

大健康与新医疗

BIG DATA Health
and New Medical

2022 年

第 09 期



上海科学技术情报研究所
上海市前沿技术发展研究中心
技术与创新支持中心 (TISC)

知识图谱与健康医疗

编者按

知识图谱在图书情报界可称为知识域可视化或知识领域映射地图，是用可视化技术对知识资源及载体进行描述，并能挖掘、构建、绘制、显示知识语义及其相互关系的图谱。知识图谱属于智能大数据研究领域的前沿性研究，它以独有的技术优势顺应了信息化时代的发展。

同时，伴随智能时代的到来，信息化正在以迅雷不及掩耳的速度开拓医疗健康的领域疆土，将临床数据、临床指南、组学数据、以及核心医学概念、医疗生态圈内全方位知识数据进行全面聚合，最终以知识图谱直观呈现，是健康医疗领域正在努力发展的方向。

因此，如何构建一个简单实用、可扩展的医疗知识图谱，从而推动更深入的AI 医疗在工业界快速落地是一个循序渐进的过程。

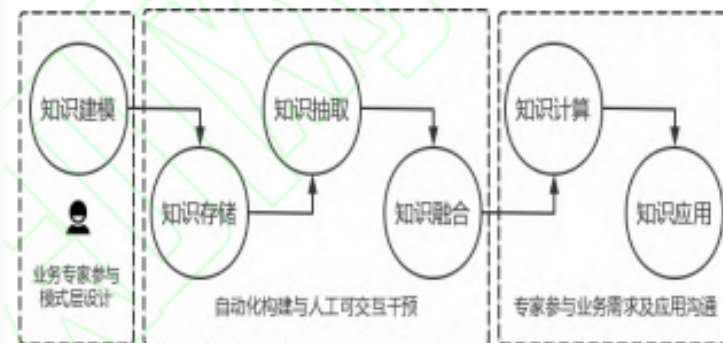
目 录

研究进展.....	3
医学知识图谱的构建	3
医学用药推荐系统研究.....	6
技术应用.....	8
医学知识图谱的应用场景.....	8
医学自然语言处理数据集.....	9
新冠图谱与病历图谱构建.....	12
相关产品.....	15
工业界医疗知识图谱产品.....	15
“汇知” 医学知识图谱简介.....	19

研究进展

医学知识图谱的构建

通用知识图谱知识覆盖范围广且数据量大，通常采用自底向上的方式构建，自动化程度较高。医学知识图谱属于领域知识图谱，构建的关键技术与通用知识图谱存在共性，但构建流程则有所区别。领域知识图谱构建的流程见下图所示，其中模式层对后续领域知识的获取和组织有着重要的指导意义。



领域知识图谱构建流程图

由于医学领域具有相对完备的知识体系，而且医学知识具有术语多样、结构复杂、专业性強且应用场景容错率低等特点，因此，由医学专家参与构建的医学本体对医学知识图谱模式层的构建极具参考价值，也对医学知识图谱的快速发展起到了很大的促进作用。文章将从医学本体构建、全科医学知识图谱构建和单病种医学知识图谱构建三个角度对中文医学知识图谱构建的相关工作进行总结。

一、医学本体构建

本体这一概念最早来源于哲学领域，后在知识工程领域作为知识组织的一种形式被广泛使用。本体指利用基本术语表达领域知识，确定领域内共同认可的概念和概念间的关系，以用于领域内不同主体之间的交流与知识共享的形式化规范说明。目前本体采用国际通用的形式化语言 OWL 来规范描述领域的概念及其语义关系，使得这些知识可被人机共同理解，从而解决了人机之间、机器之间信息传递和交流的障碍。本体作为一种重要的知识组织方法，为医学领域的知识图谱构建工作提供了重要基础，生物医学领域也一直处于本体研究的前列。国外成熟的医学本体包括 SNOMED-CT、基因本体(Gene Ontology, GO)、疾病本体(Disease Ontology, DO)和人类表型本体(Human Phenotype Ontology, HPO)等。近年来，仍

有不少学者在这些本体的基础上对医学领域本体、疾病本体和中医药本体等更细粒度的本体构建方面进行着深入研究。一些学者通过参考或复用国际权威的医学术语构建本体，如 Shepherd 等基于 SNOMED-CT 构建了一个本体并将其作为边界对象，以解决照顾慢性病患者多学科卫生保健小组成员之间的语义互操作鸿沟。国内也有学者对医学本体构建进行了积极的探索。牟冬梅等基于 SNOMED-CT 和形式概念分析构建了甲状腺疾病本体，并利用该本体对电子病历进行标注，验证了其有用性。李晓瑛等复用了 UMLS 和 SNOMED-CT 中的语义关系，并结合从文献中获取的疾病与药物之间的治疗关系，构建了呼吸系统肿瘤本体。任慧玲等构建了中医疾病本体，并完成了与 ICD-11 中文版的语义映射，为中医疾病分类统计的规范化和标准化奠定了基础。由于国内还尚未形成权威的医学术语标准，目前学者大多参考 UMLS、SNOMED-CT 以及 MeSH 词表等国际权威术语构建中文医学本体。近年来，国内也有机构致力于中文医学术语标准化的研究并取得了一定成效，如开放医疗与健康联盟(Open Medical and Healthcare Alliance, OMAHA)于 2019 年在 HiTA 知识服务平台上发布了“七巧板”医学术语集；中国医学科学院医学信息研究所于 2020 年在 BioPortal 平台公开了他们的研究成果精准医学本体(Precision Medicine Ontology, PMO)，这都为中文医学领域知识图谱的构建工作提供了极富价值的参考。

