是将文本中的生物医学实体（如疾病、基因、药物等）识别出来，关系抽取则是提取不同实体之间的关系。目前常用的方法包括规则匹配、机器学习和深度学习等[16]。知识表示和存储：生物医学知识图谱需要将提取出来的生物医学知识转换为结构化的数据，并将其存储在知识图谱数据库中。目前常用的知识表示方法包括 RDF（Resource Description Framework）、OWL（Web Ontology Language）等[17]。应用领域：生物医学知识图谱在多个领域都有广泛的应用，如疾病预测、药物开发、精准医疗等。例如，通过生物医学知识图谱，可以发现基因、疾病、药物之间的关系，从而辅助药物开发和治疗方案的设计[18]。知识图谱质量评估：对于构建好的生物医学知识图谱，需要进行质量评估。目前的评估方法主要包括知识图谱完整性、准确性、一致性等方面的评估[18]。总的来说，生物医学知识图谱是一个重要的研究领域，目前的研究主要集中在数据收集、实体识别和关系抽取、知识表示和存储、应用领域和知识图谱质量评估等方面。未来，随着技术的不断发

展，生物医学知识图谱的研究和应用将会不断得到深入和扩展。

新型冠状病毒肺炎（COVID-19）疫情的爆发对全球政治经济格局产生了重大而深远的影响。业界和学界的合作努力下，各种具有针对性的新冠肺炎知识图谱研究项目相继展开。其中包括清华大学 Aminer 团队和智谱.AI 团队发布的新冠肺炎知识图谱"COKG-19"，该图谱构建了一个大规模、结构化的中英文双语新冠肺炎知识图谱，帮助研究人员识别和连接文本中的语义知识，并开展智能服务和应用。同时，开放医疗与健康联盟（Open Medical and Healthcare Alliance）与 OpenKG 合作完成了"新型冠状病毒肺炎诊疗知识图谱项目"，并在 HiTA 知识图谱服务平台上发布了相关数据集以进行开放共享。这些新冠肺炎知识图谱在疫情防控、药物研发、辅助诊断和智能问答等关键领域展示出了明显的应用效果。借助这些研究成果，基于知识图谱的上层应用逐渐丰富，为疫情防控工作的决策和部署提供了重要的支持。

1.2.2 国内生物医学知识图谱相关研究

中文医学知识图谱（Chinese Medical Knowledge Graph, CMKG）是由中科院计算所人工智能研究中心医疗知识图谱团队开发的一个基于中文医学文献的知识图谱。CMKG 主要包括三个部分：实体、属性和关系。实体部分包括药品、疾病、症状、检查、治疗方法等；属性部分包括药品功效、药品成分、疾病症状等；关系部分包括疾病和症状的关系、药品和功效的关系、症状和检查的关系等。CMKG 的构建依赖于大量的中文医学文献，采用了深度学习和知识表示学习等技术，通过实体识别、关系抽取和知识推理等方法来提高知识图谱的质量和精度。CMKG 的应用包括基于知识图谱的疾病风险预测、药物副作用预测、中药方剂推荐等[19]。

CMeKG 是中国医学科学院阜外医院开发的中文医学实体知识图谱。CMeKG采用了基于深度学习的实体识别和关系抽取技术，并且通过对医学文献和医学本体进行自动化分析和挖掘，实现了医学知识图谱的构建。CMeKG中包含了超过4.6万种医学实体，涵盖了疾病、症状、疾病分型、药物、药品分子、

人名、组织和解剖学结构等多个医学领域的实体，以及它们之间的复杂关系。除了实体和关系信息，CMeKG 还提供了详细的实体属性和本体信息，以支持更多的医学应用。CMeKG 已经被应用到了医学信息检索、疾病风险预测、药物副作用预测、基于知识的临床诊断等多个医学领域。总之，CMeKG 是一种基于深度学习和本体技术的中文医学知识图谱，其开发和应用在中文医学信息处理和医疗领域具有广泛的应用价值[20]。

中医药知识图谱（Traditional Chinese Medicine Knowledge Graph，简称 TCMKG）。TCMKG 整合了大量的中医药相关数据，包括中药、方剂、证候、病症等，并通过知识图谱的方式将这些信息进行关联。TCMKG 提供了可视化的查询界面，并支持自定义查询和可视化分析。它包含了中药材、中成药、方剂、证候等多个方面的知识，可为中医药的研究和应用提供支持[21]。

