计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.2018.20180623
55(12):2587-2599,2018

知识图谱研究综述及其在医疗领域的应用

侯梦薇 卫 荣 陆 亮 兰 欣 蔡宏伟
(西安交通大学第一附属医院网络信息部 西安 710061)
(houmengwei777@xjtu.edu.cn)

Research Review of Knowledge Graph and Its Application in Medical Domain

Hou Mengwei, Wei Rong, Lu Liang, Lan Xin, and Cai Hongwei
(*Network and Information Department, the First Affiliated Hospital of Xi'an Jiaotong University, Xi'an* 710061)

Abstract With the advent of the medical big data era, knowledge interconnection has received extensive attention. How to extract useful medical knowledge from massive data is the key for medical big data analysis. Knowledge graph technology provides a means to extract structured knowledge from massive texts and images. The combination of knowledge graph, big data technology and deep learning technology is becoming the core driving force for the development of artificial intelligence. The knowledge graph technology has a broad application prospect in the medical domain. The application of knowledge graph technology in the medical domain will play an important role in solving the contradiction between the supply of high-quality medical resources and the continuous increase of demand for medical services. At present, the research on medical knowledge graph is still in the exploratory stage. The existing knowledge graph technology generally has several problems such as low efficiency, multiple restrictions and poor expansion in the medical domain. This paper firstly analyzes the medical knowledge graph architecture and construction technology for the strong professionalism and complex structure of big data in the medical domain. Secondly, the key technologies and research progress of the three modules of knowledge extraction, knowledge expression, knowledge fusion and knowledge reasoning in medical knowledge map are summarized. In addition, the application status of medical knowledge maps in clinical decision support, medical intelligence semantic retrieval, medical question answering system and other medical services are introduced. Finally, the existing problems and challenges of current research are discussed and analyzed, and its development is prospected.

Key words knowledge graph; medical wisdom; big data; knowledge fusion; natural language processing

摘 要 随着医疗大数据时代的到来,知识互联受到了广泛的关注.如何从海量的数据中提取有用的医学知识,是医疗大数据分析的关键.知识图谱技术提供了一种从海量文本和图像中抽取结构化知识的手段,知识图谱与大数据技术、深度学习技术相结合,正在成为推动人工智能发展的核心驱动力.知识图谱技术在医疗领域拥有广阔的应用前景,该技术在医疗领域的应用研究将会在解决优质医疗资源供给不足和医疗服务需求持续增加的矛盾中产生重要的作用.目前,针对医学知识图谱的研究还处于探索阶段,现有知识图谱技术在医疗领域普遍存在效率低、限制多、拓展性差等问题.首先针对医疗领域大数据

收稿日期:2018-09-12;修回日期:2018-10-18
通信作者:卫荣(weirong@xjtu.edu.cn)

专业性强、结构复杂等特点，对医学知识图谱架构和构建技术进行了全面剖析；其次，分别针对医学知识图谱中知识表示、知识抽取、知识融合和知识推理这4个模块的关键技术和研究进展进行综述，并对这些技术进行实验分析与比较. 此外，介绍了医学知识图谱在临床决策支持、医疗智能语义检索、医疗问答等医疗服务中的应用现状. 最后对当前研究存在的问题与挑战进行了讨论和分析，并对其发展前景进行了展望.

关键词 知识图谱；智慧医疗；大数据；知识融合；自然语言处理

中图法分类号 R391；TP399

随着信息技术和互联网技术的持续发展和不断变革，人类先后经历了以文档互联为主要特征的“Web 1.0”时代与数据互联为特征的“Web 2.0”时代，正在迈向基于知识互联的崭新“Web 3.0”时代[1]. 知识互联的目的是构建一个人与机器都能够理解的万维网，使网络更加智能化. 然而，由于万维网上的内容多源异质、组织结构松散，给大数据环境下的知识互联带来了极大的挑战[2]. 因此，人们需要依据大数据环境下的组织知识原则，探索既符合网络信息资源不断变化规律又能切合适应用户认知需求的知识互联方法[3]，使其更加深刻地展示整体而相互关联的人类认知世界[4]. 知识图谱(knowledge graph)就是在这样的大背景下产生的一种知识表示和管理的方式，它是实现智能化语义检索的基础和桥梁，为万维网上知识互联奠定坚实的基础[5].

知识图谱的概念于2012年5月被Google正式提出[6]，其原始目的是为了提高搜索引擎的能力，提高搜索结果质量并提升用户的搜索体验. 2013年之后，随着智能信息服务和应用的不断发展，知识图谱已在学术界和工业界普及，并在智能搜索、智慧问答大数据风控、推荐系统等应用中发挥着重要的作用[7]. 目前，医学是知识图谱应用最广的垂直领域之一，也是目前国内外人工智能领域研究的热点，在如疾病风险评估、智能辅助诊疗、医疗质量控制及医疗知识问答等智慧医疗领域都有着很好的发展前景[8]. 目前很多公司均构建了自己的知识图谱，如IBM的Watson Health①、阿里健康的“医知鹿”医学智库②、搜狗的AI医学知识图谱APGC等医学知识图谱的应用近2年来也开始进入人们视线. 在医学领域，典型的医疗知识图谱包括SNOMED-CT③，IBM的Waston Health以及国内的诸如上海曙光医院的中医药知识图谱[9]等. 随着区域卫生信息化及医疗信息技术的发展，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应用，是推进智慧医疗辅助的关键[10]，也是医学知识检索、辅助诊疗、医疗质量控制、电子病历及健康智能化管理应用的基础，对于提高医生诊疗水平、减轻医生负担具有非常重要的意义.