二、全科医学知识图谱构建

全科医学知识图谱旨在搜集各类医学知识，通常包含大量的疾病、症状、检查、治疗、用药等多方面的实体及语义关系，通过对这些数据进行分析和整合构建成知识图谱，为医学领域的智能化发展提供帮助。开放资源是早期全科医学知识图谱构建的主要数据来源，有学者以公共医疗网站上爬取的数据和真实医院的电子临床数据为数据源，搭建了 MEDLedge 系统，该系统可对医学数据进行有效的处理和分析，并绘制成知识图谱以支持各种真实的医疗保健应用。刘燕等和魏自强等以垂直性医疗网站中的医学知识为数据基础构建了医疗知识图谱，并对其应用进行了探讨。此外，Shi 等还利用某城市卫生信息系统中的医疗服务数据构建了一个语义健康知识图谱，以便从这些离散的医疗文本数据中挖掘有价值的信息，使医疗信息系统中积累的数据得以充分利用。近年来，不断积累的电子病历数据也引起了学界的广泛关注，其中有些学者就利用电子病历构建了知识图谱，为临床决策提供支持。聂莉莉等还以权威的医学文献和书籍为数据源，在医学专

家的帮助下梳理了呼吸系统常见疾病及其症状之间的关系，以“疾病—症候—特征”为模型构建了呼吸系统医疗诊断知识图谱。阮彤等利用上海曙光医院已有的中医临床知识库作为数据基础，利用文本信息抽取和关系数据转换 (Relational Database to RDF, D2R) 等信息技术，通过领域专家构建的模式层将疾病库、症状库、中草药库和方剂库进行了融合，构建了一个中医药知识图谱。全科医学知识图谱的构建是医学数据向知识化转变的重要尝试，不同渠道的医学信息的积累也使得全科医学知识图谱的规模不断扩大。目前，中文全科医学知识图谱相关研究已有一定的积累，但如何提升数据的质量以满足医学具体应用场景的需求仍是亟待解决的关键问题。

三、单病种医学图谱构建

由于医学应用场景对知识精确度要求较高，全科医学知识图谱在数据精度方面的缺陷导致了其应用的局限性。近年来许多学者展开了对单病种医学知识图谱构建的研究。单病种医学知识图谱往往以某一疾病为核心节点，通过梳理该疾病的临床指南构建某疾病的知识模型，再结合一系列技术手段完成知识图谱的构建。目前单病种知识图谱涉及的疾病种类已经非常丰富，如有学者提出一种基于语义分析的医学知识图谱自动构建框架，并基于此框架利用 886 例高血压患者病历构建了高血压知识图谱。糖尿病知识图谱的构建也有学者进行了研究。精神疾病方面，有学者依据 UMLS 的概念层级和医学术语对从科研文献、临床指南、维基百科和电子病历中获取的抑郁症相关数据进行了整合，构建了抑郁症知识图谱，并开发了相应的系统对图谱进行管理和更新。还有学者基于癫痫患者的电子病历构建了癫痫知识图谱。此外，有人利用某三甲医院的甲状腺疾病患者的电子病历，结合医院已有的知识库进行甲状腺疾病相关实体和关系的抽取，构建了甲状腺疾病知识图谱，并采用样例数据测试了其辅助诊断的可用性。Fang 等从电子病历和医学网站（寻医问药网、百度百科和春雨医生）中抽取了垂体腺瘤相关信息，在临床专家的帮助下构建了垂体腺瘤知识图谱，为临床决策提供支持。另外，慢性肾脏病、心血管疾病以及近年突发的新冠肺炎，均有学者在其知识图谱构建及应用方面进行了研究。然而中文的单病种医学知识图谱大多针对较常见的疾病，国外已经有学者对罕见病知识图谱的构建与应用展开了研究，这也是中文医学知识图谱未来值得研究的方向。近年来，中文医学知识图谱构建的研究成果不断增加并呈现以下特点：一是图谱构建的数据来源趋于多样化，包括科研文

献、临床指南、医疗百科、电子病历等，二是图谱类型从全科医学知识图谱发展到单病种医学知识图谱，且涉及的疾病种类日益丰富，在应用层面也取得了较好的成果。然而中文医学知识图谱的研究仍存在一些难点和挑战。首先，医学本体对医学知识图谱的构建具有重要的指导意义，然而目前国内尚未形成通用的中文医学术语，不同研究采用的知识结构并不统一，这阻碍了现有医学知识图谱的融合，不利于研究的深入。其次，现有的中文单病种医学知识图谱大多针对较常见的疾病，如何利用知识图谱辅助罕见病的诊断和治疗也是未来亟待解决的问题。

资料来源：范媛媛，李忠民. 中文医学知识图谱研究及应用进展[J/OL]. 计算机科学与探索. <https://kns.cnki.net/kcms/detail/11.5602.TP.20220524.0921.002.html>.

