学　　号　2014302580045
密　　级

武汉大学本科毕业论文

面向多源异构的医学知识融合技术研究

院 (系) 名 称：　计算机学院

专 业 名 称：　软件工程

学 生 姓 名：　熊燚铭

指 导 教 师：　李兵　教授

二〇一八年五月

BACHELOR'S DEGREE THESIS
OF WUHAN UNIVERSITY

Research on Fusion Technology Oriented to Multi-Source Heterogeneous Medical Knowledge

School (Department): Computer School

Major: Software Engineering

Candidate: Yiming Xiong

Supervisor: Prof. Bing Li

Wuhan University

May, 2018

郑 重 声 明

本人呈交的学位论文, 是在导师的指导下, 独立进行研究工作所取得的成果, 所有数据、图片资料真实可靠. 尽我所知, 除文中已经注明引用的内容外, 本学位论文的研究成果不包含他人享有著作权的内容. 对本论文所涉及的研究工作做出贡献的其他个人和集体, 均已在文中以明确的方式标明. 本学位论文的知识产权归属于培养单位.

本人签名: _____ 日期: _____

摘　　要

随着人工智能的不断发展，其在医疗领域的应用始终是研究的热点之一。智能医疗的实现离不开医学知识的支撑，而作为知识来源的医学数据集种类丰富且规模庞大，面对大量的多源异构数据集，如何有效地获取和融合知识，并对知识进行合理的表示、解释与推理是重要的研究课题。

心血管疾病作为致死率最高的慢性疾病，对智能医疗有着急切的现实需求。因此，本课题以心血管疾病为切入点，主要包括如下几方面的研究内容：

(1) 知识获取。研究如何将半结构化的网络数据、病历数据、文献资料进行结构化的方法，以及从描述文本中抽取实体建立关系的方法。

(2) 知识融合。研究将数据映射到标准的知识结构的方法，对同构知识进行对齐与归类的方法，以及将异构知识进行结构化融合的方法。

(3) 知识推理。研究所构建的知识图谱如何满足实际的应用需求。

本课题最终构建了一个心血管疾病知识图谱，包括 173781 个实体和 2132977 个关系，并基于图谱实现了简单的语音问答系统，可根据问题查询相关医学知识。

关键词: 智能医疗; 知识图谱; 心脑血管疾病; 知识融合技术;

ABSTRACT

With the continuous development of artificial intelligence, its application in the medical field has always been one of the hotspots of research.The realization of intelligent medical treatment can not be separated from the support of medical knowledge, and medical data sets as a source of knowledge are rich in variety and large. Facing a large number of multi-source heterogeneous data sets, how to effectively acquire and fuse knowledge, and to reasonably express, explain and reasoning knowledge is an important research topic.

Cardiovascular diseases, as the highest mortality rate of chronic diseases, have urgent practical needs for intelligent medicine. Therefore, this topic takes cardiovascular diseases as the breakthrough point, mainly including the following aspects:

(1) Knowledge Acquisition. This paper studies how to structurate semi structured network data, medical records, and literature, and how to extract entities from the description text.

(2) Knowledge Fusion. The method of mapping data to a standard knowledge structure, the method of aligning and classifying isomorphic knowledge, and the method of structured fusion of isomeric knowledge.

(3) Knowledge Reasoning. Use the knowledge graph constructed by the research to meet the actual application requirements.

This project has finally constructed a knowledge graph of cardiovascular disease, in-

cluding 173781 entities and 2132977 relationships, and a simple voice question answering system based on the graph, which can be used to query related medical knowledge according to the problem.

Key words: Intelligent medical; artificial intelligence; cardiovascular

目　　录

1 绪论

1.1 背景与意义

智能医疗是人工智能最有希望取得突破之处。医疗数据种类齐全，规模庞大，且医疗领域研究成果丰硕。但与此同时，由于医学知识通常服务于各类医学从业人员，其结构化程度、多源融合程度较低，不便于知识的智能化表达。就医学与计算机跨领域研究而言，融合多源异构且规模庞大的数据集并对医学知识进行合理的表达、解释与推理是一个十分重要的研究课题。

医学知识海量复杂的特点意味着丰富多样的自动化方法，表达相同概念的知识需要相互补充与对齐，表达不同概念的知识需要结构化并融合。另外，知识本身是动态发展、不断扩充的，对知识时效性与跨媒体表示也是重要的研究内容。随着语义网到知识图谱的发展，可处理的知识规模日渐增大，以知识图谱为主的知识表示方法成为主流。在知识自动化过程中，以知识图谱为理论基础的知识融合方法可以有效的保证知识融合过程中的完整正确性。

知识图谱是用于表示知识结构关系的图结构模型，用于描述知识的资源及知识载体，并挖掘、分析、构建、绘制和显示知识及它们之间的相互联系 [13]。知识图谱的构建是一个反复迭代的过程，图谱中节点个数影响着网络的结构复杂度及推理的效率和难度，构建针对性的知识图谱再进行扩展完善比起直接建立覆盖全面、内容广泛的知识图谱更具操作性。

现代医学领域，心血管疾病一直位居各种死因首位，严重威胁人类的正常生存。心血管疾病具有高患病率、高致残率和高死亡率的特点，全世界每年死于心血管疾病的人数高达 1500 万人 [14]。本研究将以心血管疾病为切入点，研究多源异构的医学知识融合技术。

在心血管医学方面，知识图谱相关融合技术可以将复杂的医学知识通过实体关系展示出来，揭示心血管疾病的知识关系和动态发展规律。同时，知识图谱创

造的检索条件将为医学研究者进一步的研究提供切实的、有价值的参考。在计算机方面，研究为其他领域的知识融合提供了技术条件。

1.2 研究现状

1.2.1 智能医疗

近年来，智能医疗相关研究成果丰硕。涉及的周边产业范围很广，设备和产品种类繁多 [15]。其影响将不仅仅限于医疗服务行业本身，还将直接触动包括网络供应商、系统集成商、无线设备供应商、电信运营商在内的利益链条，从而影响通信产业的现有布局 [15]。

由于医疗本身对数据统计的要求，医学知识数据集准备相对完善，各种类型医学知识开源数据集应有尽有，医学领域与大数据分析，数据挖掘等领域的结合也成为未来发展的必然趋势。

1.2.2 知识图谱

知识图谱由语义网发展而来，其融合了语义网对知识组织、表达与推理的方法，通过关系的形式构建知识，使得知识更易在计算机之间和计算机与人之间进行交换、流通与加工 [16]。具体而言，知识图谱由模式图、数据图以及两者关系组成：模式图对人类知识领域的概念层面进行描述，强调概念及概念关系的形式化表达，模式图中节点是概念实体，边是概念间的语义关系，如 part-of；数据图对物理世界层面进行描述，强调一系列客观事实 [16]。数据图中的节点有两类，一是模式图中的概念实体，二是描述性字符串，数据图中的边是具体事实的语义描述；模式图和数据图之间的关系指数据图的实例与模式图的概念之间的对应，或者说模式图是数据图的模具 [16]。

著名的通用知识图谱有很多，如 DBpedia，freebase 等，它们规模庞大、领域宽泛，包含大量常识。医学领域知名的知识图谱有 IBM Watson Health 等。知识图谱是智能大数据的前沿研究问题，它以独有的技术优势顺应了信息化时代的发展，比如渐增式的数据模式设计；良好的数据集成；现有 RDF、OWL 等标准支持；语义搜索和知识推理能力等。在医学领域，随着区域卫生信息化及医疗信息系统的

发展，积累了海量的医学数据。如何从这些数据中提炼信息，并加以管理、共享及应用，是推进医学智能化的关键问题，是医学知识检索、临床诊断、医疗质量管理、电子病历及健康档案智能化处理的基础。

1.2.3 知识融合

知识融合是在信息融合的基础上发展起来的一个新的融合概念。早期对于它的研究多将其作为知识工程的一部分与其他相关内容结合起来。知识融合在含义上与信息融合有交叠的部分，因此，知识融合的研究可以借鉴信息融合已有的一些研究成果 [11]。此外，本体技术、语义网、多 Agent、知识组织和数据挖掘等领域的快速发展也促进了知识融合的研究和应用 [11]。

现有知识图谱的研究与应用大致可以分为两类：第一类定义以 KRAFT 项目的相关文献为代表，该定义认为知识融合是指从众多分布式异构的网络资源中搜索和抽取相关知识，并转换为统一的知识模式，从而为某一领域的问题求解构造有效的知识资源 [17]；第二类定义强调集成过程的结果是新知识的产生，认为知识融合是一种服务，它通过对来自分布式信息源的多种信息进行转换、集成和合并等处理，产生新的集成化知识对象，同时可以对相关的信息和知识进行管理 [6]。

1.3 研究主要内容

研究将以心血管疾病为切入点，用现有的医学文献、网站数据、电子病历等多源异构数据形式融合成为一个心血管疾病的知识图谱。深度探寻心血管疾病的发病、患病、病死与生物环境、临床特征、遗传等因素的相关关系。研究的主要内容包括三个部分，知识获取、知识融合和知识表示与推理。

知识获取部分主要阐述将半结构化的网络数据、病历数据、文献资料结构化的方法和从描述文本中抽取实体建立关系的方法。

知识融合部分是研究的重点，包括了将数据映射到标准的知识结构的方法、对同构知识进行对齐与归类的方法以及将异构知识进行结构化融合的方法。

知识推理是研究的落地阶段，该部分更关注融合后的知识图谱直接的或潜在的应用价值。

1.4 研究框架

研究框架作为指导研究过程的基础，记录了研究使用的基本技术方案及处理流程。研究从原始数据到形成应用系统的过程中，会用到多种处理技术，具体技术选型需要根据研究所处理数据的实际情况而定。

图 1.1　模块架构

如图1.1 所示研究的模块架构。除数据源模块外，技术模块包括数据获取、图谱构建与应用系统，技术选型仅在本章说明，以后章节将以研究内容为主。在研究的技术选型过程中，需选择最适合数据源的技术底层。

在数据获取模块中，使用 Scrapy 建立爬虫系统，随后将数据交给分布式业务

逻辑系统，使用 MapReduce 的方式抽取实体、建立关系，并以实体图谱映射模型（OGM）转化保存于 Neo4j 图数据库。控制迭代过程的词典数据以 ORM 模型保存于 KV 数据库，控制数据的获取过程。