1 知识图谱的定义与架构

1.1 基本定义

知识图谱是语义网(semantic Web)技术之一，是一种基于图的数据结构，由节点(实体)和标注的边(实体间的关系)组成[11]，它本质上是一种揭示实体之间关系的语义网络，可以对现实世界的事物及其相互关系进行形式化地描述[12]. 和传统的Web页面网络相比，知识图谱中的节点由网页形式变成了不同类型的实体，而图中的边也由连接网页的超链接(hyperlink)变成了实体间丰富的各种语义关系.

三元组是知识图谱的一种通用表示形式[13]，由2个具有语义连接关系的医疗实体和实体间关系组成，是医学知识的直观表示，即$G=(head, relation, tail)$，其中，$head$为三元组中的头实体，$tail$为三元组中的尾实体，$relation=\{r_1, r_2, \cdots, r_{|R|}\}$是知识库中的关系集合，共包含$|R|$种不同关系. 三元组的基本形式主要包括实体1、关系、实体2和概念、属性、属性值等. 实体是知识图谱中的最基本元素，例如疾病、药物、症状、辅助检查、科室、手术、部位等；关系存在于不同实体之间，例如临床表现、病因、发病机制、科室、预防、药动学、药理作用、鉴别、诊断等；概念主要指集合、类别、对象类型、事务的种类等，例如慢性病、二级手术、抗凝血药物等；属性主要指对象

① https://www.ibm.com/watson

② https://www.mdeer.com

③ http://www.snomed.org

可能具有的特点、特征及参数,例如疾病特征、药品规格、手术类型等;属性值指对象特定属性的值,例如前面提到的疾病特征为多尿、药品规格为 0.2 mg/支等.通过一个全局唯一的 ID 号来标识实体,实体间内在特征通过属性-属性值对来刻画,实体之间的关联通过关系来描述.三元组的存在表示一个已有的事实,即实体处于给定类型的关系中.例如支气管扩张症的描述为:支气管扩张症(bronchiectasis)多见于儿童和青少年,临床表现为慢性咳嗽、咳浓痰和(或)反复咳血.

支气管扩张症的描述通过表 1 中的三元组表示:

Table 1 Ternary Representation of Bronchiectasis Description
表 1 支气管扩张症描述的三元组表示

Entity	Relation	Entity
Bronchiectasis	Called	BE
Bronchiectasis	HappenIn	Children
Bronchiectasis	HappenIn	Teenager
Bronchiectasis	Cause	Chronic Cough
Bronchiectasis	Cause	Phlegm
Bronchiectasis	Cause	Hemoptysis

我们可以将所有三元组合并构成一个多图(multigraph),其中节点表示实体,有向边表示实体之间的关系,边的方向表明了实体是作为主体还是对象出现.不同的关系通过不同类型的边来表示(也被称为边标签),知识图谱的结构有时也被称为异构信息网络(heterogeneous information network)[14].如图 1 所示:

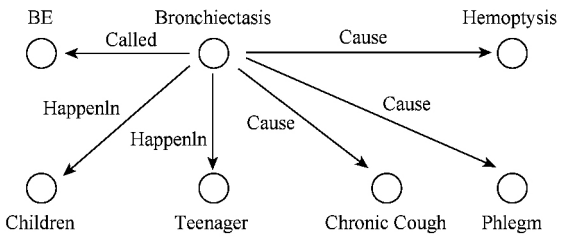


Fig. 1 Knowledge graph example
图 1 知识图谱示例

知识图谱按照覆盖范围可分为通用知识图谱和行业知识图谱.通用知识图谱强调融合更多实体,其准确度不够高,且很难借助标准知识库规范其实体、属性和关系等,主要被应用于智能搜索等领域中;行业知识图谱通常依靠特定行业的数据进行构建,对特定行业有重要的意义.行业知识图谱需要考虑从不同的业务场景和使用人员,所以实体的属性与数据模式比较丰富.本文所探讨的医疗知识图谱就属于行业知识图谱.

1.2 体系架构

构建医学知识图谱的主要目的是抽取大量的、让计算机可读的医学知识[15].在医疗信息技术飞速发展的今天,医学知识大量存在于非结构化的文本数据、半结构化的表格、网页以及部分医疗信息系统的结构化数据中,因此现有的医学知识图谱模型均为判别模型.通过训练使该模型能够区分不同关系的实体对,或者从随机抽样的无关负实体中识别有意义的实体对.为了阐述如何构建知识图谱,本文给出了医学知识图谱的体系架构,即其构建模式结构,如图 2 所示.

医学知识图谱主要有自顶向下(top-down)与自底向上(bottom-up)两种构建方式.自顶向下方式首先构建顶层关系本体,然后将抽取到的实体匹配更新到所构建的顶层本体中.自底向上的方式直接将抽取数据中发现的类别、实体、属性以及关系合并到知识图谱中.目前大部分知识图谱都采用自底向上的方式进行构建.

医学知识图谱的构建流程可以被归纳为 3 个模块,即医学知识抽取、医学知识融合以及医学知识计算.医学知识抽取通过从大量结构化、半结构化或非结构化的医学数据中提取出实体、关系、属性等知识图谱的组成元素,并选择合理高效的方式将元素存入知识库中.医学知识融合对医学知识库的内容进行整合、消歧、加工,增强知识库内部的逻辑性和表达能力,并为医学知识图谱更新旧知识或补充新知识.医学知识计算借助知识推理,推断出缺失事实,自动完成疾病诊断与治疗.