以上生物医学知识图谱案例在国内具有较高的知名度和影响力，它们的研究和应用为生物医学领域的发展提供了重要的支持和帮助。

1.2.3 国际生物医学知识图谱相关研究

人类表型本体（Human Phenotype Ontology，HPO）是一个用于描述人类表型的计算机化本体。它包含超过 1 万个独立的表型条目，每个表型都由一个唯一的标识符、定义和有关表型的注释组成。HPO 为医学领域的研究和临床实践提供了一种标准化的表型描述方式，为基因变异与表型之间的关系建立了桥梁，促进了对疾病发病机制的深入理解和相关治疗策略的制定。例如，使用 HPO，可以对患者的表型进行描述和分类，帮助医生对疾病的诊断和治疗做出更加准确的判断。HPO 还定义了不同表型之间的关系，以及表型和基因、疾病之间的关系。HPO 在遗传学和临床诊断方面得到了广泛的应用[22]。

药物基因组知识库（PharmGKB）是一个用于存储和分享与药物响应相关的基因和表型信息的数据库。PharmGKB 包含了丰富的数据资源，包括对药物代谢和药物作用的基因、基因型和表型信息，以及与药物剂量、药物不良反应、药物作用机制、临床指南等相关的文献和注释。其中，包含了来自药物研发和

临床实践中的大量数据，这些数据被整合并标准化为可搜索的、结构化的知识图谱，为药物研发和临床决策提供了基础数据支持。PharmGKB 包含超过 2 万个基因和表型条目，以及与超过 2 千种药物相关的信息。PharmGKB 中的数据可以用于预测药物响应，帮助医生做出更好的治疗决策[23]。

疾病基因关联数据库（DisGeNET）是一个用于存储和分享疾病和基因之间关系的数据库。DisGeNET是一个用于研究人类疾病的知识图谱，收集了来自各种生物医学数据库、文献和其他公开资源的疾病基因关联信息。该知识图谱提供了大量的疾病与基因之间的关联信息，其中包括遗传变异、基因表达、蛋白质互作和药物治疗等方面的信息。DisGeNET还提供了用于分析和可视化这些关联的工具和接口，方便研究人员使用和探索这些数据。DisGeNET 包含超过 12 万个疾病和基因之间的关系，以及这些关系的证据来源。DisGeNET的数据可以用于疾病研究和基因注释[24]。

1.3 本文的研究目的和意义

知识图谱作为一种新兴的语义 Web 技术，可以通过在各个领域的知识和数据之间建立关联，提供一种全面的、有机的知识组织和信息管理方式。在心血管疾病研究领域，知识图谱可以起到整合、标准化、深度挖掘和知识推理的作用，从而帮助研究人员更好地理解心血管疾病的本质、机制、诊断和治疗方法[25]。首先，知识图谱可以整合不同来源的数据，例如来自公共数据库、医疗机构、科研机构等的多样化数据，包括心血管疾病的基因、蛋白质、代谢物等生物信息数据、临床表现数据、病例报告、医学图像等，实现全面化数据管理。在这个基础上，通过知识图谱建模技术，可以构建出心血管疾病的知识图谱，从而使得研究人员可以更加全面、直观地了解心血管疾病的各个方面的知识[26]。其次，知识图谱还可以提高心血管疾病研究数据的可信度和可重复性。通过标准化数据的定义和管理，知识图谱可以避免数据存在冗余、不一致、缺失等问题，从而提高数据质量。同时，知识图谱的知识表示方式也可以为数据的重复利用提供了便利[27]。此外，知识图谱还可以为心血管疾病的深度挖掘和知识推

理提供支持。通过知识图谱上的关系和属性，可以实现心血管疾病的各种因素之间的关联分析、知识推断等，为深入探索心血管疾病的发病机制和预防、治疗方法等方面提供支持[28]。因此，构建心血管疾病领域的知识图谱具有重要的意义，可以为研究人员提供一个综合性、高质量的数据资源，帮助深入探索心血管疾病的本质和机制，为疾病预防、治疗和管理提供理论和实践支持。

1.4 论文研究内容和组织安排

本文的主要工作：针对心血管疾病医学文献，研究实现对心血管病知识图谱自动构建与更新的模型框架。实现对心血管疾病医疗文本的自动获取和相关实体和关系的自动抽取，将相关结果整合构建并更新心血管病知识图谱，并且基于预测和冲突消解相关知识实现自省过程。

第 1 章，介绍本文有关知识图谱相关工作的研究背景，具体的生物医学图谱研究内容，阐述本文研究内容对相关医学领域研究重要性。同时介绍本文所研究内容的国内外研究现状，寻找研究切入点。并针对各章节进行总结。