知识图谱构建过程是流程化、模块化且循环迭代的，人类的认知过程不断的发展进步，知识图谱也需不断的扩充。

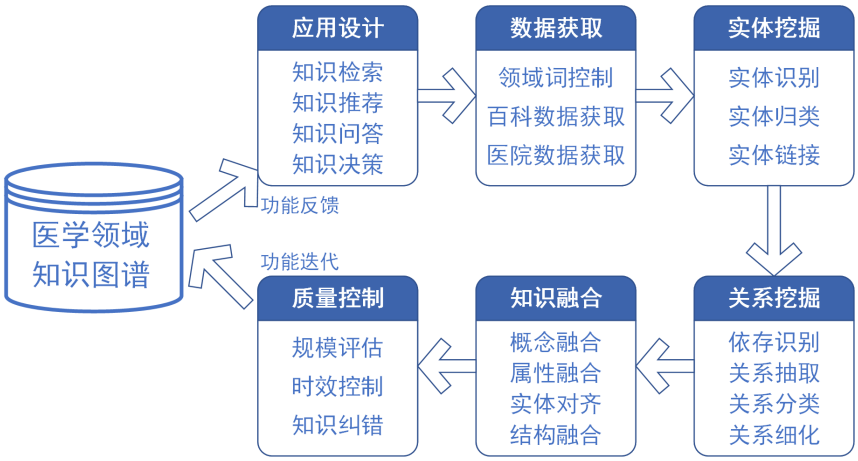

图 1.2　知识图谱构建流程

如图1.2 所示知识图谱迭代流程及相关技术，知识图谱需要不断的进行迭代与更新，以保持知识的完整性与时效性。应用设计决定数据获取处理方向，构建起的知识图谱通过使用时的功能反馈，在应用层设计新的功能，然后再次从数据获取阶段开始融合知识，迭代构建更大规模，更高质量的知识图谱。

2 知识获取

面对海量复杂且不同来源的医学文献、影像、病历、诊断方法和网络资料，以图结构为底层的数据表示方法可以有效的实现多源异构知识数据的融合。研究以心血管疾病为构建起点，从多源异构数据中自动、迭代的抽取医学实体、关系及属性，构建中英文对齐的医学知识图谱。

在处理海量文献资料时，实体关系提取技术希望有效的过滤与领域无关的文本数据。当得到文本数据后，需要通过自然语言处理来识别文章中的领域实体，逐步将非结构化数据结构化。对于病历、文献、专家经验等结构化资料，通过构建结构化知识映射来转化为可表达的实体关系数据。

实体识别技术需要多种成熟的技术支持，其中包括分词技术、词性标注、以及词向量处理等技术。研究要求较高的知识准确性和领域性，在当前识别技术召回率无法达标的情况下，借助三方或者自建的词汇知识库十分必要。通过词汇库，文本的处理结果可以保持较高的领域针对性，在此基础上进行语义抽取时，可以获得更优的抽取结果。

文本实体关系识别目前仍然是一个研究难题，通常研究会使用句法结构拆分，依存分析或者语义解析等技术，从而识别实体间的关系、事件的触发条件或是实体在关系中的角色。本研究将抽取几种明确、清晰的语义关系作为研究的基础。丰富的语义关系可以为知识图谱提供更详细准确的表达，比如症状到达某种程度与某疾病相关。当前阶段，复杂语义关系抽取不在研究考虑范围，但处理后的实体关系保留完整的描述文本。如果想要获得详细的语义关系，对抽象模糊的相关关系进行细化时，可以对图谱实体关系描述文本进行进一步分析。

2.1 启动要求

为确保医学知识图谱有较高的领域性，初始知识获取工作必须包含以下几条原则：

- 知识来源权威
- 高医学相关性
- 知识描述清晰

本研究所使用的数据集主要包含了无结构的专家诊疗经验记录；无结构的医学文献资料；半结构化的 Wiki 词条，医学百科词条的文本描述；结构化的百科医学词典，以及来自同济医院的部分结构化病历。研究以半结构化的网络数据为主要数据来源，结构化的医学词典为对齐标准，融合病历、文献、专家诊断等多种来源的医学知识数据。

2.2 存储标准制定

数据来源广泛多样，为避免数据源的的结构化或半结构化特征在存储过程被破坏，需要按数据源制定相应的存储标准。数据存储标准的制定是保证数据多源异构特性的基础，决定未来数据融合以及对齐的方法，高效的数据结构基础可以加速知识图谱建立工程的进度。

就当前医学领域数据现状而言，多数来自医院的数据集并未很好的结构化，来自网络的数据集存在不标准、不完整等问题。建立知识图谱要求底层数据满足结构化，且尽可能标准完整。不同类型的数据适合不同的数据存储标准，处理过程中根据实际情况来确定。

表 2.1　领域词表存储标准

字段	中文解释	可空	示例
name	词汇名称	true	冠心病
is_searched	是否已查	false	false

2.2.1 领域词典存储

领域词表为领域概念知识获取的基础。如表2.1 所示，词表保留最简单的逻辑结构，用于控制信息的查询过程。词表里只存储领域词汇，其他非领域词汇不进行保存。

2.2.2 领域概念存储

如表2.2 所示，以冠心病为存储示例，描述了的领域概念的存储方案。词汇实体用 name 字段记录，研究以中文名为标准。除此之外，记录与实体相关的中英名词信息，用于相同概念识别的工作。

表 2.2 领域概念存储标准

字段	中文解释	可空	示例
name	词汇实体	false	冠心病
name_zh	中文标准	true	冠心病
name_en	英文标准	true	coronary disease
name_alias	中英别名	true	{冠状动脉疾病}
category	类别	true	{I25.101, 疾病, 循环系统疾病, 急救}
desc	描述	true	冠心病，全称冠状动脉粥样硬化性心脏病...
desc_mention	提及领域词	true	冠心病，心脏病...
sub_concept	子结构集合	true	{病因, 诊断,...}
sub_concept_des	子结构描述	true	{病因描述, 诊断描述,...}

许多领域概念的相关语义在医学领域已经被很好的总结（比如疾病的病因、诊断等），这些已经结构化的数据在获取时需要合理的存储。实际操作时，领域概念以 KV 结构存储，可以方便的存储子结构的结构化特征，如在子结构“诊断”中包括“问诊”、“叩诊”等更深层次的子结构，这里为便于展示领域概念，不将子结构

展开。

获取到网络上的文本描述信息后，将已经半结构化的信息按照规则映射到领域概念的存储标准中。如与冠心病鉴别诊断相关的描述文本存储在冠心病实体下子结构鉴别诊断的描述字段内，建立起最基本的语义关系，不同数据源统一映射，相互补充，获得关于领域概念的实体数据集。

领域概念包括疾病、药品、辅助检查、手术知识等等，其整体结构相同，子结构有所差异。子结构所描述的语义关系有所变化，如药品的子结构中会包括发展历史、化学结构、理化性质、作用机理、不良反应、使用注意事项等语义描述字段，这里不再赘述。

2.2.3 病历存储

医学领域已有完备、准确的电子病历存储方案，国内国外均有来自官方的存储标准。这些标准用于医生快速理解病历、获取重点，在医疗实践过程中发挥了重要作用。

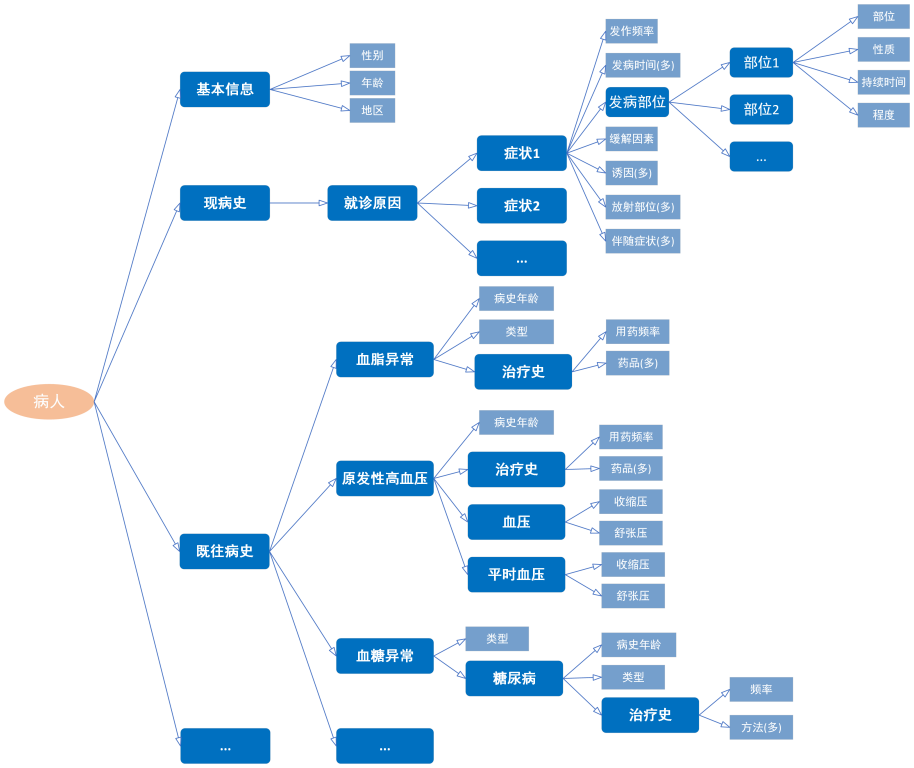

图 2.1　结构化病历存储标准

然而，为了方便医生理解，这些电子病历结构化不彻底，且在执行过程中，不同科室的病历存在不同的描述模板（如内科与五官科进行的检查不同，描述模板侧重点存在差异），为了计算机可以更好的理解病历中所记录的与病人疾病相关的知识，需要采用结构化的存储标准或将文本映射到结构化的存储标准上。

病历本身结构复杂，数据种类多且稀疏，结构化后内容深度无法控制，关系型存储逻辑不能完全满足病历存储要求，故病历采用 KV 存储，主体结构如图2.1所示，部分冗余内容省略，省略内容包括危险因素、体格检查、常规检查和特殊检查等病历子条目，详细结构可在附录中查阅。病历作为带标签（确诊的疾病）的知识，在后面知识融合表示过程中，采用逻辑上分离，结构上融合的基本处理方式。

2.2.4 文献存储

文献是一种无结构的数据来源，文献资料中包含了当前医学研究的最新进展，对医学从业人员的参考价值巨大。然而，对文献本身进行结构化是一件困难的工作，当无法很好的理解作者观点时，结构化过程可能出现预料不到的结果。因此，对医学文献的存储以跨媒体无结构的融合方式最佳，这种处理方式的落地目的是为医学工作者提供合适的参考资料，不对文献内的内容进行推理。

表 2.3 文献存储标准

字段	字段解释	可空
name	文献名称	false
author	文献作者	false
time	发表时间	false
abstract	文献摘要	true
mention	提及领域词汇	true
location	存储位置	false