2 医学知识图谱的关键技术

理解、推理和归纳能力是人类智力的核心[16].然而对于机器而言,想要理解和推理出 2 个实体之间的关系具有很大的挑战.具体来说,现实世界中的医疗关系实体具有非常复杂的属性:首先,医学实体通常使用不同的语言表达.例如 nose plugged, blocked nose 和 sinus congestion 都代表着鼻塞,却有不同的表达方式;其次,实体之间的关联关系可能拥有不同粒度和强度.例如 $Disease \xrightarrow{Cause} Sympton$ 这一关系会包含例如〈鼻炎,导致,鼻塞〉的粗粒度实

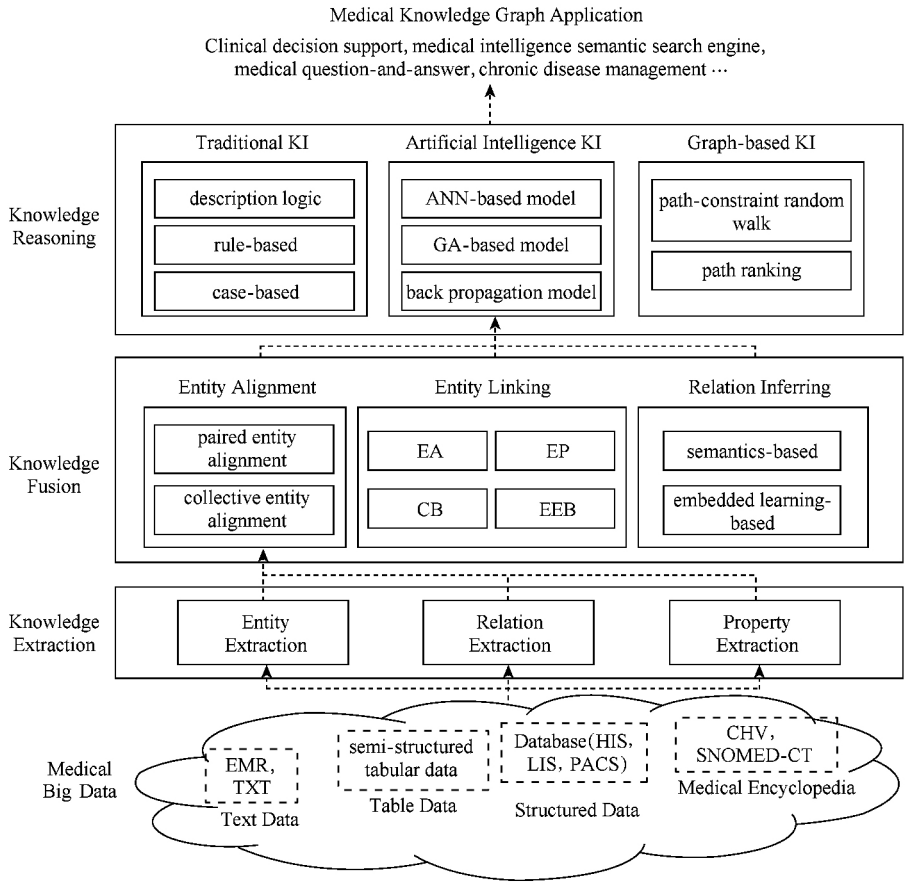


Fig. 2 Medical knowledge graph architecture

图 2 医学知识图谱架构

体对,也会包含例如〈急性鼻炎,导致,鼻塞〉,〈慢性鼻炎,导致,鼻塞〉的细粒度实体对. 对于关系的强弱程度,〈感冒,导致,疲劳〉的关系比〈感冒,导致,耳部感染〉的关系要强,因为感冒很少会引起严重的耳部感染. 这些知识对于人类来说非常简单,但是对于机器来说,要深刻理解实体之间关系的共性仍然是一个挑战.

医学知识图谱的构建与应用需要多种智能信息处理技术的支持[17]. 通过知识抽取技术,可以从半结构化、非结构化数据中提取知识要素. 借助知识融合技术,可以消除实体、关系、属性与对象之间的歧义,形成高质量医学知识库. 医学知识计算是在已有知识的基础上进一步挖掘隐含知识,从而丰富、扩展医学知识库. 本节将从医学知识表示、医学知识抽取、医学知识融合和医学知识推理所运用的关键技术为重点,详细说明其中的相关研究.

2.1 医学知识表示

三元组知识表示形式虽然受到了广泛的使用和认可,在应用于医学领域时却会出现计算效率低等问题. 近年来随着人工智能、机器学习、深度学习等表示学习技术的重大进展[18-19],医学实体中的语义信息可以表示为稠密低维实数值的向量,从而在低维度空间中计算实体和关系中的复杂语义关联,对于医学知识库的构建过程有重要意义. 医学知识表示按照计算方式不同可以分为距离平移模型(translational distance model)和语义匹配模型(semantic matching model). 其中距离平移模型利用基于距离的评分函数对事实的合理性进行判断,代表包括翻译模型(TransE)[20]及其延伸出的复杂关系模型(TransH, TransR, TransD, TransG, KG2E 等). 语义匹配模型的代表包括单层神经网络模型(single layer model, SLM)[21]、双线性隐变量模型(latent factor model, LFM)[22]、神经张量模型(neural tensor model, NTM)[23]、矩阵分解模型(matrix factorization, MF)[24]等.

距离平移模型 翻译模型

TansE 是最具代表性的距离平移模型,它将实体和关系表示为同一空间的矢量. 三元组中的关系矢量 $l_{relation}$ 可以被看作头实体矢量 l_{head} 到尾实体矢量 l_{tail} 的翻译,并满足关系:

$$\boldsymbol{l}_{head}+\boldsymbol{l}_{relation}\approx\boldsymbol{l}_{tail}, \tag{1}$$