第 2 章，这一章主要介绍了本研究涉及的相关理论知识和基本方法。首先，对知识图谱的构建和更新模型进行了简要概述，然后对涉及到研究工作各模块的具体模型进行了详细介绍。最后，对知识图谱自动构建和更新任务做了总体概述。

第 3 章，本章为本文提出心血管病自动构建和更新框架 ASK。第一部分是自动构建模块，该模块可以实现自动获取数据并抽取图谱。第二部分是更新模块，可以不断地更新和完善知识图谱。第三部分是自省模块负责对知识图谱进行自我审查和评估。

第 4 章，本章是实验部分。实验部分重点介绍了所使用的知识图谱构建和预测实验数。其中构建采用的数据集包括构建时用的生物医学命名实体识别数据集和构建的生物医学关系抽取数据集。选取了具有代表性的工作，对构建出的知识图谱的评估结果进行了展示。

第 5 章，本章总结了研究内容，对研究的缺陷和不足进行了总结分析，并提出了未来的研究方向。

第 2 章 相关理论介绍

2.1 知识图谱构建相关研究

2.1.1 知识图谱构建方法

知识图谱构建方式可以分为以下几种：人工构建：人工构建是最早期也是最基础的一种方式。该方式需要人工从文本、数据库等来源中收集数据，并手动建立实体、属性、关系等元素，再进行知识图谱构建。优点是可以精细化控制知识图谱的构建过程，缺点是时间、人力成本较高，构建速度较慢。自动构建：自动构建是近年来知识图谱构建领域的研究热点之一。该方式主要利用自然语言处理、机器学习等技术对大量文本数据进行处理，提取实体、属性、关系等元素，并自动构建知识图谱。优点是构建速度快，成本较低，缺点是精确性需要进一步提高。半自动构建：半自动构建是一种介于人工构建和自动构建之间的方式。该方式利用人工标注或领域专家知识等手段对自动提取的实体、关系等元素进行校验和修正，以提高知识图谱的精确性。优点是结合了人工和自动的优点，缺点是需要领域专家的参与和投入较多的成本[29]。

按照构建的方向来分，则是分为自底向上和自顶而下的构建方法。自底向上是指从底层数据开始，逐步构建起一个完整的知识图谱，这个过程通常是通过从多个数据源中提取数据，然后通过各种算法和技术进行处理和集成，最终形成一个完整的知识图谱。而自顶而下则是先定义知识图谱的高层结构和主题，然后再逐步添加和填充具体的实体和关系，最终形成一个完整的知识图谱。

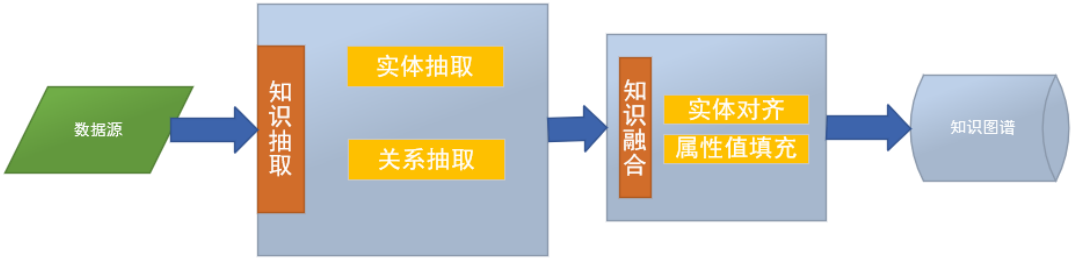

图 2.1 知识图谱自底向上构建方法

自顶而下的知识图谱构建方法则是先定义一个结构化的知识模板或者领域本体，然后基于这个模板或本体构建知识图谱。这种方法的优点在于能够保证知识图谱的准确性和完整性，但是缺点是需要手工定义知识模板或本体，对相关领域专业知识要求较高。