如表2.3 所示，文献只需保存几种基本的字段，包括文献的名称、作者等。文

献的领域结构化特征在实体抽取过程中进行，主要获取文献所提及的领域词汇及频率，用于在图谱中通过拓扑结构快速找到合适的可参考文献。

2.3 迭代获取

对于网络数据而言，研究的数据获取过程是迭代的，领域性的。如图2.2 所示，从一个可以控制的领域起点开始，逐步扩大规模到一个可描述的知识图谱。

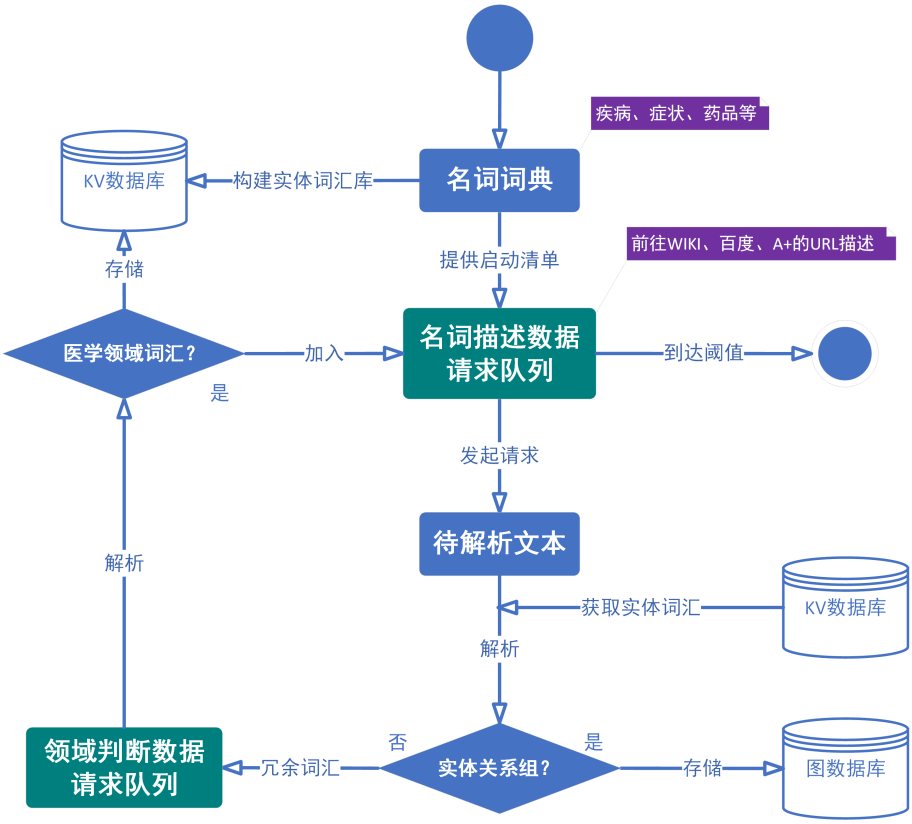

图 2.2　迭代检索过程

以心血管疾病相关的词汇表为起点，从词汇表所描述的实体开始逐层迭代，获取知识描述，即在确保图谱词汇领域针对性的同时确保相关领域词汇拓扑结构准确完整。获得知识描述信息后，交给实体抽取模块抽取实体，抽取出来的领域实体词汇放入词典并标记为待查状态，并对其与描述文本之间模糊的 MENTION 关系进行存储。

迭代知识获取的详细流程如下：

1. 起始词典建立

面对海量的网络数据，确定研究的起点和数据迭代获取的深度会有效的提高数据获取的效率，帮助我们快速建立领域相关的知识图谱。

2. 文本描述解析

在获取半结构化的文本描述信息后，需要抽取出其中与医学领域相关的词汇，从而确定下一步解析的方向并建立基本的实体关系。本研究将会抽取以下内容：

(1) 实体文本描述中所提及的医学词汇，将该医学词汇与实体之间建立 MENTION 关系并进行存储。

(2) 实体文本描述中与实体直接相关的“的”字语法关系，如冠心病的鉴别，将随后的医学文本描述与实体之间建立从属语法关系。

3. 多元关系存储

解析过程完成后，将解析得到的提及关系与子结构存储到词典中，提及词汇加入请求队列作下一步迭代检索的方向，从属关系直接进行存储，在后续研究描述的过程中进行实体关系组判别，并保存在知识图谱中。

2.4 实体抽取

实体抽取的最大障碍在于如何准确的找到目标词汇，如何保证词汇不被二次拆分。研究引入庞大的词表信息，以本地 17 万实体词汇的词表信息与网络标记共同协助分词，词表信息对目标词汇的查找作用显著。

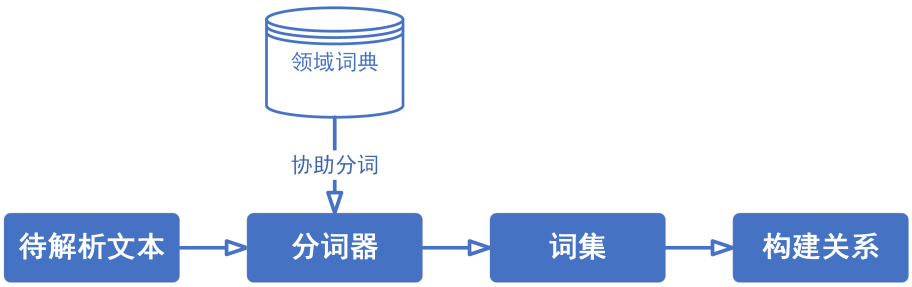

图 2.3 实体抽取过程

如图2.3 所示研究抽取实体的过程，使用领域词典来协助分词，分词结果属于领域词典时，放入词集，无关词汇不进行其他处理。获得词集后，建立带解析文本

与词集中词汇之间的 MENTION 关系。

2.5 关系抽取

由于人类语言抽象复杂的特性，目前高质量的实体关系的自动抽取依旧是自然语言处理领域一个亟待解决的课题。随着深度学习方法的发展，很多半监督式的算法模型在关系抽取的研究上取得了不错的成果。这些算法通过对大量语句文本进行分析，训练出可以分析句法依存关系、可以标注语义角色的模型。然而对于领域知识而言，因其对准确性的较高标准，通过深度学习抽取的关系暂时只能作为医学从业人员的参考。

由于对先验经验的依懒性，具体关系抽取将会作为未来研究的内容之一，当前研究采用模糊关系抽取的方案。当描述文字中存在医学实体时，建立该医学实体与该段落之间的提及关系，这种提及关系不考虑具体的语义，不考虑描述文字中对该医学实体的褒贬情绪。

在单独使用模糊关系进行推导时，这种模糊的关系可能导致不准确或者具有偏见的结果。但通过整体进行推导时，这种模糊关系也可以表现出十分优秀的结果，如在比较两个疾病相似性时，越相似的疾病需要考虑的因素和需要排查的疾病也会十分相似。

3 知识融合

知识融合相关技术起于对本体匹配[1]的研究，通过对本体属性和结构的定义，使用本体匹配的方式，实现概念消歧。随着语义网到知识图谱的发展，知识融合相关技术方案更加丰富，以结构化知识为基础的知识图谱不仅可以实现同构概念消歧，还可以完成多源异构知识的融合。

知识融合过程存在很多不确定性，这种不确定性存在于同类实体的知识的融合过程，同时也存在于完全异构的知识结构融合过程。知识融合不仅需要将不同来源的数据整合进一个整体构架中，同时需要对不确定性问题进行处理，消除带有歧义的语义关系，从而对外提供统一的表现形式。

以知识图谱为基础的知识融合技术灵活性大大增强，同构知识的融合使用结构化关系属性存储，异构知识的融合使用结构化关系链接。为保证融合质量，知识融合过程完成后知识图谱满足以下基本要求：

- 知识融合后保持领域针对性
- 知识融合后便于使用
- 知识融合后便于组织和管理
- 知识融合后便于理解与实现

基于知识图谱的知识融合过程分为四部分，第一部分是将结构化存储的数据映射到图结构上，实现知识图谱的知识存储基础；第二部分是同构知识融合，合理抽取并链接知识之间的关系，实现知识对齐和链接两个目标；第三部分是异构知识融合，设计合理的融合结构，对两类异构知识进行融合，融合后不破坏原有知识结构并实现异构知识之间的推导；第四部分是对融合的质量控制，要求知识图谱能表达出知识的语义概念、能识别表达同一概念的知识、能跨结构表达知识、能表达出知识的时空有效性，当融合多媒体知识时还需要跨媒体表达知识。

[1]本体匹配：本体匹配指两概念间相似度计算过程，并根据相似度值判断概念间的语义关系，实现同概念融合或相似概念的关联。

3.1 知识结构选取

实体结构是数据组织融合的基础逻辑组成，本研究提出两种基本的目标逻辑结构。选择目标实体结构时需要考虑不同结构对存取速率、逻辑关系、数据扩展等多方面的影响。

3.1.1 细粒度结构

目标结构实体划分具体明确时，领域性更强。如图3.1 医学领域实体有疾病，症状，药品三部分，并各自建立相应的分类结构。包含疾病实体与疾病分类，药品实体与药品分类以及症状实体与症状分类。这种结构特征明显，分类清晰，领域性特征强。