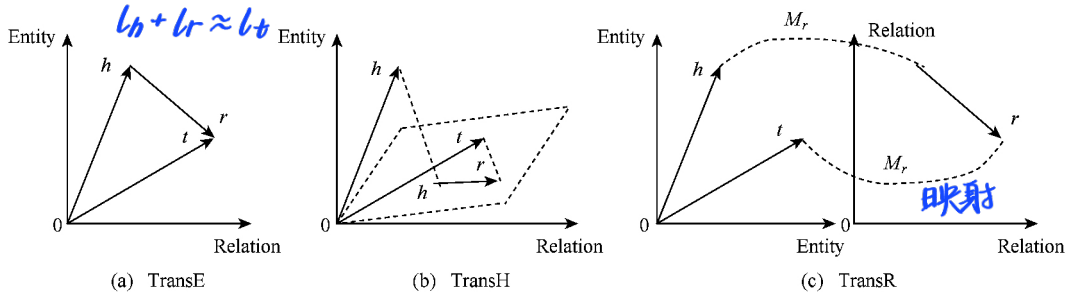
评价函数为

$$f_{relation}(head,tail)=|\boldsymbol{l}_{head}+\boldsymbol{l}_{relation}-\boldsymbol{l}_{tail}|_{L_1/L_2}, \tag{2}$$

翻译模型的参数较少,计算复杂度低,且适用于大规模稀疏医学知识库,性能和扩展性都比较好.图 3(a)为 TransE 模型的示例.

2) 复杂关系模型

复杂关系模型主要针对实体之间 1-to-N,N-to-1,N-to-N 的关系类型[25-26].这里简要介绍这 5 项代表性模型的原理.TransH 针对不同关系下的同一医学实体的角色问题进行研究,试图采用不同的形式表示不同关系中的医学实体.TransR 将不同的关系进行更细致的划分,保证了映射后模型的表达能力.TransD 分别对头实体和尾实体在盥洗空间进行投影矩阵的定义,保证了实体和关系之间的交互.TransG 和 KG2E 均采用高斯分布对医学实体和关系进行描述,具有较高的实体区分度.图 3(b)(c)为 TransH 和 TransR 模型的示例.



(a) TransE (b) TransH (c) TransR

Fig. 3 Schematic diagram of distance translation model

图 3 距离平移模型示意图

3) 单层神经网络模型

非线性的单层神经网络模型为医学知识库中的三元组$(head,relation,tail)$定义了评价函数:

$$f_g(head,tail)=\boldsymbol{\mu}_{tail}^{\mathrm{T}}\tanh(\boldsymbol{M}_{relation,1}l_{head}+\boldsymbol{M}_{relation,2}l_{tail}), \tag{3}$$

其中,$\boldsymbol{\mu}_{tail}^{\mathrm{T}}\in\mathrm{R}^k$为关系 $relation$ 的向量化表示;$\boldsymbol{M}_{relation,1}$和 $\boldsymbol{M}_{relation,2}$是通过关系 $relation$ 定义的矩阵;l_{head}和 l_{tail}是头实体和尾实体的向量化表示.图 4(a)为单层神经网络模型的示例图.

单层神经网络模型基于实体之间的关系,刻画了医学实体的语义相关性,从而解决了医学实体之间协同性较差的问题,但计算复杂度较高,不适用于大规模医学知识图谱的表示.

4) 双线性隐变量模型

双线性隐变量模型基于医学实体间关系的双线性变化对实体的语义相关性进行定义.评价函数:

$$f_{relation}(head,tail)=l_{head}^{\mathrm{T}}\boldsymbol{M}_{relation}l_{tail}, \tag{4}$$

其中,$\boldsymbol{M}_{relation}$是通过关系 $relation$ 定义的双线性变换矩阵;l_{head}和 l_{tail}是头实体和尾实体的向量化表示.图 5(a)为双线性隐变量模型的示例.

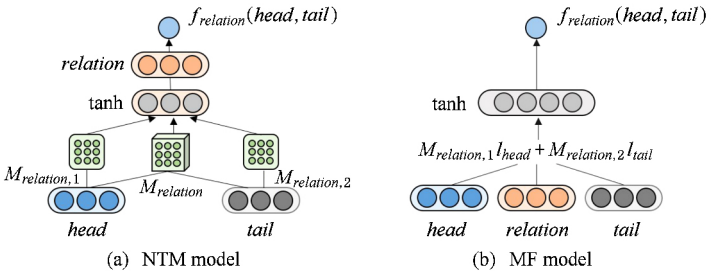
双线性隐变量模型形式简单,降低了计算复杂度,并有效刻画实体间相关性关系.

5) 神经张量模型

神经张量模型通过将医学实体中单词的向量取平均值来表示实体间语义联系.评价函数:

$$f_{relation}(head,tail)=\boldsymbol{\mu}_{tail}^{\mathrm{T}}\tanh(l_{head}\boldsymbol{M}_{relation}l_{tail}+\boldsymbol{M}_{relation,1}l_{head}+\boldsymbol{M}_{relation,2}l_{tail}+b_{relation}), \tag{5}$$

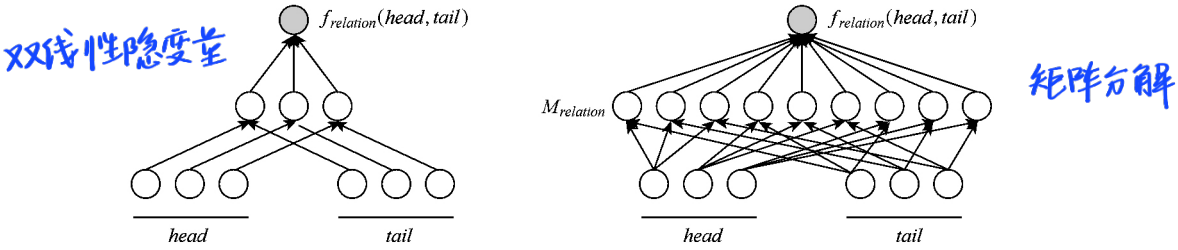
其中,$\boldsymbol{\mu}_{tail}^{\mathrm{T}}\in\mathrm{R}^k$为关系 relation 的向量化表示;$\boldsymbol{M}_{relation}$是一个三阶张量;$\boldsymbol{M}_{relation,1}$和 $\boldsymbol{M}_{relation,2}$是通过关系 $relation$ 定义的 2 个投影矩阵.图 4(b)为神经



(a) NTM model (b) MF model

Fig. 4 Schematic diagram of neural tensor model and single layer model

图 4 神经张量模型、单层神经网络模型示意图



(a) LFM model (b) MF model

Fig. 5 Schematic diagram of matrix factorization model and latent factor model

图 5 矩阵分解模型、双线性隐变量模型示意图

张量模型示意图.