医学用药推荐系统研究

医学上的用药推荐与一般的推荐算法不同，一般的推荐算法是根据用户的历史记录，利用数学算法推测出用户可能的需求，已被广泛应用于电商等互联网场景。而用药推荐则是基于循证医学的原则，结合患者的具体患病情况以及医学专业知识，推荐适合的用药方案。一般的推荐算法的推荐结果对准确率的容忍度较高，即使部分推荐结果与用户需求不符，也能够接受。但用药推荐在实际应用中要求达到百分之百的准确率，即药品一定能够起到作用，且不能产生不良反应或药品间的相互作用。

知识图谱能够更加清晰准确地表达疾病与药品之间的适应关系以及药品间的相互作用，基于知识图谱的用药推荐与其他人工智能方法相比，能够取得更好的效果。目前基于知识图谱的用药推荐研究进展与其他基线水平相比有所提升，但还无法达到实际应用的要求。

1. 医学用药的方法

目前医疗用药推荐系统使用的方法主要有以下两种。

第一种是图卷积网络的方法，即在图上使用卷积神经网络。2018年 Shang J Y 等人通过一个存储模块将药物相互作用（drug-drug interaction, DDI）的知识图谱集成为一个图形卷积网络，并将纵向患者向量建模作为查询，该方法在所有有效性度量方面都优于所有基线方法，并且在现有电子健康记录（electronic health record, EHR）数据中实现了 3.60% 的 DDI 率降低（即推荐药品之间有相

互作用的概率降低 3.6%)。2019 年, Wang S S 等人提出了一种针对药物组合预测 (medicine combination prediction, MCP) 的图卷积强化学习模型。其将 MCP 任务转换为无序马尔可夫决策过程 (Markov decision process, MDP) 问题, 并设计了一个深度强化学习机制来学习药物之间的相关性和不良相互作用。相比于 GAMENet, CompNet 在 Jaccard 和 F1 分数标准上分别提高了 3.74%、6.64%。2020 年, Kwak H 等人构造了一个药物疾病图谱, 使用图神经网络学习节点表示, 根据学习到的节点表示来预测药物节点和疾病节点是否具有药物不良反应 (adverse drug reaction, ADR) 关系。与其他算法相比, 该模型的接受者工作特征曲线下的面积 (area under curve of receiver operating characteristic, AUROC) 和精度召回率曲线下的面积 (area under curve of precision recall curve, AUPRC) 性能分别提高到 0.795 和 0.775。

第二种是知识图谱嵌入的方法, 包括将实体和关系转化为连续的向量空间, 从而简化操作, 同时保留知识图谱的原有的结构。2017 年 Wang M 等人构建了患者疾病-药品图谱, 将其嵌入低维空间后, 进行用药推荐。首先构建疾病和药品图谱, 通过 EHR 分别连接疾病和药品图谱, 形成两个二分图, 通过患者数据将两个二分图连接起来, 并构建了一个高质量的异构图, 该方法的预测准确度 (即 Jaccard 系数)、药物相互作用发生率、冷启动 (即没有患者数据时的使用)、临床专家评分均高于基线水平。2019 年 Wang X Y 等人构建了疾病-药品图谱, 将其嵌入低维空间后, 进行用药推荐, 并提出一种基于知识图谱嵌入增强主题模型 (knowledge graph enhanced topic model, KGETM) 的中药推荐模型。在中药基准数据集上的实验结果表明, 该方法优于当时最新的方法, 中药知识图谱嵌入在中药推荐中有很好的应用前景。

2. 医学用药的研究方向

医学知识图谱在用药推荐系统应用领域的未来研究方向主要有以下几方面。

① 构建完整的医学知识图谱。人类对疾病与药品的认识是动态变化的, 结合疾病、症状、药品、药品间的相互作用及患者的临床数据、患病的时间序列信息等, 构建一个完整的医学动态知识图谱, 确保知识的完整性、准确性和时效性。

② 知识图谱嵌入学习是将实体和关系映射到低维连续向量空间的表示方法, 在保留知识图谱结构信息的同时, 还能够改善数据稀疏问题, 提高计算效率, 因此在进行后续用药推荐任务之前, 先对知识图谱进行表示学习是很有必要的。

③ 考虑到构建动态医学知识图谱的必要性，而目前大多数知识嵌入表示研究建立在静态的知识图谱上，如何对动态知识图谱进行有效的知识表示是一个待解决的问题。将图时空网络与动态知识图谱相结合的知识嵌入表示用于用药推荐是一个颇具价值的研究方向。

资料来源：谭玲, 鄂海红, 匡泽民, 等. 医学知识图谱构建关键技术及研究进展[J]. 大数据, 7(4): 2021040.

技术应用

医学知识图谱的应用场景

知识图谱按领域可分为通用知识图谱和领域知识图谱，医学知识图谱是一种重要的领域知识图谱，它在语义搜索、知识问答和临床决策支持等智慧医疗领域都有很好的发展前景。

一、语义搜索

语义搜索不拘泥于用户所输入请求语句的字面本身，而是透过现象看本质，准确地捕捉到用户所输入语句后面的真正搜索意图，从而更准确地返回最符合用户需求的搜索结果。将知识图谱应用于搜索是当前实现语义搜索的有效解决方案，知识图谱描述了事物的分类、属性和关系，具有丰富的语义信息，可以为语义搜索提供极大的底层支持。基于医学知识图谱的语义搜索目前被用于医学百科知识、临床指南/文献、医学健康资讯、医疗保健信息等内容的推荐。

二、知识问答

基于知识库的问答（Knowledge-Based Question Answering, KBQA）也称知识问答，主要依托于大型的知识库，将用户的自然语言问题转化成结构化查询语句，直接从知识库中导出用户所需的答案。KBQA 已经成为各种智能问答系统的标准组件配置，知识图谱由于具有知识丰富、结构化程度高和易于推理等特点，是 KBQA 的知识库中最重要组成部分之一。