这种较细粒度的结构相当于关系型数据库存取过程中的水平分表，便于数据库底层建立索引，加速存取速度。

但是对于医学领域而言，还有急救方法、生物医学、生理健康等众多领域的医学实体，相同概念实体不易对齐，扩展方法更为复杂。

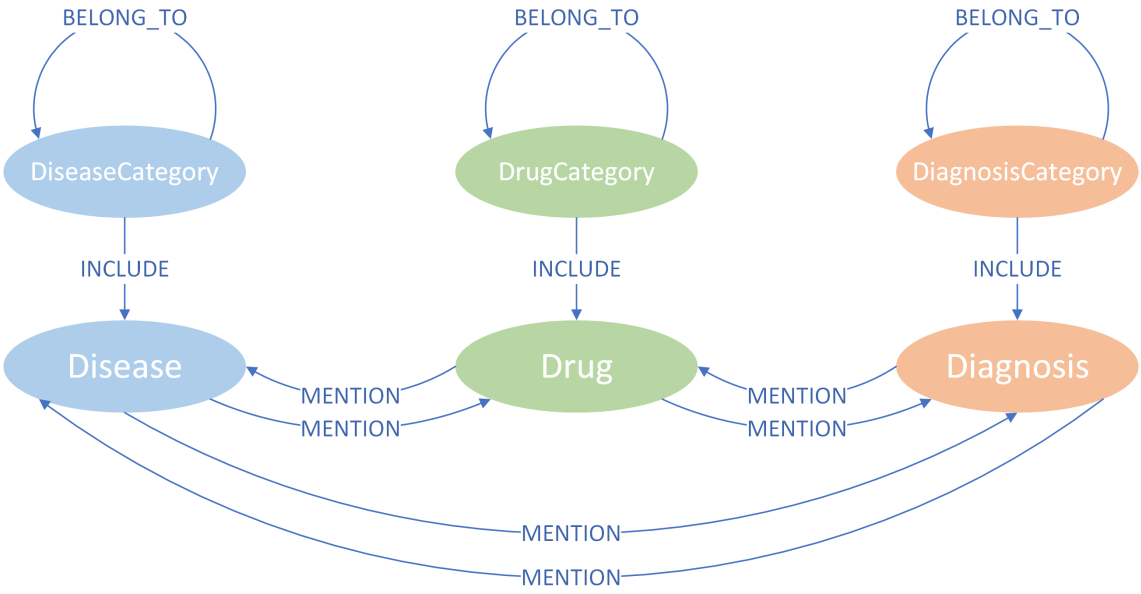

图 3.1　细粒度结构

3.1.2 高抽象结构

目标结构实体具有较高抽象程度时，领域性变弱，而实体对齐、数据扩展、引入更多专家数据更为便利，是多源异构医学知识融合的不二选择。

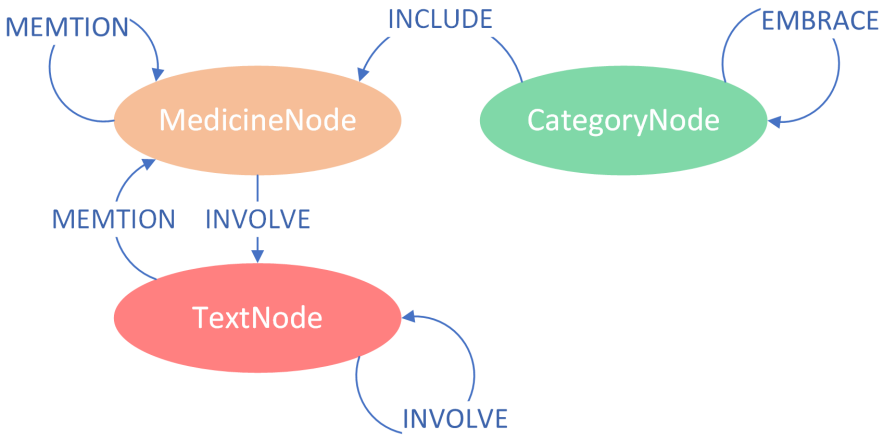

图 3.2　高抽象结构

如图3.2 所示，MedicineNode 之间为 MENTION 关系，根据初始文献或网络数据挖掘医学实体之间的 MENTION 关系，未来将根据专家数据，准确的医疗诊断方法提供更加紧密，精准的数据关系。

TextNode 为医学实体提及到的医学子结构，包括医学实体的详细结构信息，比如对于分类属于疾病的医学节点，INVOLVE 的 TextNode 包括疾病的成因，诊断，并发症等。

分类实体结构之间为 EMBRACE 精准关系，根据 ICD-10 国际标准，提取医学领域分类结构关系。CategoryNode 与 MedicineNode 之间同样为 EMBRACE 精准关系。CategoryNode 实体决定 MedicineNode 医学实体所属类别，是使得 MedicineNode 实体结构抽象化的关键。

3.1.3 结构选取

对于实体划分明确的目标结构，将会更加容易提取类别之间的逻辑结构关系，也会方便数据库底层建立索引；对于高抽象程度的目标结构，将会更方便数据的扩展。

表3.1 是在 2.5GHz 的 CPU，8G 内存环境下两种结构读取速率之间的对比。划分明确的逻辑结构由于方便构建索引，在直接查询[2]和按分类查询[3]上均优于高抽象程度的目标结构，但由于实际查询过程中不会超过 10 种不同的关系结构，最终产生的时间差异不会大于 1 秒。

表 3.1　查询速度对比

	数量 (个)	划分明确的逻辑结构		高度抽象的逻辑结构	
		直接查询 (ms)	按分类查询 (ms)	直接查询 (ms)	按分类查询 (ms)
疾病	14422	140	132	174	150
药品	26310	146	135	171	150
症状	12891	137	132	169	149
其他	173781	171		171	

为了多源异构的数据有效融合，本研究将采用高抽象程度的目标结构，同时以 MedicineNode(医学实体节点) 为主要数据实体，构建医学本体结构，充分保证目标结构的领域针对性。

3.2　知识映射

医学知识从获取，到关系解析，再到实体映射到标准结构并进行图数据存储后才算是完成了一次完整的知识映射。医学实体之间的相关关系错综复杂，在对存取速率要求不严苛的情况下，本研究选择高抽象程度的目标结构。

由于医学知识本身的多源异构性，需要对不同的数据源设计不同的映射结构，确保对已有知识利用率的最大化。下文将描述几种典型的实体映射过程，图3.3 描述了冠心病的知识映射结果。

领域概念映射为典型的哈希结构映射，即通过概念名可以唯一的确定医学实

[2]直接查询：通过节点名字或者 ID 直接检索实体

[3]按分类查询：通过节点所属的分类与节点之间的关系间接查询节点

体。在对描述文本分词解析完毕后，由哈希映射转变为一对多的关系映射。保存在图数据中之后转变为多对多关系映射。

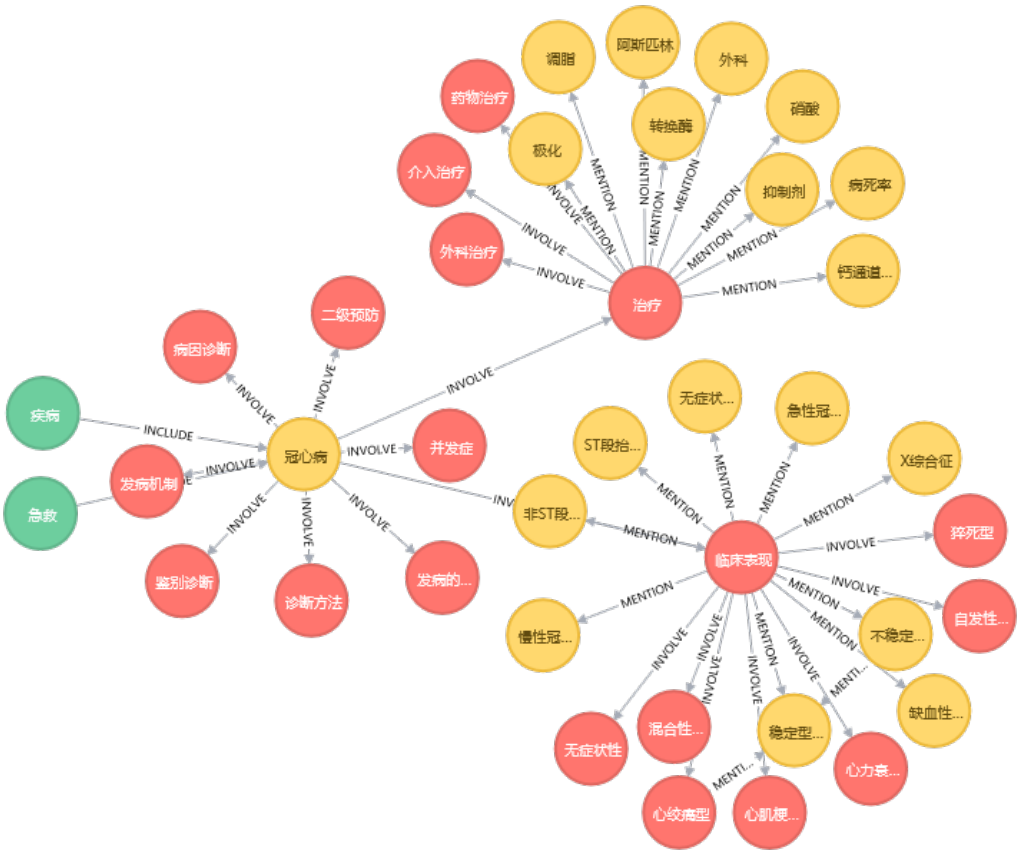

图 3.3　冠心病映射结果

病历映射为典型的树形结构映射，一个病历中包含 N 个结构化的子树结构，根节点保存病历信息，子结构保存病历涉及的检查与结论。在存取过程中，始终需要保证根节点的独立性，查询数据时，由根节点遍历到各子节点上。

文献映射为改进的树结构，子树结构依次相连，记录文献中上下文的相关性。在推导过程中，既可以推导出文献前面的步骤章节，也可以推导出文献后面的结论章节。

专家经验本就为图结构映射，但是这种图结构并非传统静态知识图谱，而是有向的过程性的事理图谱，其中记录了专家处理问题时的流程化信息。

对于其他种类的医学数据，如影像数据、时序数据，可以转化为影像特征-症状（疾病）、时序特征-症状（疾病）的哈希映射结构，融合过程与疾病类似。在本研究提及的映射以及融合的过程中不进行过多的讨论。

3.2.1 领域概念映射

1. 标准结构

如图3.4 左侧所示，左图为疾病在图数据库中高抽象程度的存储标准结构，其中 CategoryNode 是用于划分分类的实体节点，MedicineNode 表示核心实体概念，TextNode 为 MedicineNode 所包含的基本语义关系。

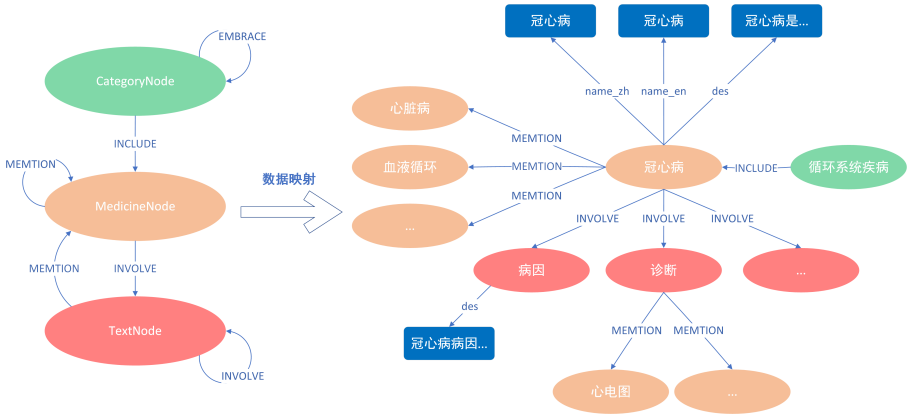

图 3.4　疾病映射

2. 映射逻辑

如图3.4 右侧所示，右图为数据在标准结构上的映射结果，右图以冠心病为例展示了疾病数据映射的基本元素对应关系。其中名为冠心病的实体映射到中心 MedicineNode 节点，循环系统疾病映射到 CategoryNode 节点，图3.3 展示了冠心病实体在图数据库中实际的存储结构。

3.2.2 病历映射

1. 标准结构

如图3.5 左侧所示，左图为病历在图数据库中高抽象程度的存储结构标准，所有树结构的信息均可以该结构进行存储。

图中 MedicalRecord 节点用于存储病历的根节点，包含病历的基本信息，SubNode 节点是与病历相关的历史以及各项检查，包含不同检查的病历即包含不同的 SubNode，SubNode 的具体信息以属性的方式存储。