神经张量模型通过取均值的方式解决了低维向量的"稀疏性"问题,并可以重复使用同一单词向量进行医学实体构建.

6) 矩阵分解模型

矩阵分解模型的主要作用是对向量进行降维,将每个三元组$(head, relation, tail)$表示为一个三阶张量 $\boldsymbol{X}_{head, relation, tail}$,如果该三元组存在,则张量对应位置值为 1,否则为 0. 张量值 $\boldsymbol{X}_{head, relation, tail}$ 分解为 $l_{head}^{\mathrm{T}} \boldsymbol{M}_{relation} l_{tail}$,并使 $|\boldsymbol{X}_{head, relation, tail} - l_{head}^{\mathrm{T}} \boldsymbol{M}_{relation} l_{tail}|_{L_2}$ 尽量小. 矩阵分解模型的示例如图 5(b)所示.

7) 模型性能对比

为了验证各模型的效率,本文对比分析了前文讨论过模型的时间复杂度和空间复杂度,如表 2 所示. 这里分别用 n 和 m 表示实体和关系的数量,d 和 k 分别表示实体的维数和关系嵌入空间的维数,TransG 中的 c 指每个关系的平均语义组件数量.

Table 2 Comparison of Time and Space Complexity

表 2 模型在时空复杂度上的比较

Model	Space Complexity	Time Complexity
SLM	$O(nd+md)$	$O(d^2)$
LFM	$O(nd+md)$	$O(d)$
NTM	$O(nd+md^2k)$	$O(d^2k)$
MF	$O(nd+md^2)$	$O(d^2)$
TransE	$O(nd+md)$	$O(d)$
TransH	$O(nd+md)$	$O(d)$
TransR	$O(nd+mdk)$	$O(dk)$
TransD	$O(nd+mk)$	$O(\max(d,k))$
TransG	$O(nd+mdc)$	$O(dc)$
KG2E	$O(nd+md)$	$O(d)$

2.2 医学知识抽取

医学知识抽取是面向开放的医疗数据,通过人工或自动化技术抽取出可用的知识单元,知识单元包括实体、关系及属性这 3 个知识要素,并以此为基础,形成一系列高质量的事实表达,为上层模式层的构建奠定基础.

人工抽取方式是依据一定规则收集并整理相关医学信息并提取知识,目前包括 ICD-10[27]、临床医学知识库、SNOMED-CT[28]等都是通过人工构建的医学知识库;自动抽取方式是利用数据挖掘、人工智能、机器学习等技术从医学信息中自动提取基本元素,一体化医学语言系统 UMLS 是通过自动提取方式构建的. 自动抽取方式是目前的研究重点,也是未来知识抽取的趋势. 本节主要介绍如何自动从医疗数据源中抽取知识,按照要素类型分为实体抽取、关系抽取和属性抽取.

1) 实体抽取

实体是医学知识图谱中的最基本元素,实体抽取的准确率和召回率等将直接影响知识库的质量,所以实体抽取是医学知识图谱技术的重点研究方向.

早期的实体抽取方法是在限定文本领域、限定语义单元类型的条件下进行,采用基于医学规则和医学字典的方法,使用已经定义好的医学规则,抽取出文本中的疾病名、药物名、症状名等实体. 例如文献[29]中通过 CHV 和 SNOMED-CT 两个医学词典对医疗诊所笔记中的医学信息进行识别,得到了不错的实验结果. 但是,这种方法在实施过程中具有极大的难度. ①目前没有一个完整的医学字典囊括所有类型的实体,所以无法使用文本匹配的方法对实体进行识别;②中文医学短语的含义根据上下文的改变而指代不同的实体;③需要疾病或药物实体拥有多个名称. 因此,基于医学规则和字典的实体抽取方法仅在最早期被广泛应用,难以适应数据不断变化的现实要求.

随后,研究者们尝试将机器学习和统计学算法应用到实体抽取问题上,利用医学数据的特点对模

型进行训练,然后识别实体.常用的方法包括支持向量机、人工神经网络、隐 Markov 模型、条件随机场等.文献[30]使用支持向量机模型进行生物医学命名实体识别,为了提高训练效果,引入词缓存、无监督训练等方法,实验结果表明:该方法在 GENIA 医学数据集中的准确率高于基准算法,并能高效地应用于大规模知识库中.文献[31]提出一种最大熵算法作为机器学习算法和基于规则字典的抽取方法的混合算法,并在 Medline 数据集进行实验,实验的准确率和召回率都在 70%以上.基于机器学习的实体抽取方法在运用于医学领域时面临着数据质量的良莠不齐及人工标注专业性不高等问题,目前的解决方法是利用海量未标注数据持续提升模型性能,从小样本中学习,形成一个交互学习过程,从而提升实体抽取的准确率.

深度学习是机器学习研究中的一个新的领域,其目的在于建立、模拟人脑进行分析学习的神经网络[32].它模仿人脑的机制来解释数据,例如图像、声音和文本,近年来被广泛应用于实体抽取中.目前 BiLSTM-CRF 是医学领域实体抽取中最主流的深度学习模型.文献[33]通过实验对比 BiLSTM-CRF 与其他机器学习模型在医学电子病历的实体抽取的效果,实验结果表明 BiLSTM-CRF 对提高结果的准确率是有效的.