医学知识较其他领域相比专业性更高，非专业人士很难通过自主理解一堆资源文档，来精准地找到相关问题的答案。因此，基于医学知识图谱的 KBQA 可以帮助患者更加快捷、便利地获得问题的答案，适用于医学知识科普、智能导诊、自诊等领域。

三、临床决策支持

临床决策支持（Clinical Decision Support, CDS）是指运用相关的、系统的临床知识和患者基本信息及病情信息，向临床医务工作者提供加强医疗相关的决策和行动的信息，提高医疗质量和医疗服务水平。具有临床决策支持功能的系统，称为临床决策支持系统（Clinical Decision Support System, CDSS），一般由知识库、推理机和人机交互接口三个部分组成，其中知识库是核心。

医学知识图谱中包含丰富的医学基础知识，可以作为 CDSS 知识库中的重要组成部分，它为 CDSS 的推荐结果提供了可解释的依据，能够协助系统从辅助诊断、治疗方案推荐、合理用药检测等方面为临床医务人员提供决策支持。

四、 药物研发

近年来，人工智能辅助药物研发逐渐成为研究的热点。知识图谱可应用于药物研发的知识的聚类分析，帮助提出新的可以被验证的假说，从而加速药物研发的过程，降低研发成本。

五、 公共卫生事件应对

在流行病学调查和疫情发生事件的分析和预警等公共卫生事件场景中，知识图谱采用图存储数据的理念可以起到非常大的帮助。例如，利用知识图谱的形式可以直观地表示流行病调查中的人员分布、人员活动轨迹、发病时间等信息，基于图展示出的信息可以更方便的用于病例之间相关性的分析，更快地分析和梳理出感染源头。另外，对疫情发生事件的脉络进行分析，通过找到多个事件存在的因果关系，构建疫情相关事件知识图谱，帮助发现潜在的公共威胁，从源头上预防和降低舆情风险。

资料来源：医学知识图谱的价值与应用场景[EB/OL].（2022-04-08）.[2022-10-09].

<https://zhuanlan.zhihu.com/p/495429556>.

医学自然语言处理数据集

国内外医学自然语言处理相关数据集有以下一些：

一、 中文评测数据集

1. Yidu-S4K：医渡云结构化 4K 数据集

数据集描述：

Yidu-S4K 数据集源自 CCKS 2019 评测任务一，即“面向中文电子病历的命名实体识别”的数据集，包括两个子任务：

1) 医疗命名实体识别：由于国内没有公开可获得的面向中文电子病历医疗实体识别数据集，本年度保留了医疗命名实体识别任务，对 2017 年度数据集做了修订，并随任务一同发布。本子任务的数据集包括训练集和测试集。

2) 医疗实体及属性抽取（跨院迁移）：在医疗实体识别的基础上，对预定义实体属性进行抽取。本任务为迁移学习任务，即在只提供目标场景少量标注数据的情况下，通过其他场景的标注数据及非标注数据进行目标场景的识别任务。本子任务的数据集包括训练集（非目标场景和目标场景的标注数据、各个场景的非标注数据）和测试集（目标场景的标注数据

数据集地址

<http://openkg.cn/dataset/yidu-s4k>

2.瑞金医院糖尿病数据集

数据集描述：

数据集来自天池大赛。此数据集旨在通过糖尿病相关的教科书、研究论文来做糖尿病文献挖掘并构建糖尿病知识图谱。参赛选手需要设计高准确率，高效的算法来挑战这一科学难题。第一赛季课题为“基于糖尿病临床指南和研究论文的实体标注构建”，第二赛季课题为“基于糖尿病临床指南和研究论文的实体间关系构建”。

官方提供的数据只包含训练集，真正用于最终排名的测试集没有给出。

数据集地址

<https://tianchi.aliyun.com/competition/entrance/231687/information>

3.Yidu-N7K：医渡云标准化 7K 数据集

数据集描述：

Yidu-N4K 数据集源自 CHIP 2019 评测任务一，即“临床术语标准化任务”的数据集。

临床术语标准化任务是医学统计中不可或缺的一项任务。临床上，关于同一种诊断、手术、药品、检查、化验、症状等往往会有成百上千种不同的写法。标准化（归一）要解决的问题就是为临床上各种不同说法找到对应的标准说法。有了术语标准化的基础，研究人员才可对电子病历进行后续的分析。本质上，

临床术语标准化任务也是语义相似度匹配任务的一种。但是由于原词表述方式过于多样，单一的匹配模型很难获得很好的效果。

数据集地址

<http://openkg.cn/dataset/yidu-n7k>

4. 中文医学问答数据集

数据集描述：

中文医药方面的问答数据集，超过 10 万条。

数据说明：

questions.csv：所有的问题及其内容。answers.csv：所有问题的答案。

train_candidates.txt, dev_candidates.txt, test_candidates.txt：将上述两个文件进行了拆分。

数据集地址

<https://www.kesci.com/home/dataset/5d313070cf76a60036e4b023/document>

数据集 github 地址

<https://github.com/zhangsheng93/cMedQA2>

5. 平安医疗科技疾病问答迁移学习比赛

数据集描述：

本次比赛是 chip2019 中的评测任务二，由平安医疗科技主办。chip2019 会议详情见链接：<http://cips-chip.org.cn/evaluation>