2. 映射逻辑

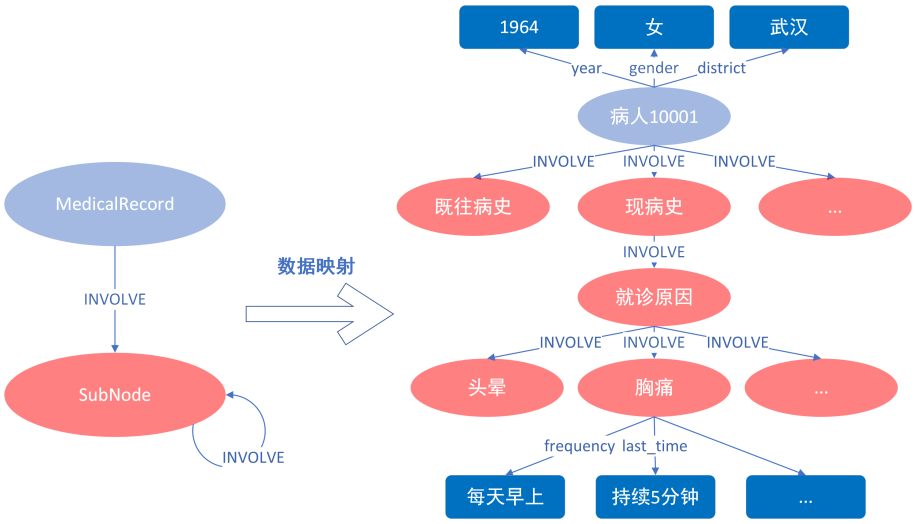

图 3.5　病历映射逻辑

如图3.5 右侧所示，右图为病历数据在标准结构上的映射结果。右图以 id 为 10001 的病人为例展示了病历数据映射的基本元素对应关系，将脱敏后的病人信息看做一种知识，根节点保存病人的基本信息，SubNode 用一种迭代的关系处理复杂的 KV 结构数据。

3.2.3　文献映射

1. 标准结构

如图3.6 左侧所示，左图为文献在图数据库中高抽象程度的存储结构标准，需要保留上线文关系的树结构数据均可以使用该结构进行存储。

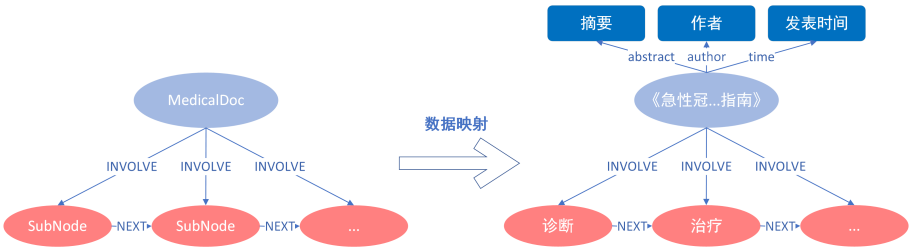

图 3.6　文献映射逻辑

图中 MedicalDoc 节点用于存储文献类资料的根节点，包含了文献的基本信息。在此结构中，SubNode 节点是保留上下文关联性的，记录了文献记录的基本逻辑流程，可以辅助医生快速推导到其需要的文本信息。

2. 映射逻辑

如图3.6 右侧所示，右图以《非 ST 段抬高型急性冠状动脉综合征诊断和治疗指南》为例展示了文献资料映射时基本元素的映射关系。其中根节点保留了改文献的摘要、作者等信息，使用 SubNode 结构化存储文献资料的子章节以及子章节之间的推导关系。

3.3 同构知识融合

来自不同数据源结构相似的知识融合的前提是两组数据执行统一的融合标准，下文以领域概念为示例，探究来自不同网站的领域概念数据的融合方法。

3.3.1 规范制定

研究针对数据集建立了两个规范，标准映射规范和相似度提取规范。标准映射规范要求异构的多源数据集以同构的方式存储，相似度提取规范要求多源数据集保持统一的特征抽取原则。以下为详细处理流程：

1. 确定两组数据都能接受的融合标准，对不同结构的来源进行标准映射 (MAPPING) 工作, 表2.2 展示了本研究对疾病医学实体的映射标准。多源数据集按来源建立多个数据库，每个数据库的表结构相同，无映射结果时数据可以为空。
2. 多源数据映射到数据标准后，确立统一标准的相似度识别规范，把用于比较相似度的数据进行存储，用于抽取从属关系和等价关系，最终实现实体分类和实体消歧。本研究在分词后抽取描述文本中所有领域相关的词汇。即用于比较的词汇均为领域相关词，确保相关性的同时带有强领域针对性。

3.3.2 相似对比数据抽取

实体词汇相似度比较的基本条件是得到比较所需要的特征向量（特征词汇）。对于医学领域而言，由于人体各器官本身具有高度的相关性，单纯通过词汇文本所描述的内容进行相似度对比不能完整的展现实体之间的相似度关系。本研究提出一种基于拓扑结构比较的领域实体词汇相似度提取办法，并在实验过程取得不错的结果。

1. 文本分词

分词前首先建立具有专业领域性的医学基础词典，本研究采用万方医学词典。基础词典越丰富，分词效率越高。

除掉中文停用词后，对实体词汇的描述信息进行尽可能细粒度的分词处理，分词结果用词典做粒度比较，确定具体词汇，不确定词汇加入检索队列。通过三方专业数据标签，确定属于医学领域的词汇即加入分词结果并更新领域词汇数据库，不属于医学领域的词汇不做处理。

2. 实体关系存储

分词后得到实体-实体描述的领域词汇数据集，实体与领域词汇间的确切关系暂不考虑，统一组织为提及 (MENTION) 关系，并将此提及关系按照提及方向存储在图数据中。

3. 拓扑结构查询

实体关系存储完成后，可以得到大量实体之间清晰的提及关系结构。由于图数据库本身对拓扑子结构检索的优异性，对实体词汇的多级拓扑关系查询快速可行。本研究将分别比较多级拓扑的相似性比较结果。

多级拓扑: 一级拓扑指与实体直接相关的实体所产生的拓扑子图结构，二级拓扑指与实体通路长度为 2 的拓扑子图结构，在关系型数据库中表现为一次 join 关系。

3.3.3 对比方法

在图谱中，以节点 v 为中心, 深度为 d 的子网络可以建模成数据图 $G(v,d) = \{V(G(v,d)), E(G(v,d))\}$, 其中 $V(G)$ 指在数据图 $G(v,d)$ 中的所有节点形成的点集，$E(G)$ 指在数据图 G 中的所有节点链路产生的边集，另定义节点 v_0 在数据图 $G(v,d)$ 中的深度为 $D(G(v,d),v_0)\quad v_0 \in V(G(v,d))$，从而得出对任意节点 v_0 的权重为:

$$\begin{cases} W(G(v,d),v_0) = d - D(G(v,d),v_0) + 1 & v_0 \in V(G(v,d)) \\ W(G(v,d),v_0) = 0 & v_0 \notin V(G(v,d)) \end{cases}$$

当考虑由两张子网络 $G(v_1,d)$ 和 $G(v_2,d)$ 共同构成的拓扑结构时，节点 v_0 在

其中的权重描述为：

$$W(G(v_1,d),G(v_2,d),v_0)=W(G(v_1,d),v_0)+W(G(v_2,d),v_0)$$

本文针对点集 $V(G(v,d))$，提出基于拓扑节点集的带权结构对比方法。

定义 3.3.1 基于拓扑节点集的覆盖率对比。给定数据图 $G_1(d)=G(v_1,d)$、$G_2=G(v_2,1)$, 节点集 $V_1(d)=V(G_1(d))$，节点集 $V_2=V(G_2)$，数据图 G_1、G_2 可能包含相同节点，也可能完全异构。$G_1(d)$ 对 G_2 覆盖率定义如下：

$$cov(G_1(d),G_2)=\frac{|V_1(d)\cap V_2|}{|V_2|}$$

定义 3.3.2 带权拓扑节点集的覆盖率对比。 计算过程与以上定义相似，不同的是加权节点集覆盖率计算过程中为集合中每一个节点赋予了一个权重，其中 $W(G_1(d),v_0)\in(0,d)$, 离中心节点越近权重越高。$G_1(d)$ 对 G_2 的带权覆盖率定义如下：

$$w\_cov(G_1(d),G_2)=\frac{\sum_{v_0\in(V_1(d)\cap V_2)}W(G_1(d),v_0)}{\sum_{v_0\in V_2}d}$$

w_cov 的取值范围为 (0,1)，越是靠近 1，$G_1(d)$ 对 G_2 的带权覆盖率越高。

定义 3.3.3 基于拓扑节点集的相似度对比。给定数据图 $G_1(d)=G(v_1,d)$、$G_2(d)=G(v_2,d)$, 节点集 $V_1(d)=V(G_1(d))$，节点集 $V_2(d)=V(G_2(d))$。$G_1(d)$ 对 $G_2(d)$ 相似度定义如下：

$$sim(G_1(d),G_2(d))=\frac{|V_1(d)\cap V_2(d)|}{|V_1(d)\cup V_2(d)|}$$

定义 3.3.4 带权拓扑节点集的相似度对比。 计算过程与以上定义相似，不同的是加权节点集相似度计算过程中为集合中每一个节点赋予了一个权重，其中 $W(G_1(d),G_2(d),v_0)\in(0,2d)$,$v_0$ 离两中心节点越近权重越高。$G_1(d)$ 对 G_2 的带权相似度定义如下：

$$w\_sim(G_1(d),G_2(d))=\frac{\sum_{v_0\in(V_1(d)\cap V_2(d))}W(G_1(d),G_2(d),v_0)}{\sum_{v_0\in(V_1(d)\cup V_2(d))}W(G_1(d),G_2(d),v_0)}$$

3.3.4 内容覆盖率对比

内容覆盖率对比要求获取两实体之间的领域包含程度，并对对实体从属提供数学指导。根据上文定义公式，得到不同 d 下的内容覆盖率情况。图3.7 描述了实体间的覆盖率沿深度的曲线。图3.8 描述了实体间的带权覆盖率沿深度的曲线。

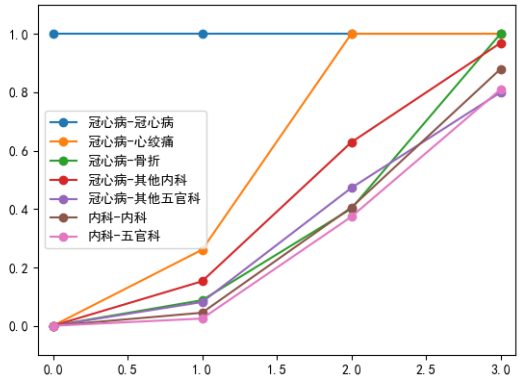

图 3.7　覆盖率对比

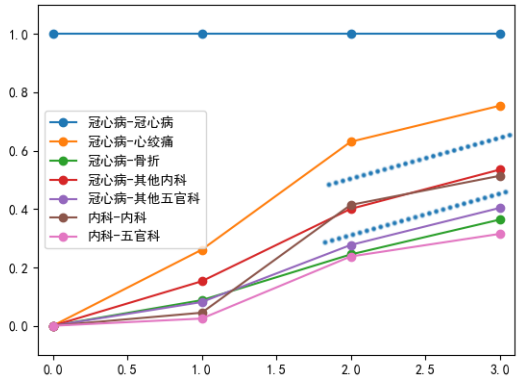

图 3.8　带权覆盖率对比

图中共有 7 条曲线，对曲线的解释如下：

- 蓝色曲线代表等价实体间的（带权）覆盖率
- 黄色曲线代表从属实体间的（带权）覆盖率
- 绿色曲线代表不同实体间的（带权）覆盖率
- 红色曲线代表特例对同类实体间的平均（带权）覆盖率
- 紫色曲线代表特例对异类实体间的平均（带权）覆盖率
- 褐色曲线代表同类实体间的平均（带权）覆盖率
- 粉色曲线代表异类实体间的平均（带权）覆盖率