知识学习和深度学习的方法大多需要搜集大量语料,或过多依赖于专家的标注.文献[34]提出利用已标注的实体三元组在自然语言表述上的共性和差异,对多种医疗实体关系类内的数据分布进行联合编码,进而从生成模型的角度去发现未被标注的关系实体三元组.该方法减轻了传统判别模型对于外部资源的过度依赖,并且不依赖于医疗实体关系之间的差异进行建模.实验表明:算法不仅能够在外部资源有限的条件下,以 92.91%的支持度生成属于某个特定医疗关系的实体三元组,其生成的结果拥有 77.17%的准确率且生成结果中有 61.93%的样本未曾出现在训练数据中.

2) 关系抽取

医学实体关系抽取的目标是解决实体间语义链接的问题,早期的关系抽取主要是通过人工构造语义规则以及模板的方法识别实体关系.之后,医学实体间的关系模型逐渐替代了人工预定义的语法与规则.本文按照医学实体的类型将医学实体关系归结为 2 类:同类型医学实体层关系抽取以及不同类型医学实体关系抽取.

同类型医学实体层关系比较简单,主要为 is-a 和 part-of 关系,此类关系通常在医学词典、百科、信息标准中定义.在实际应用中,可通过网络爬虫、正则表达式等技术从标准医疗数据库中抽取分层结构,ICD-10,SNOMED 等医疗数据库的医学专业分类和标准化工作比较权威且覆盖范围广,被广泛使用.非同类型医学实体关系的抽取方法是先定义好 2 实体间要抽取的关系类型,再将抽取任务转换为分类问题进行处理.

在知识图谱的构建过程中,远程监督(distant supervision)能够减少对标注数据的需求,因此被大量应用于从非结构化医学文本中进行关系抽取.文献[35]首先证明由于医学知识库的不完整,大量标记过程产生的否定标签为假否定,并基于此提出一种仅从实体对正标签进行学习的远程监督提取算法,并通过实验证明了此算法的有效性.文献[36]提出一种基于远程监督的卷积神经网络模型,利用卷积神经网络抓取实体的描述特征,丰富实体表示,并通过计算实体间关系与句子间的相似度赋予句子不同的权重.

然而,远程监督算法虽然从一定程度上减少了模型对人工标注数据的依赖,但该类方法也存在明显的缺点.其一是此种假设会引入大量噪音,其二是此算法的数据构造过程依赖于自然语言处理工具,中间过程出错会造成错误传播问题.为解决这个问题,文献[37]提出了一种协同消噪的模型,该方法由 2 个神经网络和一个协同模块组成,充分利用了医疗领域中丰富的医疗文本、医疗影像等信息数据.模型中的 2 个神经网络分别在文本语料库和知识图谱领域进行学习,再通过一个自适应的双向协同模块完成它们间的相互学习,达到消除噪声的目的.实验表明:该方法在噪声较大的数据上有较明显的效果提升.

3) 属性抽取

属性抽取针对医学实体而言,如药品的熟悉包括规格、剂量、适应症等,通过属性可以对实体进行完整勾画,如二甲双胍是二型糖尿病患者适用.由于实体的属性可以看成实体与属性值之间的一种名称性关系,因此可以将属性抽取问题转换为关系抽取问题.

2.3 医学知识融合

由于医学数据库中的知识来源复杂,存在知识质量良莠不齐、不同数据源知识重复、知识间关联关系模糊等问题[38],所以必须将来自不同数据源的多

源异构、语义多样、动态演化的医学知识在同一框架规范下进行异构数据的整合、消歧、加工、推理验证、更新等，对知识进行正确性判断，去粗取精，达到数据、信息、方法、经验与人思想的融合，将验证正确的知识通过对齐关联、合并计算有机地组织成知识库. 通过知识融合的定义可以看出，知识融合建立在知识抽取的基础上. 如何消除知识理解中的不确定性，发现知识的真值，并将正确的知识更新扩充到知识库中是知识融合研究中关注的重点[39]. 医学知识融合的关键技术有实体对齐技术、实体链接技术和关系推演技术. 其中，实体对齐技术用于消除本体和数据源的异构性；实体链接是医学知识融合的基础，通过消歧等操作消除知识中的不一致；关系推演用于发现隐含知识，从而扩充和补全医学知识库.

1) 实体对齐

实体对齐用于消除异构数据中的实体冲突、指向不明等不一致问题，从而从顶层创建一个大规模的统一知识库，从而帮助机器理解多源异质数据，形成高质量知识.

在医疗大数据的环境下，受医学知识库规模的影响，实体对齐会面临 3 个方面的挑战.

① 计算复杂度大. 算法计算复杂度会随知识库规模呈二次增长，计算复杂度难以接受.

② 数据质量良莠不齐. 由于不同医疗知识库的构建目的和方式不同，可能存在相似重复数据、孤立数据、数据时间力度不一致等问题.

③ 训练数据缺失. 大部分医疗数据库中并没有先验数据，通常需要研究者手工对数据进行标签等操作构造训练数据，这也是一项庞大的工作.

现有的实体对齐算法可分为成对实体对齐和集体实体对齐 2 类. 成对实体对齐方法只考虑实例及其属性相似度，常用方法包括概率统计模型、回归分类树模型、支持向量机分类模型、集成学习模型、层次图模型等. 集成实体对齐方法是在成对实体对齐的基础上，在计算实体相似度时加入了实体间相互关系，常用方法包括向量空间模型、bootstrapping 算法、贝叶斯网络模型、LDA 分配模型、Markov 逻辑网模型等.