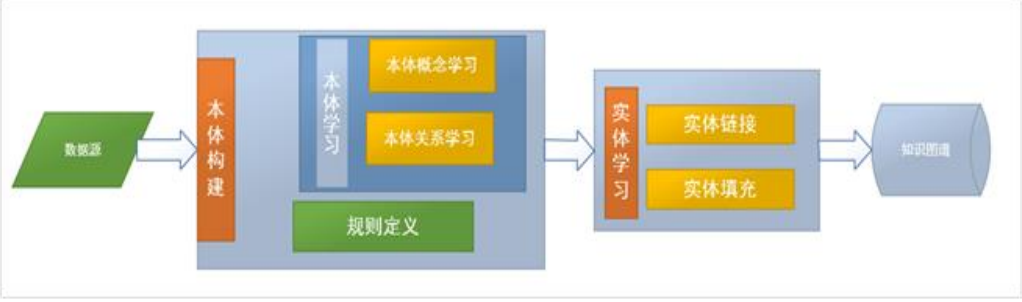

图 2.2 知识图谱自顶而下构建方法

2.1.2 知识图谱构建过程相关方法

知识图谱构建过程通常包括以下步骤：1）数据收集和清洗：从各种结构化和非结构化数据源中收集和提取数据，对数据进行清洗和预处理以消除噪声和冗余。 2)实体识别：从原始数据中识别出实体，如人物、组织、地点等，并对它们进行命名实体识别。3）关系抽取：从数据中抽取实体之间的关系，如“某个基因导致某种疾病”，“某个化合物能够治疗某种疾病”等，并进行关系类型分类。 4）知识表示：将实体和关系表示为图谱中的节点和边，同时为它们分配唯一的标识符，并创建相应的元数据。5）知识存储和查询：将知识图谱存储在特定的图数据库中，以便快速检索和查询，并使用图数据库的查询语言（如Cypher）执行各种查询操作。6）知识质量评估和改进：对知识图谱的完整性、准确性和一致性进行评估，并使用各种技术（如基于规则的方法、机器学习方法等）改进知识图谱的质量。

2.2 知识图谱更新相关方法

知识图谱的更新是一个持续的过程，需要及时更新和维护，以保证知识图谱的准确性和实用性。知识图谱的更新主要包括以下方面：实体识别和关系抽

取：通过自然语言处理技术，从文本中识别出新的实体，并抽取实体之间的关系，更新知识图谱中的实体和关系信息。数据融合：将来自不同来源的数据进行融合，以更新知识图谱的内容。知识补全：根据已有的知识，推断出一些未知的实体和关系信息，并将其添加到知识图谱中。知识推理：通过逻辑推理和推断，发现新的实体和关系信息，以及验证已有的实体和关系信息的正确性。

2.2.1 知识图谱融合相关方法

知识图谱融合是将多个知识图谱集成为一个更大、更完整的知识图谱的过程。它可以帮助整合不同领域、不同来源的知识，从而更全面地描述知识领域。下面介绍一些知识图谱融合的技术：实体对齐（Entity Alignment）：实体对齐是指将不同知识图谱中表示同一实体的实体进行匹配，并将它们映射为同一实体，从而达到知识图谱融合的目的。实体对齐的算法有许多，如基于相似度的算法、基于语义的算法和基于图匹配的算法等[30]。关系对齐（Relation Alignment）：关系对齐是指将不同知识图谱中相似或相互关联的关系进行匹配，并将它们映射为同一关系，从而达到知识图谱融合的目的。关系对齐的算法也有许多，如基于语义的算法和基于规则的算法等[31]。知识映射（Knowledge Mapping）：知识映射是指将不同知识图谱中的实体和关系映射到一个共同的语义空间中，从而进行融合。知识映射的方法包括基于词向量的方法、基于本体的方法和基于深度学习的方法等[32]。

2.3 知识图谱预测

链接预测是指根据现有的知识图谱中的实体和关系，预测未知实体之间的关系，或者在现有实体和关系的基础上，预测新的实体和关系。链接预测在知识图谱补全、推荐系统、社交网络分析等领域中都有应用。相关的知识主要包括：基于矩阵分解的方法：该方法将知识图谱表示为一个实体-关系矩阵，然后使用矩阵分解的方法对矩阵进行分解，得到实体和关系的低维度嵌入表示，从而预测新的实体和关系。代表性的方法包括 TransE[33]、TransH[34]、TransR[35]等。

基于图卷积网络的方法：该方法基于图结构对实体和关系进行嵌入表示学习，通过图卷积网络对实体和关系进行非线性变换，从而预测新的实体和关系。代表性的方法包括 GCN[36]、GAT[37]、KGAT[38]等。基于路径推理的方法：该方法通过从知识图谱中提取实体之间的路径，学习实体之间的关联信息，从而预测新的实体和关系。代表性的方法包括 PTransE[39]等。基于注意力机制的方法：该方法通过引入注意力机制，将实体和关系的重要性进行加权，从而预测新的实体和关系。代表性的方法包括 ComplEx[40]、ConvE[41]等。

链接预测（Link Prediction）是一种基于图谱的方法，用于预测知识图谱中尚未揭示的潜在关系。在图谱中，实体之间的关系可以表示为边（Edge），每个实体和关系可以表示为一个节点（Node），形成了一个复杂的图结构。