迁移学习是自然语言处理中的重要一环，其主要目的是通过从已学习的相关任务中转移知识来改进新任务的学习效果，从而提高模型的泛化能力。

本次评测任务的主要目标是针对中文的疾病问答数据，进行病种间的迁移学习。具体而言，给定来自 5 个不同病种的问句对，要求判定两个句子语义是否相同或者相近。所有语料来自互联网上患者真实的问题，并经过了筛选和人工的意图匹配标注。

数据集地址(需注册)

<https://www.biendata.com/competition/chip2019/>

6. 天池新冠肺炎问句匹配比赛

数据集描述：

本次大赛数据包括：脱敏之后的医疗问题数据对和标注数据。医疗问题涉及“肺炎”、“支原体肺炎”、“支气管炎”、“上呼吸道感染”、“肺结核”、“哮喘”、“胸膜炎”、“肺气肿”、“感冒”、“咳血”等 10 个病种。

数据共包含 train.csv、dev.csv、test.csv 三个文件，其中给参赛选手的文件包含训练集 train.csv 和验证集 dev.csv，测试集 test.csv 对参赛选手不可见。

每一条数据由 Category, Query1, Query2, Label 构成，分别表示问题类别、问句 1、问句 2、标签。Label 表示问句之间的语义是否相同，若相同，标为 1，若不相同，标为 0。其中，训练集 Label 已知，验证集和测试集 Label 未知。

数据集地址(需注册)

<https://tianchi.aliyun.com/competition/entrance/231776/information>

二、中文知识图谱

CMeKG

地址

<http://cmekg.pcl.ac.cn/>

简介：CMeKG（Chinese Medical Knowledge Graph）是利用自然语言处理与文本挖掘技术，基于大规模医学文本数据，以人机结合的方式研发的中文医学知识图谱。CMeKG 的构建参考了 ICD、ATC、SNOMED、MeSH 等权威的国际医学标准以及规模庞大、多源异构的临床指南、行业标准、诊疗规范与医学百科等医学文本信息。CMeKG 1.0 包括：6310 种疾病、19853 种药物（西药、中成药、中草药）、1237 种诊疗技术及设备的结构化知识描述，涵盖疾病的临床症状、发病部位、药物治疗、手术治疗、鉴别诊断、影像学检查、高危因素、传播途径、多发群体、就诊科室等以及药物的成分、适应症、用法用量、有效期、禁忌证等 30 余种常见关系类型，CMeKG 描述的概念关系实例及属性三元组达 100 余万。

三、外文数据集

PubMedQA: A Dataset for Biomedical Research Question Answering

数据集描述：基于 Pubmed 提取的医学问答数据集。PubMedQA has 1k expert-annotated, 61.2k unlabeled and 211.3k artificially generated QA instances.

资料来源：医学自然语言处理相关资源整理[EB/OL]. (2021-04-01). [2022-10-09].

<https://blog.csdn.net/demm868/article/details/115388303>.

新冠图谱与病历数据集的构建

和强调知识广度的通用知识图谱不同，领域知识图谱面向特定领域，突出专业深度，通过数据和知识双重驱动，从问题建模分析为出发点，对特定行业场景的多源异构信息进行整合和分析。行业数字化的推进促使越来越多的行业知识图谱落地。

一、新冠图谱

自 2019 年新型冠状病毒（COVID-19，下文简称新冠）爆发以来，人们的生活乃至生命健康都受到了极大的影响。随着近年来关于新冠疫情的不断发展，有关疫情的各类信息也在不断更新。为更好地助力抗“疫”，分析疫情发展始末，OpenKG 紧随疫情发展，联合国内数所顶级高校和科研机构、企业，构建出针对多种应用场景的知识图谱资源。新冠百科图谱[Openkg.Org, 2020a]以病毒、疾病、细菌为主体，以百度百科、中文维基百科、互动百科等作为主要数据源，并扩展治疗，疾病等相关内容，构建出实例规模 5,000 余个，三元组 35,000 余组，总计包含 250 余个类别的知识图谱，可用于面向新冠相关术语 的语义检索、智能问答，以及相关文档的智能搜索和推荐。新冠科研图谱[Openkg.Org, 2020b]搜集、整理并关联与新冠病毒相关的科研基础数据和科研文献，以 NCBI（美国国家生物技术信息中心网站）中的 Taxonomy 板块为基础构建知识图谱，实例超 20 万，三元组数量接近 200 万的规模，可用于病毒的生物学分类、病毒变异性、潜在治疗药物，病毒传播途径与 种类等的预测任务，具有很强的潜在应用价值。目前仍在不断进行数据补充和完善，扩展出病毒分类图谱，新冠基本信息图谱，抗病毒药物图谱，病毒亲缘关系图谱等具体细分图谱。新冠临床图谱[Openkg.Org, 2020c]从目前已有的规范文件入手，基于诊疗规范、研究进展、发病统计，以新冠肺炎为核心延展至病毒、治疗方案、症状、方剂等各类概念，形成新冠临床知识图谱，实现可基于该图谱进行知识问答的应用落地。新冠英雄图谱[Openkg.Org, 2020d] 149 包括了医疗专家组、因公殉职英雄、武汉当地和全国各地的意见领袖等，涉及生平事迹和基本属性，并与新冠百科、新冠科研、新冠临床等图谱中的一些概念或实体关联。图谱以新冠病毒专家为核心延展至履历、成

果、事件、战役等各类概念，形成新冠英雄知识图谱，可基于图谱进行英雄人物动态展示。除此之外，OpenKG 还开源了新冠热点事件图谱[Openkg.Org, 2020e]，新冠健康图谱，新冠物资图谱，新冠流行病学图谱等。