2) 实体链接

实体链接的主要作用是利用医学知识库中的实体对从医疗大数据的文本中获取的实体指代进行消歧，识别每一个实体指代在医学知识库中与其对应的映射实体. 这里的实体指代指的是实体的一种文本表示形式[40]，一个医学实体可能有多种不同的表达，如全名、别名、缩写等. 按照实体链接利用的信息不同，现有工作主要分为基于实体属性(entity attributes based, EA)的实体链接方法[41]、基于实体流行度(entity popularity based, EP)的实体链接方法[42]、基于上下文(context based, CB)的实体链接方法[43]和基于外部证据(external evidencebased, EEB)的实体链接方法[44].

基于实体属性的实体链接方法通过计算实体的名字属性中字符串的相似度来判断实体是否相同. 实体名称和属性的相似度主要通过 Consine 距离、Jaccard 相关系数等方式进行计算：

$$Sim_{\text{Cosine}}(e_1,e_2)=\frac{|A(e_1)\bigcap A(e_2)|}{\sqrt{|A(e_1)||A(e_2)|}}, \tag{6}$$

$$Sim_{\text{Jaccard}}(e_1,e_2)=\left|\frac{A(e_1)\bigcap A(e_2)}{A(e_1)\bigcup A(e_2)}\right|, \tag{7}$$

其中，同 e_1 和 e_2 为给定的医学实体，$A(e)$ 表示医学实体 e 的属性字符串.

基于实体流行度的实体链接方法认为，对于给定的实体指代，与其对应的映射实体最有可能是医学数据库中最为公认的实体，计算为

$$P(e)=\frac{\#(e\text{ 出现的文本})}{\#(\text{医学数据库中的文本})}, \tag{8}$$

其中，e 表示给定的医学实体，$P(e)$ 表示医学实体 e 的流行度，$\#()$ 表示次数.

基于上下文的实体链接方法通过计算给定医学实体的上下文之间的相似性判断 2 个实体之间是否为同一实体.

基于外部证据的实体链接方法认为同一文本中的医学实体并不是独立的，它们之间存在语义相关性，而这种相关性有助于提升实体链接的准确率.

表 3 对比了以上 4 种实体链接方法的特点.

3) 关系推演

通过实体对齐和实体链接，可以得到初步的本体雏形，但构建知识库时需求和设计理念的不同会导致知识库中数据的多样性和异构性，因此要形成高质量的医学知识，还需要不断进行关系推演，将动态产生的关系扩展到已有的医学知识库中，从层次上形成一个大规模的医学知识体系，统一对知识进行管理，对提高医学知识库的时新性、覆盖能力至关重要. 由于医学自然语言表达的随意性，关系存在大量同义或多义表达，这给关系的扩充带来了巨大的挑战.

关系推演的主要目标是将从医疗大数据文本中获取的实体关系动态扩展到知识库中. 医学实体关系存在 2 种可能情况：1) 医学知识库中存在与目标

文本实体关系相同或等价的实体关系,只需找到文本实体关系在医学知识库中与之对应的实体关系;2)医学知识库中不存在与目标文本实体关系相同或等价的实体关系,则需要将实体关系扩展合并到知识库中,完成医学文本实体关系和医学知识库实体关系的关联合并.

Table 3 Advantages and Disadvantages of Entity Linking Model

表 3 实体链接方法分类汇总

Model	Rule	Advantages	Disadvantages
EA	Compute the similarity between the attributes.	Simple realization and the accuracy rate is high when the medical attribute is rich.	Poor anti-noise ability, the accuracy cannot be guaranteed when the entity attribute is sparse.
EP	Based on probability statistics and the frequency in medical encyclopedia.	Reliable and simple heuristic rules	Poor robustness, the ambiguity of specific entities is not considered.
CB	Based on the similarity between the entity context	High accuracy when the text is long enough and relatively clean	Less flexibility, the accuracy rate cannot be guaranteed when the text is sparse.
EEB	Based on semantic correlation	Strong expansibility, rich feature information is introduced	Method effectiveness depends on the quality of external evidence.

关系推演的关键在于判定 2 个实体关系是否表示同一种关系.目前有 2 种方法:①传统的基于语义的方法,通过对比描述关系的词汇之间语义相似度来验证是否是同一种关系;②基于嵌入学习的方法,这种方法通过在嵌入空间中寻找一个恰当的能量函数学习实体的嵌入表示,利用实体的嵌入表示表达实体关系,并判断 2 个描述实体的关系是否表达同一种关系,从而实现将实体关系进行结构映射.

目前医疗领域的知识融合技术虽有一些发展,但仍需要大量人工干预,高效的知识融合算法仍然有待研究.

2.4 医学知识推理

知识推理是在已有医学知识库的基础上进一步挖掘隐含知识,从而丰富、扩展知识库.在医学知识图谱中,知识推理能够帮助医生完成患者数据搜集、疾病诊断、治疗方法、避免医疗差错等.然而,医学领域拥有其特殊性,即使对于相同疾病,医生也会根据患者个体情况做出不同的诊断,所以医学知识图谱必须处理大量相同或矛盾的信息,大大增加了构建医学知识推理模型的复杂性.

传统的知识推理方法包括基于描述逻辑推理(description logic reasoning)[45]、基于规则推理(rule-based reasoning)[46]与基于案例推理(case-based reasoning)[47]等.传统知识推理方法虽然在一定程度上推动了医学知识图谱的发展,但是也存在准确率不高、数据利用率低、学习能力不足等缺陷,并未达到实际应用的要求.

随着医疗大数据规模的飞速增长,传统知识推理方法会出现信息遗漏、诊断时间延长等问题.而人工智能技术对于从海量医疗数据中挖掘有用信息有着天然优势,可以提升知识推理的效率和准确度,常用模型包括了人工神经网络模型(artificial neural networks model)[48]、遗传算法(genetic algorithm)[49]和反向传播网络模型(back propagation)等.