链接预测的基本思想是：在已知部分图谱中，我们可以找到已知实体之间的相似性模式（如共同的关系、相同的实体类型等），并将这些模式应用于未知实体，从而预测未知实体之间的关系。这个过程类似于基于相似性的推荐系统，即根据用户历史行为推荐其可能感兴趣的内容。

在知识图谱中，链接预测的主要任务是预测两个实体之间是否存在某种关系。预测关系的方法包括传统的机器学习方法（如逻辑回归、支持向量机等）和基于神经网络的方法。这些方法通过学习实体和关系的向量表示，捕捉实体和关系之间的相似性和差异性，并在此基础上进行预测。

链接预测可以被应用于各种领域，如社交网络分析、推荐系统、生物信息学等。在知识图谱中，链接预测可以用于实体关系发现、图谱补全、数据修正等任务，有助于完善和更新知识图谱，提高知识图谱的质量和准确性。

2.4 冲突消解相关内容

生物医学知识图谱可以用于许多应用，例如药物研发、疾病预测、基因表达等。例如，在药物研发中，可以使用生物医学知识图谱来查找潜在的治疗靶点，并预测药物的副作用。在基因表达方面，可以使用生物医学知识图谱来研究基因之间的相互作用，并预测基因调节网络。

当存在冲突的时候，可以使用冲突消解方法执行知识图谱的冲突消解，以消除知识图谱中的不一致和不完整。例如，在生物医学知识图谱中，可能存在多个实体之间的不一致关系，这些关系可以通过冲突消解来解决。

在知识图谱中，通常使用一组三元组（Subject-Predicate-Object）来表示实体之间的关系。其中 Subject 和 Object 表示实体，Predicate 表示它们之间的关系。例如，可以使用“疾病-症状-疼痛”三元组来表示“疼痛”是“疾病”的“症状”。在构建知识图谱时，可能会存在不一致和不完整的问题，例如，可能存在多个实体之间的不一致关系或缺失的关系等。解决这些问题的方法之一是使用冲突消解技术。冲突消解用的就是冲突消解规则[42]。冲突消解规则可以看作是一种约束，它规定了知识图谱中的知识之间的关系，限制了知识图谱中的冲突产生。在 PySAT 中，冲突消解规则被转化为 CNF 公式，成为 SAT 求解器需要满足的约束条件之一。

冲突消解技术的目标是在知识图谱中寻找不一致的三元组，并将它们进行合并或删除，从而消除知识图谱中的不一致和不完整。这个过程可以通过谓词逻辑来表示。谓词逻辑是一种形式化的逻辑系统，用于表示命题、关系和量词等。在知识图谱中，谓词逻辑可以用来表示三元组和它们之间的关系。将冲突消解方法引入生物医学知识图谱处理中的创新点在于能够解决知识图谱中不一致、矛盾或错误的问题。传统的知识图谱处理方法主要依赖于规则和模式匹配，但这些方法无法解决知识图谱中存在的冲突问题。

利用冲突消解方法可以检测和解决知识图谱中的矛盾、错误、重复等问题，提高知识图谱的准确性和质量。而将 miniSAT 等 SAT 求解器引入冲突消解处理中，不仅能够快速高效地解决知识图谱中的冲突问题，还能够利用其灵活性和可扩展性对知识图谱进行定制化处理，提高知识图谱处理的效率和可靠性。此外， SAT 求解器还可以通过自动化的方式检测和修复知识图谱中的错误和矛盾，提高知识图谱的可维护性和可靠性。