二、病历数据集

目前，越来越多的医疗领域知识图谱数据集开源，从细分内容来看，现有的医学数据集可分为电子病历数据集、中医疗法剂方数据集、专科疾病数据集和常见疾病数据集等。高效规范地书写病历，一直是医务工作者的痛点。Yidu-S4K [Yiduccloud.Com, 2019a]和 Yidu-N7K [Yiduccloud.Com, 2019b]是由医渡云提出的标准化电子病历数据集，两个数据集根据真实病例分布再经过医学人工编辑而成，前者面向中文电子病历的命名实体识别任务，包含医疗命名实体识别和医疗实体及属性抽取两个子任务，后者针对临床术语标准化子任务，是语义相似度匹配任务的一种。目前国内最大规模的电子病历知识图谱由之江实验室提供，覆盖了 18 大类医学标准术语集、包含 479 万医学概念实例、3,531 万概念相互关系以及 9,600 万篇文献知识关联，临床术语覆盖范围达到国际领先水平。此外，为了推动 CNER 系统在中文临床文本上的表现，CCKS 从 2017 年开始，每年都组织了面向中文电子病历的命名实体识别评测任务，并推出了相应数据集[Ccks, 2017; Ccks, 2018; Ccks, 2020]。151 由中医科学院中医药信息研究所搭建的中医药知识服务平台在中医理论的指导下，从中医古籍文献、病案中提取经典名方及其治法，系统收集中医理论和方法，集成八大知识库，包含中医药领域的信息标准、领域本体、术语系统、文献库、知识库等多种知识资源，内容涉及中药、方剂、针灸、临床、养生等领域，提供知识检索、知识问答、知识浏览、知识推荐等多种服务。例如，中成药知识图谱（TCMLS）是一个包含 10 余万个中医概念以及 100 余万个语义关系的大型语义网络，基本覆盖了中医药学科的概念体系，在完整性方面处于中医界的领先地位。它以中成药领域海量文献为基础，构建了以中成药应用为主题的大规模知识库，建立了以病、症、证为核心的囊括组成、适应症、禁忌属性等属性的中成药知识图谱，设计了面向中成药推荐的全局最优图谱路径算法，研发了中成药知识问答系统，同时嵌入中医临床辅助系统进行探索应用，有助于实现中成药准确性、有效性、经济性、安全性的应用目标。疾病数据集的典型代表有 DiaKG（糖尿病知识图谱数据集）、Disease KG（常见疾病

信息知识图谱)等。DiaKG 由两位经验丰富的内分泌专家设计标注指南,侧重“实体”和“关系”,定义了 18 类实体类型和 15 类医学关系,从 41 篇公开发表的糖尿病指南中收集了共计 22,050 个实体和 6,890 个关系,涵盖了近年来糖尿病垂直领域的热点研究内容。DiseaseKG 从“寻医问药”医疗网站上爬取原始数据,对爬取的数据进行预处理后筛选适合做知识存储的相关信息,共定义了 8 类实体(4.4 万实体量级),7 类疾病属性和 11 种关系(31 万关系量级),覆盖了常见的疾病。

资料来源:知识图谱发展报告(2022)[EB/OL].(2022-08).[2022-10-10].

www.cipsc.org.cn/sigkg/wp-content/uploads/知识图谱发展报告 2022.pdf.

相关产品

工业界医疗知识图谱产品

工业界有名机构已经开发并推出了自己的医疗知识图谱产品。分别如下:

一、百度——灵医智惠

灵医智惠是由百度大脑技术驱动的 AI 医疗品牌。灵医智惠技术中台以医学数据结构化及医学知识图谱为基础,构建多项医疗专项能力,覆盖临床、科研、管理、患者服务等多环节。

以灵医智惠“爱助医”为例,该产品能提供医学知识服务、临床辅助决策、智能随访服务和区域健康驾驶舱等四项核心应用,且服务范围辐射院内、院外和区域三大场景。在临床辅助决策方面,通过学习海量教材、临床指南、药典及三甲医院优质病历,基于百度自然语言处理、知识图谱等多种 AI 技术,打造遵循循证医学的临床辅助决策系统,用以提升医疗质量,降低医疗风险。包含辅助问诊、辅助诊断、治疗方案推荐、医学知识查询等多种功能。



灵医智惠“爱助医”解决方案整体架构

辅助诊断: 基于主诉、现病史中提到的症状、疾病，以及体格检查、检验结果推荐相关的疾病、相关症状及体征，按照诊断结果由高到低推断潜在疾病可能性，实时抓取信息并提示，涵盖 4000 余种常见病，且全病种 Top 3 推荐准确率 89%。

相似病历推荐: 基于病历文书实时分析，帮助临床医生从历史病例特点、诊疗方法、治疗效果中学习及借鉴诊疗思路。全病种 Top 3 历史病例匹配准确率 89%。

医学知识查询: 通过认知计算、知识推理等技术通过查询关键词提供候选知识列表。所有知识来源均来自可循证的权威知识书籍、临床指南、药品说明书等，为医务人员提供实时可靠的医学知识，其中精选疾病知识 9700+，典型病例 3600+，国家临床路径 1200+，医学词典 17 万+，医学计算器 270+，心电图 760+，临床检验 330+，多维临床决策知识 2600+。