从整体来看，内容覆盖率对具有从属关系的实体更加敏感，一级拓扑已经可以比较实体覆盖关系，这也是目前大多数文本比较的方法，但覆盖程度整体低于0.3，很难找出不同类别之间的差异性。

从二级拓扑开始覆盖率大幅度上升，在左图中由于整体上升趋势均较大，无法显著的划分出同类实体（褐色曲线）与异类实体（粉色曲线）覆盖率之间的区别，由此提出带权覆盖率对比。

由于二级以上相关性权重下降，而分母不变，使得在拓扑结构大范围靠近（同

类）的实体相比拓扑结构大范围疏远（异类）的实体在带权覆盖率对比过程中更具优势。因此可以明显的划分出同类实体间与异类实体间的曲线间隔，并可以以此为基准为实体划分从属关系。

在带权覆盖率对比过程中，三级拓扑与二级拓扑没有明显差异性，而三级拓扑规模远大于二级拓扑，故在本研究中选择二级拓扑为主要计算对象。

3.3.5 领域相似度对比

领域相似度对比要求获取两实体之间的领域相关性程度，并对对实体消歧提供数学指导。根据相似度获取实体节点 Node1 与 Node2 之间的相似关系，相似关系可以被诠释为基本等价，程度相似和完全异构三种实体组织关系。

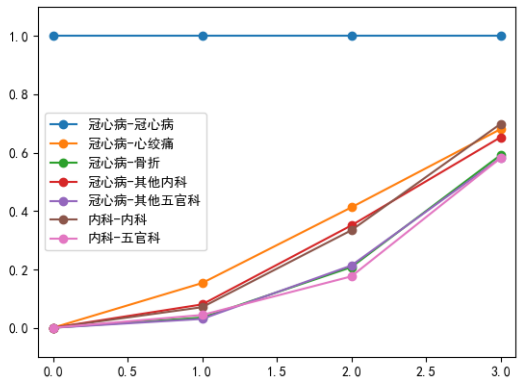

图 3.9 相似度对比

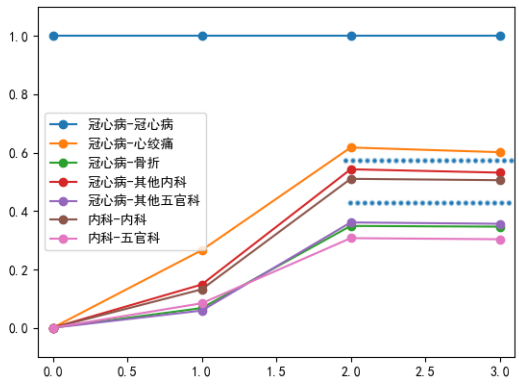

图 3.10 带权相似度对比

拓扑相似度对比从另一个角度对实体间的关系进行诠释，更注重诠释二者之间的相似性而非从属性，用于挖掘可能隐含的相似性。其中对 7 条曲线的解释与上文类似，在此不再赘述。

一级拓扑相当于直接抽取出来的文本特征。从整体来看，相似度对比对相似性更加敏感，一级拓扑无法很好地比较实体相似关系，即比较相似度需要借助更多的特征数据来完成。

从二级拓扑开始，左图中同类实体（褐色曲线）实体（粉色曲线）相似度已有明显界限。但由于没有权重限制，分子扩大规模小于分母扩大规模，从三级拓扑开始，左图中界限重新变得模糊，相似度对比变得不再直观。

右图中，分子分母扩大规模受到权重限制，随着深度的变化，带权相似度在达到极值后逐渐下降。使得在拓扑结构大范围靠近（同类）的实体相比拓扑结构大范围疏远（异类）的实体更容易达到较高的极值，之间的差异性可以通过一条近似直线快速划分。

在带权相似度对比过程中，大部分数据在二级拓扑达到极值，从而没有比较再对三级拓扑进行计算，以大规模减少计算量。

3.3.6 实体融合

同名不同义的医学领域词汇颇为少见，研究主要针对医学同义词与近义词。实体融合根据上文所述理论基础，同名实体直接映射到同一结构中，别名实体先考虑类别对实体的内容覆盖率，再对比与其他实体的领域相似度关系，确定进行同义消歧或者是进行相似关联。

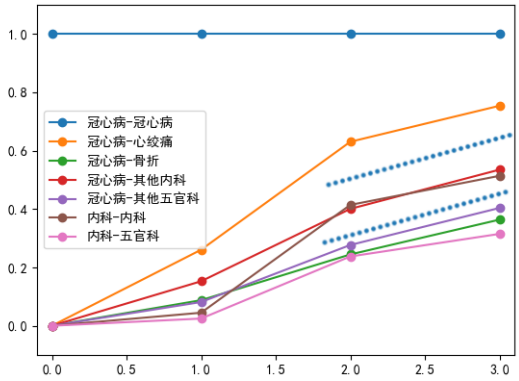

图 3.11 带权覆盖率对比

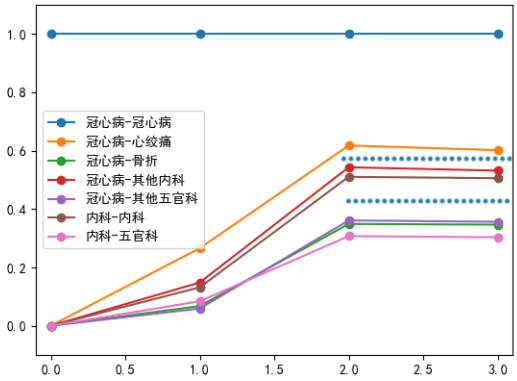

图 3.12 带权相似度对比

以上章节论证表明，带权覆盖率对实体间的从属关系更为敏感，而带权相似度对实体间的相似性关系更为敏感。如图3.11与图3.12 所示，综合两图的最大间隔信息，可以挖掘出实体间存在从属与相似关系的概率大小并进行存储。

人体各大系统相互关联，医学领域实体即使属于完全不同的类别，也可能会有很高的从属关系与相似性关系。这种关系是不完全确定的，可以在在推导过程中作为辅助，在存储过程中以概率的形式继承 MENTION 关系，将 MENTION 关系细化。

3.4 异构知识融合

来自不同数据源异构的知识融合的前提是以一种结构为基础，将不同种类的数据融合到该基础结构上。本研究以医学实体（MedicineNode）为基础结构建立起了庞大的知识图谱结构，其他异构的知识均会以此为基础结构进行融合，融合满足以下原则：

- 融合前数据均以结构化形式存储
- 融合过程不改变不破坏基础结构
- 融合后多类结构之间可以相互检索

3.4.1 规范制定

异构知识融合时，先要定义出多源数据在知识图谱中的基本结构单元以及结构之间的关联方法，实现多类知识融合的结构基础。本研究将讨论多类异构数据向医学实体数据融合的过程。

3.4.1.1 领域概念基础

如图3.13 所示，该结构为医学实体研究的基本结构。本研究所实现的医学实体均以该结构进行存储，医学实体之间以 MENTION 关系相互联系，形成一张如右图所示结构巨大的知识图谱3.14

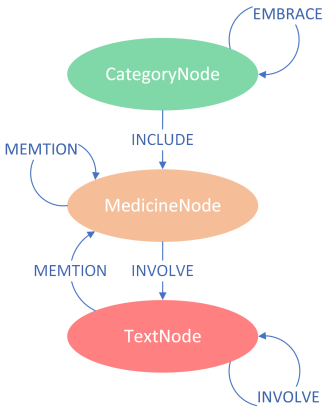

图 3.13　领域概念基础结构

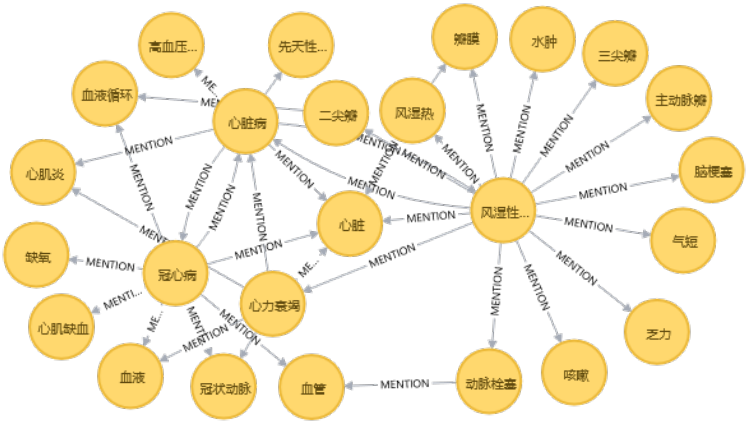

图 3.14　领域概念图谱示例

右图经过处理去掉其他类型节点和关系，只保留医学实体和医学实体之间的MENTION 关系。

3.4.1.2 病历融合规范

如图3.15 所示，左侧为医学实体基础结构，右侧为病历基础结构。病历所涉及的各项检查提及同名医学实体时，建立右侧结构到左侧结构的相关性映射，相关检查与提及节点之间以 MENTION 关系相连接。

MENTION 关系为单向的，不破坏原有结构的，且该结构并不影响在图谱中实现从医学实体向病历实体的推导。显而易见，也可以轻易的通过病历挖掘出病历映射在医学实体图谱上的拓扑结构。

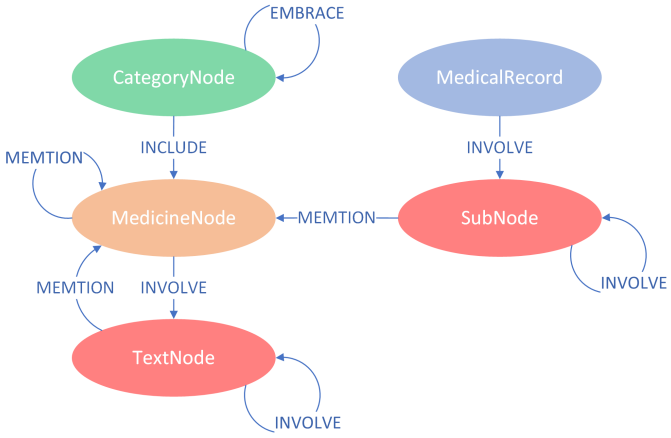

图 3.15 病历融合规范

3.4.1.3 文献融合规范

如图3.16 所示，左侧为医学实体基础结构，右侧为文献基础结构。文献的子章节中提及到同名医学实体是，建立右侧结构到左侧结构的相关性映射，使用提及关系 MENTION 存储。