2.5 本文研究相关问题

2.5.1 知识图谱构建现有问题

在知识图谱构建的过程中，仍然存在一些挑战和问题，其中一些主要问题包括：(1) 知识获取：知识获取是知识图谱构建的第一步，它通常包括结构化和非结构化数据的获取。结构化数据通常来自于数据库和表格等，而非结构化数据则包括文本、图片、音频和视频等多媒体数据。在多语言和多领域的情况下，数据的获取和处理就更加困难，需要借助机器翻译、自然语言处理等技术来处理数据[43]。 2）知识表示：知识图谱中的知识需要用一种形式化语言来表示。目前，大部分知识图谱使用的是 RDF/OWL 等形式化语言来表示知识。然而，这些语言不太适合表示复杂的语义关系，如事件、时间和空间等。为了解决这个问题，近年来出现了一些新的知识表示方法，如基于图神经网络的表示方法、基于语义网格的表示方法等[44]。 3）知识融合：知识图谱通常包含来自不同数据源的知识，如维基百科、Freebase 等。这些知识源中的信息可能存在重叠或者冲突，因此需要进行知识融合。知识融合的过程需要解决实体对齐、属性映射、关系匹配等问题[45]。 4）知识推理：知识推理是知识图谱的重要功能之一，它可以帮助我们发现知识之间的关系和规律。知识推理通常包括基于规则、基于统计的方法等。然而，当前的知识推理技术仍然存在一些局限性，如处理不确定性和推理效率等[46]。 5）知识维护：知识图谱是一个动态的结构，它需要不断地更新和维护。在知识图谱更新的过程中，可能会出现一些错误或冲突，这需要人工干预来进行修正。同时，随着知识图谱的规模不断扩大，其维护的难度也会逐渐增加。因此，需要开发出自动化的知识维护工具来解决这个问题[47]。

2.5.2 生物医学命名实体识别现有问题

生物医学命名实体识别存在许多挑战和问题。 1）实体边界识别：在生物医学文本中，实体边界的识别可能受到一些特殊字符或标点符号的影响，例如

“-”、“/”等符号，同时一些实体可能具有多个名称，需要对这些名称进行识别[48]。2）实体类型分类：在生物医学领域，实体类型的分类通常比较复杂，例如“胰腺癌”既可以归类为“疾病”类别，也可以归类为“器官”类别。此外，一些实体可能存在混淆的情况，例如“STAT3”既可以被识别为基因，也可以被识别为蛋白质[49]。 3）实体消歧：在生物医学文本中，同名实体的消歧尤为重要，例如“ACE”可以是血管紧张素转化酶，也可以是肿瘤抗原，需要根据上下文信息进行消歧[50]。 4）实体关系抽取：在生物医学领域，实体之间的关系通常比较复杂，例如“基因突变导致癌症”这一关系需要进行多个实体之间的抽取。同时，一些关系可能难以通过语言规则来表达，需要考虑结构化信息和语义信息进行抽取[51]。

2.5.3 生物医学关系抽取现有问题

生物医学关系抽取存在以下问题： 1）数据稀缺性：生物医学领域的数据通常比较稀缺，这使得使用传统的监督学习方法进行关系抽取困难。这也限制了训练模型的能力，并可能导致模型在新的数据集上的泛化性能较差[52]。 2）实体和关系的多样性：在生物医学领域，实体和关系的类型比较多样化，涉及到各种生物学实体、组织、疾病、药物、基因等等。这使得关系抽取的任务变得更加复杂，需要考虑到各种不同类型实体之间的关系[52]。 3）关系表示的多样性：生物医学领域中的关系可能具有不同的表示方式，例如文本中的语言表达、知识图谱中的关系等等。这使得关系抽取的任务更具挑战性，需要使用多种不同的技术来解决。总之，生物医学关系抽取是一个具有挑战性的任务，需要考虑数据稀缺性、实体和关系的多样性、关系表示的多样性以及领域知识的利用等问题[52]。为了解决这些问题，需要使用多种不同的技术和方法，并结合领域专业知识来提高关系抽取的准确性和效率。

2.5.4 生物医学实体对齐现有问题

在生物医学领域中，由于实体名称存在多样性、异构性等特点，导致生物

医学知识图谱实体对齐存在以下问题： 1）实体异构性：在生物医学知识图谱中，同一实体可能由不同的知识图谱构建者使用不同的名称和属性描述，这就导致同一实体在不同知识图谱中被表示为多个不同的实体。这个问题被称为实体异构性，会给实体对齐带来困难[53]。 2）实体名称多样性：在生物医学领域，同一实体可能有多个不同的名称，包括标准名称、同义词、缩写等等。这就导致在不同的知识图谱中同一实体可能被使用不同的名称表示[54]。 3）知识图谱不完整性：由于生物医学领域知识的广泛和复杂性，生物医学知识图谱往往是非常庞大和复杂的，但是其中的知识并不完整。这意味着有些实体可能没有在所有知识图谱中被表示，或者表示不完整，这就给实体对齐带来了挑战。总之，在生物医学知识图谱实体对齐中，实体异构性、实体名称多样性和知识图谱不完整性需要解决的主要问题，需要采用多种不同的方法和技术来克服这些问题。