灵医智惠包括临床辅助决策、医疗大数据解决方案、眼底影响分析和智能诊前助手四大模块。医疗大数据解决方案专注于使用 NLP 和知识图谱技术对电子病历进行多层级深层次的结构化。智能诊前助手是基于专业医疗知识图谱，采用多种算法模型与多轮智能交互理解病人病情，根据病人病情精准匹配就诊科室和专业医生。

二、中国平安-平安好医生

2019 年平安智慧医疗正式推出中文医疗知识图谱。该图谱集成了 60 万医学

概念、530 万医学关系、千万医学证据，覆盖核心医学概念。



平安知识图谱整体架构

从上图可以看出，整个知识图谱体系分为四层：

数据层：数据源主要来自于医学领域核心概念及丰富的临床证据，并通过 RWE（真实世界证据）和经典病例分析形成医学知识

图谱构建：使用知识图谱及 NLP 相关技术来构建图谱

数据平台：完成数据的获取、清洗处理、存储和分析

智能服务：提供可视化、问答、决策等服务

目前基于此医疗知识图谱和更多的数据源，平安医疗开发出多个智能服务的应用场景，组件了医疗科技平台。以旗下的平安好医生'AI 辅助诊疗系统'为例，通过智能语音语义解析及大数据等技术赋能，集合了超过 3 亿条在线诊疗及健康咨询数据，用于在线医疗咨询的预诊分诊、问诊等环节。

三、阿里巴巴-阿里健康

阿里健康还拥有一款智能医疗系统——Doctor You。目前 Doctor You 系统包括临床医学科研辅助平台、医疗人工智能开放平台以及临床医师技能训练平台等。

1. 临床医学科研辅助平台

该系统是以智能搜索为主的大数据科研辅助分析引擎，通过数据仓库、数据挖掘等方法，用海量临床科研数据来评价治疗效果，发现诊疗规律，提炼最佳治疗路径，提高医疗科研的能力和水平。其科研数据平台包括：

知识图谱：通过建立病历研究专题相关的百科知识图谱，如疾病介绍、症状

介绍等。构建诊疗之间的关系，例如：疾病与症状、体征、异常指标、检查项目、检验项目、药物、手术之间的关系；症状与疾病、伴随症状、异常指标、检查项目、检验项目、药物、手术之间的关系；药物与疾病、症状、异常指标、检查项目、检验项目、手术之间的关系。

自然语言处理：信息系统中存在很多次的同义词情况，比如冠心病、冠状动脉粥样硬化之类的词。利用自然语言处理算法，在结构化数据的同时，根据词向量形成同义词簇，由人工参与后，确认同义词之间的关系，形成确定的、准确的同义词词库

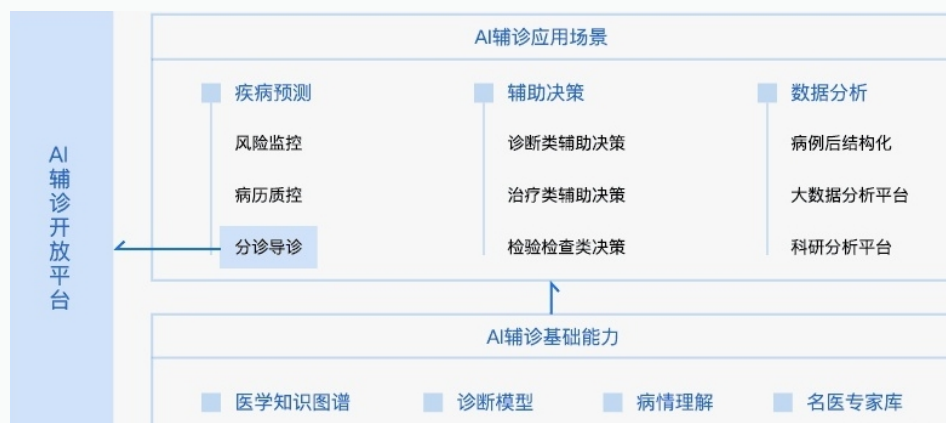
机器学习：提供一站式数据挖掘流程，数据上云之后，无需做任何线下处理即可实现业务逻辑。提供了数据预处理、自然语言处理、特征工程、机器学习算法、预测和评估、在线预测这一套流程的全部组件

2. 医疗人工智能开放平台

该系统提供医疗 AI 建模、训练及开放应用服务，针对医疗机构真实临床场景，提供智能肺、乳腺 X 射线、肿瘤靶区勾勒等多部位、多病种医学 AI 系统应用及糖尿病用药、宫颈癌筛查等 AI 辅助诊断决策系统应用平台。

四、腾讯-觅影

腾讯觅影发布 AI 辅诊开放平台，旨在依托觅影在医疗领域积累的医学知识图谱、诊断模型、病情理解、名医专家库等 AI 辅诊基础能力，提供疾病预测、辅助决策、数据分析等功能。



辅诊开放平台

疾病预测功能旨在多方位帮助医生减轻工作量、辅助医生临床诊断。疾病预测由三个模块构成：

智能导诊: 通过对医学文献、病历、问答等医疗大数据的深度学习，进行症状推理，预测范围覆盖全部科室、全部常见疾病，基本满足 90% 的导诊需求。用户与机器人以问答的形式进行交互。

病案管理: 通过深度学习技术自动提取病案特征信息，减少人力与资金的投入，提高医院病案整理的效率，同时助力医院完成工作量巨大而且推行困难的数据治理工作。

风险监控: 通过 AI 来辅助医生鉴别是否误诊，降低高危疾病漏诊带来的巨大风险。

资料来源: 初探医疗知识图谱[EB/OL]. (2020-09-28). [2022-10-10].