3.4.2 实体融合

实体融合根据以上理论基础，将异构的知识图谱与医学实体图谱相融合。融合后可以通过医学实体直接推导病历实体与专家经验实体，也可以被反向推理。病

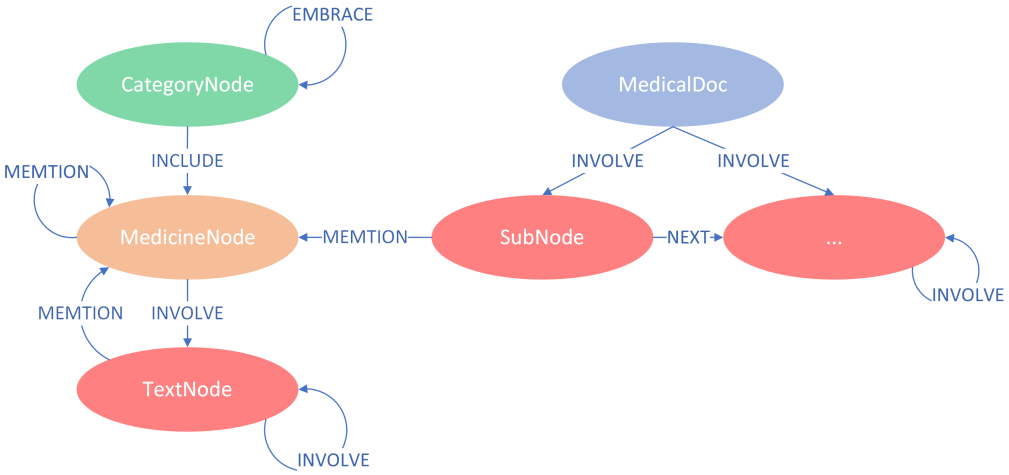

图 3.16　文献融合规范

历与专家经验之间通过医学实体间接推理。

3.4.2.1　病历知识融合

如图3.17 所示，病历结构通过 MENTION 关系融合到医学实体结构图谱上。即病历结构上的节点头晕、胸痛与医学实体图谱上的节点等价时，建立二者之间的 MENTION 关系。

这种关系不会破坏原来医学图谱中头晕和胸痛的关系组成，而通过反推 MENTION 关系可以通过头晕和胸痛的症状，找到与之相关的病历。

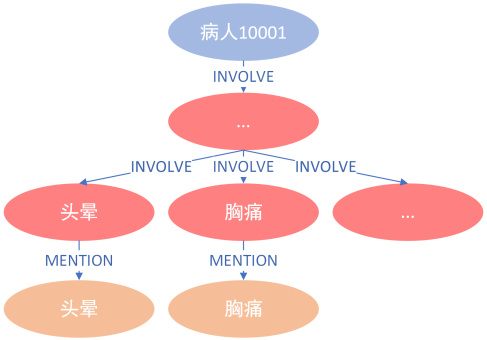

图 3.17　病历融合示例

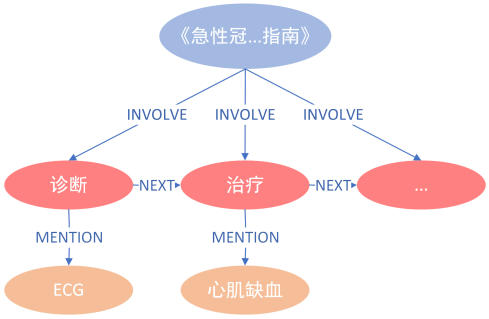

图 3.18　文献融合示例

3.4.2.2 文献知识融合

如图3.18所示，文献知识通过 MENTION 关系融合到医学实体图谱上。即文献知识中提及医学领域词汇时，建立二者之间的 MENTION 关系。结构融合后的检索方案与病历类似，故不在此赘述。

3.5 知识表示与评估

知识融合后，需要以符号化、形式化、模块化的约定表示知识，利用知识，评估融合知识的合理性。不同的表示方式可能直接影响知识转化运用的效率。医学领域知识种类丰富，存储结构存在差异，部分数据涉及生物等交叉领域，为医学知识表示带来巨大挑战。

目前知识的表示方法包括状态空间表示法、谓词逻辑表示法、框架表示法等等。其中医学领域包括 SNOMED-CT、EcoCyc 等数据库均以此类方法进行知识表示。随着语义网到知识图谱的发展，可处理的知识规模日渐增大，以知识图谱为主的知识表示方法成为主流。以知识图谱为基础的知识表示方式以三元组表示法为主，近年来有许多研究通过机器学习三元组结构的方式训练更优的知识表示方式。

目前针对知识图谱的知识表示研究包括三元组表示、时空语义表示和跨媒体表示知识等方面。三元组表示要求通过关系准确的描述概念与概念的关系；时空语义表示要求知识在当前时间和空间状态环境下是有效的；跨媒体表示要求对多媒体的数据源建立关系链接，满足更丰富的检索或科研需求。研究将针对三元组表示与时空语义表示进行研究。

3.5.1 三元组表示

知识图谱通过三元组 SPO(Subject, Predicate, Object) 的形式来描述两个实体之间的关联，对医学术语的结构和概念进行深入的建模，不仅可以表达出晦涩的医学概念，也可以轻松的表示跨语言的知识。

1. 关系表示

关系表示是知识图谱表示的基础，一切三元组均以实体-关系-实体的形式进行表示，可以十分便捷的表示实体的固有特征以及与其他实体之间的关联。

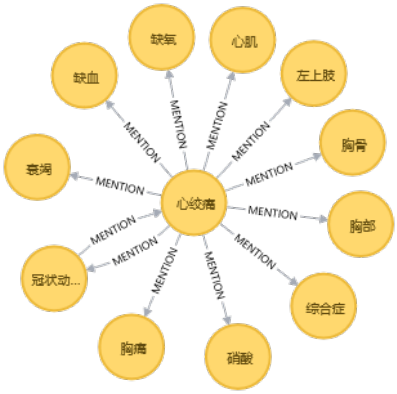

图 3.19　概念表示

2. 概念表示

概念表示如图3.19 所示，是以领域实体为中心，提及关系为中介的实体集合。在表达一个概念时，不仅要表达概念本身，还需要表达与概念相关的实体集合。

3.5.2　时空语义表示

时空语义表示将三元组 SPO(Subject, Predicate, Object) 的形式扩展为四元组 SPOT(Subject, Predicate, Object, Time) 的形式，这样不仅包含了知识的结构信息，也包含了知识的产生时间。当知识需要迭代时，无需破坏原有的知识结构，其表达形式更加灵活。

3.5.3　知识评估

知识图谱融合过程完成后，需要对图谱做基本的评估，确定图谱的领域针对性，考察当前图谱可以完成哪些功能，研究对图谱的评估包括以下几个方面：

- 图谱的知识规模与质量
- 图谱的知识时效性

在图谱规模方面，图谱融合了 17 万医学实体节点，210 万实体关系。以心血管疾病为主体，覆盖日常生活中涉及的医学知识。数据融合时未破坏原始数据结构，知识均完整准确的保留了下来，基本的医学知识检索可以保证。表3.2 是在 2.5GHz 的 CPU，8G 内存环境下对查询速度的评估。

表 3.2　查询速度

时间 (ms)	数量 (个)	查询	推理查询			拓扑查询		
			一级	二级	三级	一级	二级	三级
领域实体	173781	171	172	179	184	241	1527	9770
领域关系	2132977	209						

在图谱知识时效性方面，当前研究均采用最新的领域知识。但随着知识的动态变化，尤其是临床文献类知识的快速演化，对时效性要求高的知识不再以三元组（S, P ,O）表示，而是以带时间参数的四元组（S, P ,O ,T）的形式提供给医学从业人员，后续会为其开发知识迭代系统减少人工参与更新的过程。

4 知识推理

逻辑学指出，推理是通过前提推导结论的过程。与之类似，基于图谱的知识推理是从已有的知识结构中，由一个或者多个实体出发，通过遍历满足规则的实体关系，推导出新知识的过程。

传统推理在从已有知识中探寻隐含信息时，条件与结论之间的详细关系与推导方法相关，且推导方法需要进行额外的组织和管理。

基于图谱的知识推理重视如何获取和表达知识之间的关系，并以此为基础进行推导，从而大大减少了推导过程中选择和运用方法的过程。如三元组（心电图-鉴别-冠心病），只需要从心电图向冠心病推导，其中的推导过程显而易见，无需选择运用其他方法进行组织和管理。

知识推理的研究主要包括对知识的检索、解释、推荐、问答与决策。知识检索不同于传统信息检索，检索结果不仅要包含知识本身，还需要包含与知识相关的语义信息；知识解释是在知识检索过程完成后，如何将语义信息合理的表达给检索者；知识推荐则要求检索过程中能通过知识的拓扑关系推荐相似的知识；知识问答与知识解释相辅相成，区别在于知识问答需要分析提问者的语义再进行解释；知识决策指在问题的拓扑知识结构下为检索者提供明确的解决方案。当前研究规模暂不足以支撑决策问题，将就其他四个问题展开研究。

4.1 知识检索

知识检索是知识图谱最核心也是最基础的功能，研究考虑的检索分为两部分，一部分是对领域概念及其确定关系的直接检索，这种检索要求高度的语义准确性，另一部分是对领域概念间潜在关系检索，通过最短路等算法，探究多个领域概念之间潜在的关系。

1. 直接检索

直接检索不仅可以检索到概念本身，也可以检索与概念直接相关的关系。如图4.1 所示，通过三元组的形式，表示出与概念相关联的各个方面。

图 4.1　概念直接检索

2. 间接检索

在多数情况下，由于实体（概念）之间语义关系的模糊性，部分事实无法直接通过单个条件推理出结论或以决策树形式按序推理得到结论。但是通过分析多个实体（概念）之间的的整体关系，或者通过分析深层的实体关系，可以得到基于事实的潜在的或者概率性的结论。

图 4.2　酒与冠心病的潜在关系

如图4.2 所示酒与冠心病的潜在关系，通过最短路查询过程，找到 15 条酒

与冠心病，路径长为 2 的相关路径，对路径推导过程的解释，分别从不同层面解释了酒与冠心病的关系，可以作为医学人员研究过程的辅助参考。

4.2 知识推荐

知识推荐要求以知识检索时的状态为基础，推理出尽可能相似的知识。知识推荐作为知识检索的一种扩展，可以有效的辅助查询，减少查询工作量，快速得到相关知识的参考。

1. 同构知识推荐

对于同构知识而言，推荐相似知识可以直接使用带权拓扑相似度对比方法，对比过程将会十分有效，每次对比后需要将结果保存在 MENTION 关系内，以加速下次推荐的过程。