2.5.5 生物医学图谱补全现有问题

生物医学图谱补全是指通过自动化方法从现有生物医学图谱中识别出遗漏的实体、关系和属性，并将它们添加到图谱中的过程。虽然生物医学图谱补全可以提高生物医学研究和医疗决策的效率和准确性，但是在实践中仍然存在一些问题。 1）知识表示：生物医学知识的复杂性和多样性使得如何表示知识成为一个挑战。在生物医学图谱补全中，需要选择合适的知识表示方式，以确保表示的知识能够被有效地集成到现有图谱中[55]。 2）知识融合：生物医学领域中存在大量的异构知识源，这些知识源可能采用不同的知识表示方式和命名方式。在进行生物医学图谱补全时，需要将这些知识源进行融合，以便将信息集成到现有图谱中[56]。 3）表示学习：生物医学领域中存在大量复杂的实体和关系类型。在生物医学图谱补全中，需要对这些实体和关系进行有效的表示学习，以便能够从大量的文本和非结构化数据中进行自动化识别和抽取。综上所述，生物医学图谱补全需要克服数据质量、知识表示、知识融合和表示学习等方面的问题，才能实现对生物医学领域知识的全面识别和建模[55]。

2.6 本章小结

本章的重点在于介绍本文框架所涉及的概念和理论知识。第一小节将介绍与构建相关的技术和方法；第二小节将分类知识图谱更新所使用的知识和方法；第三小节将介绍与知识图谱预测相关的知识和经典模型；第四小节将介绍知识图谱与冲突消解相关的知识；最后，第五小节将探讨现有的知识图谱构建和更新技术中存在的问题。

第 3 章 心血管疾病知识图谱的自动构建与更新框架

在生物医学领域，知识图谱是一个重要的工具，它可以将大量的生物医学知识以一种结构化的方式进行表达和组织，从而帮助研究人员更好地理解和利用这些知识。知识图谱包含了丰富的知识信息和多样的数据类型，例如实体、属性、关系等，这些信息可以帮助研究人员进行生物医学知识的计算机化处理和分析。

然而，构建和维护一个高质量的生物医学知识图谱仍然是一个具有挑战性的问题。这是因为生物医学领域的知识非常丰富和复杂，涉及的领域也非常广泛，包括生物学、医学、药学等多个学科，因此构建一个全面且准确的生物医学知识图谱需要大量的人力和物力资源。

为了解决这个问题，本文提出了 ASK（Auto-construction and Self-reflection Framework for Biomedical Knowledge Graph，ASK）框架，它由三个模块构成，分别是图谱自动构建、图谱更新和图谱自省模块。具体地说，图谱自动构建模块采用了一系列的方法和技术来自动地从不同的数据源中提取和组织知识，以构建一个初步的生物医学知识图谱。图谱更新模块则通过监控和分析最新的数据源来不断地更新和完善知识图谱，以保证知识图谱的时效性和准确性。图谱自省模块则负责对知识图谱进行自我审查和评估，以检测和修复可能存在的错误和缺陷，从而提高知识图谱的质量和可靠性。

总之，ASK 框架是一个全面而有效的生物医学知识图谱构建和维护框架，其能够对医学文献进行智能化的抽取和整合，并且提供了可靠和可扩展的架构来支持生物医学研究和应用。为生物医学研究和应用提供了重要的支持，可以帮助研究人员更好地了解生物医学领域的知识，促进医学科学的发展和创新。