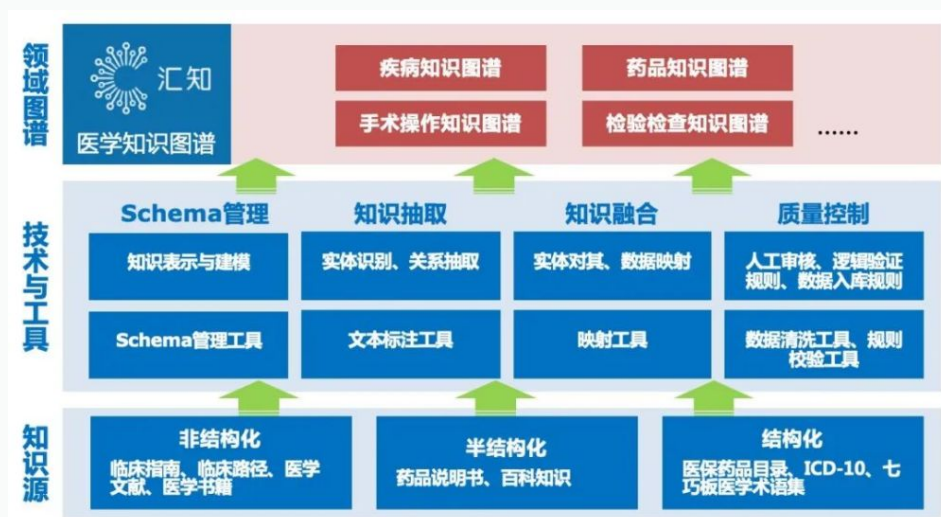
<https://zhuanlan.zhihu.com/p/258838308>.

“汇知”医学知识图谱简介

“汇知”医学知识图谱（Wisdomed Medical Knowledge Graph，简称“汇知”图谱）致力于建立符合中文环境下的规范、共享、易用、动态、系统的医学知识图谱。以满足临床智慧诊疗的知识应用需求为切入点，围绕“诊前-诊中-诊后”流程开展疾病、药品、检验检查、手术操作四大细分领域知识图谱构建，为医疗领域提供广泛的知识支撑。

一、“汇知”图谱的构建

“汇知”图谱基于临床指南、临床路径、药品说明书、医学书籍和医学文献等高质量医学资源，采用机器+人工方式进行构建，专家全流程参与，严格把关数据质量，真实还原医学知识。同时，“汇知”图谱与“七巧板”术语集建立映射关系，实现数据的标准化，从而促进知识更深度地应用。



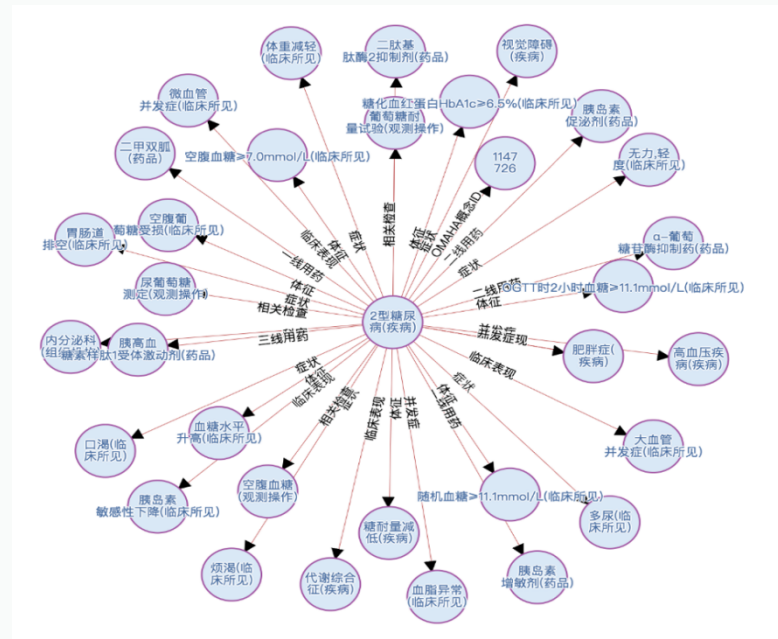
“汇知”图谱总体架构

二、“汇知”图谱资源

“汇知”图谱根据不同领域划分，可分为疾病、药品、手术操作和检验检查四大知识图谱。目前已发布疾病和药品知识图谱，共计 11.7 万实体、91.4 万三元组，后续将发布手术操作和检验检查知识图谱。

(1) 疾病知识图谱：以疾病为中心，链接超过 3 万种疾病的临床表现、科室、相关检查、治疗方式、症状等核心关系。共包含 7.5 万实体，34.5 万三元组。为辅助诊疗、知识问答等系统的研发提供高质量、结构化知识库基础，助力疾病诊疗水平的提升。

(2) 药品知识图谱：以药品为中心，链接超过 2.3 万种药品的适应证、适用人群、禁忌证、禁忌人群等核心关系，共包含 4.8 万实体，59.0 万三元组。为助力临床合理用药等场景提供可靠的智能引擎。



“汇知”医学知识图谱示例

三、“汇知”图谱获取方式

下面两幅图显示了“汇知”图谱的获取途径，及图谱资源。



资料来源：汇知医学知识图谱协作项目[EB/OL]. (2020-11-25). [2022-10-10].

<https://www.omaha.org.cn/index.php?g=&m=article&a=index&id=319&cid=45>.