2. 异构知识推荐

对于异构的知识而言，推荐相似的知识可以用提及到的词汇进行相互之间的推导，上文研究中已阐明，通过单个或者一级拓扑进行比对，并不能很好的得到相似度推荐结果。

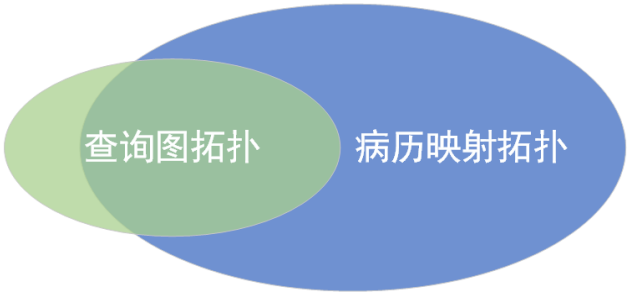

图 4.3　拓扑覆盖率比对

如图4.3 所示病历知识拓扑与查询拓扑之间的关系，对异构知识进行推荐时，使用无中心点要求的带权覆盖率对比方案，通过拓扑覆盖率对比的方式，获取更为相似的病历数据。

4.3 知识决策

基于知识图谱的知识决策是指通过决策条件向量与与决策关系向量，尽可能大概率的推导到最优解上。如图4.4 所示研究中确定解的决策过程，通过对冠心病成因关系的推导，决策得出成因描述文本。

图 4.4　知识决策

在实践环境中，许多决策过程无法像图示这样准确清晰，决策条件，决策关系均无法用单个实体或关系来表示。如何尽可能抵达目标结果，与决策条件的表示、决策算法的实现都有很大关系，对于知识决策相关研究而言十分重要。

4.4 知识问答

知识问答实际上为知识决策过程的进一步演化，目的是更好的服务人类，帮助人们快速获取需要的知识。总体而言知识问答包含两部分。首先是基于自然语言处理的决策条件与决策关系映射，这其中需要计算机理解人类语言中所蕴含的可以映射为决策条件和决策关系的成分。然后以决策条件与决策关系为基础，通过图4.4所示过程，尽可能推导出最优解。研究所实现问答系统在附录中有所展示。

结　论

本课题最终构建了一个心血管疾病知识图谱，基于图谱实现了初步的知识决策系统，并以此建立语音问答系统，可根据问题查询相关医学知识。研究成功解决了多源异构医学知识融合这一基本问题，具体实现了以下内容:

- 实体抽取与知识映射
- 同/异构知识融合
- 实体对齐与链接
- 初步的知识推理

研究发现知识图谱是解决知识结构化问题十分优秀的方案。在知识融合阶段，研究提出了合理的融合方法，且发现基于拓扑的相似度比较远优于基于文本的对比。

研究仍然有很多遗留内容，这些遗留内容对知识图谱的进一步构建有重要意义。展望未来，研究还需包括：

- 语义关系抽取
- 专家经验自动化
- 复杂知识推理

课题始终以前沿优秀的理论作指导，以技术形成应用为目标方向。随着研究所构建的知识图谱不断成熟，提供更多、更实际的落地解决方案。

参考文献

[1] Laskey K B, Costa P C G, Janssen T. Probabilistic ontologies for knowledge fusion[C] International Conference on Information Fusion. IEEE, 2008:1-8.

[2] Yang M C, Duan N, Zhou M, et al. Joint Relational Embeddings for Knowledge-based Question Answering[C] Conference on Empirical Methods in Natural Language Processing. 2014.

[3] Wang Q, Mao Z, Wang B, et al. Knowledge Graph Embedding: A Survey of Approaches and Applications[J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(12):2724-2743.

[4] Yih W T, Chang M W, He X, et al. Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base[C] Meeting of the Association for Computational Linguistics and the, International Joint Conference on Natural Language Processing. 2015:1321-1331.

[5] Berant J, Chou A, Frostig R, et al. Semantic parsing on freebase from question-answer pairs[J]. Proceedings of Emnlp, 2013.

[6] Gawriljuk G, Harth A, Knoblock C A, et al. A Scalable Approach to Incrementally Building Knowledge Graphs[J]. 2016.

[7] Szekely P, Knoblock C A, Slepicka J, et al. Building and Using a Knowledge Graph to Combat Human Trafficking[C] International Semantic Web Conference. Springer, Cham, 2015:205-221.

[8] Yan X, Mou L, Li G, et al. Classifying Relations via Long Short Term Memory Networks along Shortest Dependency Path[J]. Computer Science, 2015, 42(1):56-61.

[9] Zeng D, Liu K, Chen Y, et al. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks[C] Conference on Empirical Methods in Nat-

ural Language Processing. 2015:1753-1762.

[10] Dong X L, Gabrilovich E, Heitz G, et al. From data fusion to knowledge fusion[J]. Proceedings of the Vldb Endowment, 2014, 7(10):881-892.

[11] White H D. Pathfinder networks and author cocitation analysis: A remapping of paradigmatic information scientists[J]. Journal of the American Society for Information Science & Technology, 2003, 54(5):423–434.

[12] Xu K, Reddy S, Feng Y, et al. Question Answering on Freebase via Relation Extraction and Textual Evidence[J]. 2016.

[13] 何南洋. 图书情报学知识图谱的构建及解读 [D]. 上海交通大学, 2011.

[14] 韩芳, 乔亚京, 石艳芬. 231 例中老年心血管疾病患者动态心电图临床分析 [J]. 中国中医药科技, 2014, 02(1):269-270.

[15] 李明. 智能医疗在中国发展现状、问题和对策分析 [D]. 中国科学技术大学, 2011.

[16] 刘峤, 李杨, 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.

[17] 高俊峰. 基于形式概念分析的开放存取资源组织方法研究 [D]. 吉林大学, 2011.

[18] 荆宁宁, 程俊瑜. 数据、信息、知识与智慧 [J]. 情报科学, 2005, 23(12):1786-1790.

[19] 徐赐军, 李爱平, 刘雪梅. 基于本体的知识融合框架 [J]. 计算机辅助设计与图形学学报, 2010, 22(7):1230-1236.

[20] 唐晓波, 魏巍. 知识融合: 大数据时代知识服务的增长点 [J]. 图书馆学研究, 2015(5):9-14.

[21] 梁秀娟. 科学知识图谱研究综述 [J]. 图书馆杂志, 2009(6):58-62.

[22] 中华医学会心血管病学分会. 非 ST 段抬高型急性冠状动脉综合征诊断和治疗指南 (2016)[J]. 中华心血管病杂志, 2017, 45(5).

致　　谢

图谱构建由 0 近 1，本科生活却渐趋于 0。在学位论文将要完成之际，我向所有在此期间给予我支持、帮助与鼓励的人表示诚挚的谢意。

首先要感谢我的导师，李兵教授。李老师对我论文的研究方向做出了指导性的意见和推荐，在论文撰写过程中及时对我遇到的困难和疑惑给予悉心指点，提供可用资源，指导我完成论文。其次要感谢何扬帆老师，何老师在实验过程中与我们交换意见，指出很多问题并为我们提供解决方案。

感谢实验室的师兄师姐们。感谢胡方家师姐不辞辛劳的带着我们去同济医院学习经验，帮助我修改论文，解决课题执行过程中方方面面的问题；感谢陈健师兄帮助解决实验环境问题，平日里一同交流技术经验，一同晨跑与吐槽；感谢赵玉琦师兄对课题的不断关心与支持。

感谢我的家人们。感谢爸爸妈妈对我学习过程无条件的支持，让我毫无后顾之忧；感谢家人们对我的夸奖与鼓励，以及对课题成果的功能性建议。你们是我力量的源泉。

最后感谢我身边的伙伴儿们。感谢三位室友为知识图谱的茁壮成长提供的恶劣环境，你们的吐槽让我受益匪浅，四年感情不多言，如果再来四年一定还是你们；感谢三位研友疯狂宣传只有小熊才能写出来的 O(0) 算法（算是花式鼓励吧），以及基本上随叫随到的支持与陪伴；感谢实验室的小伙伴儿们提供毕设资料，将来学习与生活的小船说翻一起翻。

毕业在即，愿吾与诸君过上向往的生活。

附录 A　相关成果

A.1　知识图谱

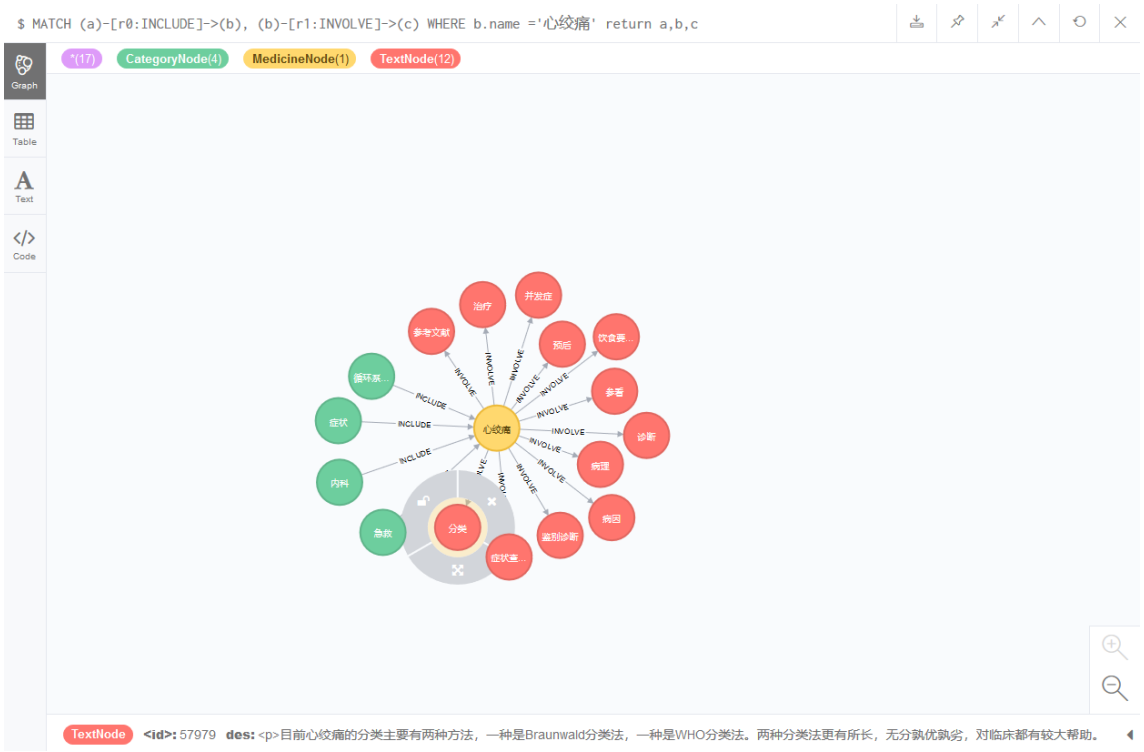

图 A.1　知识图谱

A.2　病历系统

图 A.2　病历系统

A.3　问答系统

图 A.3　问答系统