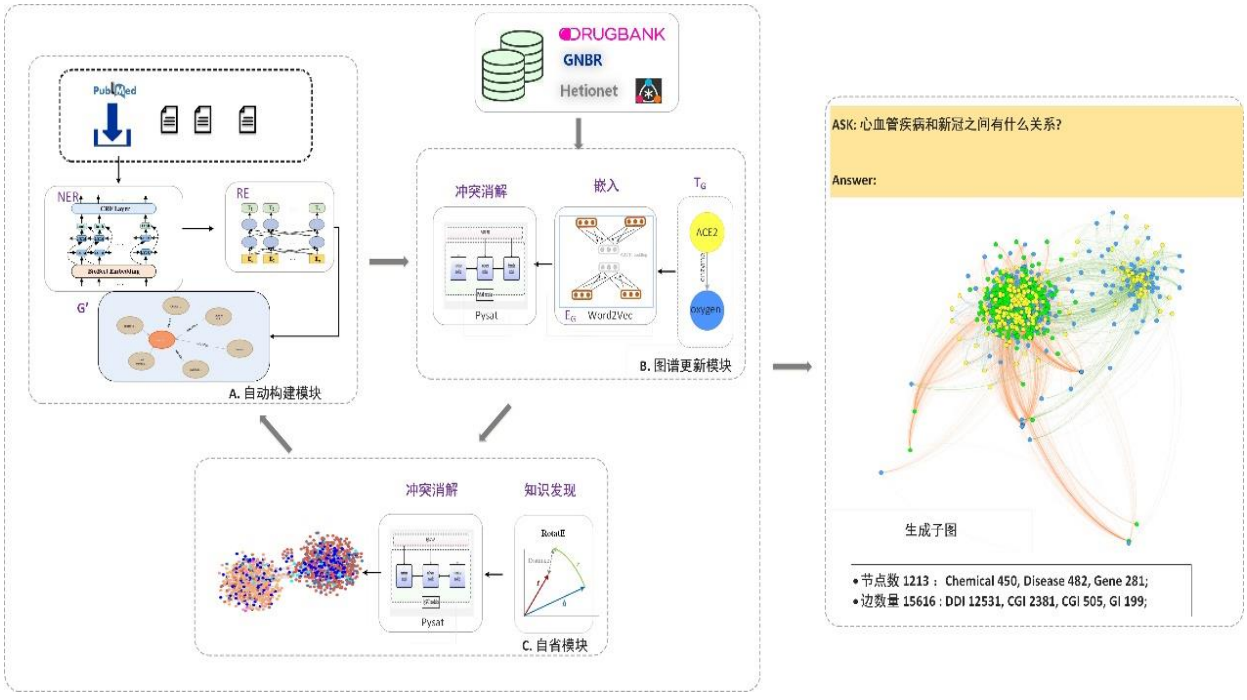

图 3.1 ASK 框架图

3.1 ASK 整体框架介绍

3.1.1 自动构建模块

图谱自动构建模块是 ASK 框架中的第一个模块，其主要任务是从大量的生物医学文献数据库中获取相关文章，并通过实体识别和关系提取技术自动构建一个新的生物医学知识图谱 G_d。这个模块的核心技术是自然语言处理和机器学习，它们能够帮助我们从非结构化的文本中提取有用的信息，并将其转化为结构化的知识图谱。通过实现对于医疗文本的自动获取和知识抽取工作，实现图谱构建的自动化实现。

具体地说，实体识别是图谱自动构建模块中的一个重要环节，其主要任务是从文本中识别出生物实体，例如基因、蛋白质、疾病等。为了实现这个目标，我们需要使用一些生物医学实体识别工具，例如 BioBERT+BiLSTM+CRF 模型，它们可以帮助我们自动地从文本中识别出生物实体，并将其映射到知识图谱中的实体节点。关系提取是图谱自动构建模块中的另一个重要环节，其主要任务是识别文本中实体之间的关系。为了实现这个目标，我们需要使用一些已有的

关系提取工具，它们可以帮助我们自动地识别文本中的关系，并将其映射到知识图谱中的关系边。

通过实体识别和关系提取技术，我们可以从大量的生物医学文献中自动构建一个初步的生物医学知识图谱 G_d。然而，在实际应用中，这个初步的知识图谱可能存在一些错误和缺陷，因此我们需要对其进行进一步的修正和优化。这就需要用到 ASK 框架中的图谱更新和图谱自省模块，以保证知识图谱的准确性和可靠性。

3.1.2 更新模块

在图谱更新模块中，ASK 框架采用嵌入方法来表示知识图谱中的实体和关系。嵌入方法是一种将实体和关系映射到向量空间的技术，它将每个实体和关系表示为一个向量，使得它们可以被更好地处理。嵌入方法有许多种实现方式，例如 Word2Vec[57]、TransE、DistMult 等。

在 ASK 框架中，我们使用的是 word2vec 模型来进行嵌入表示。Word2Vec 可以通过学习知识图谱中的实体和关系的嵌入向量，来建立实体和关系之间的语义关联。这些向量可以用于计算实体之间的相似性，从而支持知识图谱的匹配和推理。在嵌入表示完成后，ASK 框架会进行图谱更新，主要任务是冲突消解。冲突消解是指在知识图谱中出现了矛盾或冲突的情况，例如同一个实体有多种不同的属性或不同的实体有相同的属性。ASK 框架会使用冲突消解算法来解决这些问题，保证知识图谱的一致性和准确性。在完成上述任务后，将更新后图谱内容与原有图谱融合。通过图谱更新模块，ASK 框架可以不断地优化和完善知识图谱，提高其质量和可靠性。同时，由于嵌入方法的使用，知识图谱的表示也变得更加高效和便于处理。通过引入了冲突消解方法，提高了知识图谱的准确性。

3.1.3 自省模块

图谱自省模块是 ASK 框架中非常重要的一环，它能够不断地对知识图谱进