无论是传统知识推理方法还是人工智能只是推理方法都将知识图谱作为数据源,而基于图的推理则将知识图谱视作图,将医学实体看作节点,实体间关系看作边,利用关系路径中的蕴含信息,通过图中 2 个实体间的多步路径对其语义关系进行分析.常用算法包含了路径约束随机游走算法(path-constraint random walk)、路径分级算法(path ranking)等.

3 医学知识图谱应用

知识图谱为医疗信息系统中海量、异构、动态的医疗大数据的表达、组织、管理及利用提供了一种更为有效的方式,使系统的智能化水平更高,更加接近于人类的认知思维.目前医学知识图谱技术主要用于临床决策支持系统、医疗智能语义搜索引擎、医疗问答系统、慢病管理系统等.

3.1 临床决策支持

利用知识图谱技术可以辅助医疗行业和领域的大数据分析与决策,根据患者症状、检验、检查等数据,自动生成诊断、治疗方案,还可以对医生的诊疗方案进行智能化分析,有效减少误诊情况的发生.

IBM Watson 主要面向肿瘤和癌症领域的决策支持,基于巨大的知识库并使用自然语言、假设生成和基于证据的学习能力为临床决策支持系统提供帮助,供医学专业人员参考.此外,很多研究者针对这一领域进行了深入的研究.文献[50]提出一种本体驱动的、针对传染病诊断和抗生素处方的临床决策支持系统,系统包括一个医学本体知识库,其中综合

了多个医学本体资源,包括传染病、综合征、细菌、药物等相关本体;文献[51]提出一种面向重症监护室的急性心肌梗死患者的智能监测和决策支持系统.该系统的知识库由 OWL 本体和 1 组表示专家知识的规则组成,能够分析患者的情况,并给出治疗建议;文献[52]通过自然语言处理方法建立 3 层疾病结构知识图谱(疾病-症候-特征),运用正则表达式、隐 Markov 模型等人工智能技术解决了构建医学知识图谱过程中效率低、耗时长等问题.

3.2 医疗智能语义检索引擎

在大量医学数据中搜索生物医学信息是 1 项复杂的任务,医疗信息语义搜索建立大规模医学知识库对用户搜索的关键字和文档内容进行语义标注,从医学知识图谱中检索并查询相关的实体对、实体关系及属性进行扩展查询,从而改善医疗信息搜索结果;文献[53]利用医学主题词表 Mesh 对医学术语的用户检索进行扩展,改进了多模块医学信息检索系统.在 ImageCLEFmed 医学图像数据库对方法进行了验证,实验结果表明,使用医学本体扩展技术可以改善查询结果的准确性;文献[54]提出了一种基于概念匹配而非关键字匹配的电子病历检索方法,将电子病历文本从基于术语的原始文本转换为 SNOMED-CT 本体定义的医学概念,结果显示这种方法能够提升搜索精度并且为实现基于推理的医疗数据搜索系统提供了框架;文献[55]提出了一种支持多模式医学案例检索的医学信息检索系统.该系统通过提供多模态搜索、新的数据融合算法和医学同义词典的术语建议来支持医学信息检索和发现.

目前,国内外的医疗信息语义搜索引擎包括 Healthline,Google health、搜狗明医、360 良医等.其中 Google 率先提出将知识图谱应用于搜索引擎,Google 提供了超过 400 种健康状况数据,通过信息卡片的方式对疾病特征进行展示,告知用户某疾病是否具有传染性、影响主要人群等信息.Healthline 是一个基于医学知识库的医学信息搜索引擎,其知识库涵盖超过 800 000 项医疗元数据和 50 000 条相互关联的关系概念.搜狗明医和 360 良医结合了元搜索索引和知识库索引,收集权威医疗知识学术网站内容,为用户提供包括维基百科、知乎问答、国际前沿学术论文等权威的医学知识.

3.3 医疗问答

医疗问答系统是医疗信息检索系统的一种高级形式,能够以准确简洁的自然语言形式为用户提供问题的解答.多数基于知识图谱的医疗问答系统将给定的问题分解为多个小的问题,然后逐一去知识库抽取匹配的答案,并自动检测答案在时间和空间上的吻合度等,最后将答案进行合并,以直观的方式展示给用户.IBM 的 Walson、微软的“小冰”都是融合知识图谱的问答系统的代表产品.在医学领域,受限于现有医学知识图谱的推理能力,市场上尚未出现比较成熟的医疗问答系统.

研究人员针对知识图谱与医疗问答系统的融合开展了许多相关研究.文献[56]提出了一种基于自然语言处理的医疗问答系统,首先对比了先前 4 种基于医学本体的医疗问答系统,并结合医学领域知识、自然语言处理相关技术和语义关系构建知识图谱,实现了医疗问题的自动化回答;文献[57]构建了包括疾病库、症状库、中草药库等的中医药知识图谱,利用文本抽取、关系数据转换以及数据融合等技术,探索中医药知识图谱自动化构建方法与标准化流程,实现中医药知识图谱的智能应用,包括基于模板的中医药知识问答和基于知识图谱推理的辅助用药;针对已有工作在除关注上下文之外,对起着重要作用的背景知识关注很少的问题,文献[58]提出了一种具有知识感知能力的双向长短记忆模型,它利用医学知识图谱引入的背景知识来丰富问答的表征学习.模型的核心是一个上下文引导的注意力神经网络,通过将知识图谱中的背景知识嵌入整合到句子表示中,并结合知识型注意力机制模块,对问题和答案中的各个部分进行有效的相互关联.通过实验验证了该方法在 WikiQA 和 TREC QA 数据集上的效果,实验结果证明:该方法对于医疗问答准确性的提升具有一定有效性.

4 讨 论

知识图谱技术是对语义网技术的一次改造和升华.自 Google 提出知识图谱这一概念至今,其热度仍然只增不减.通过对知识图谱构建技术体系进行深入观察和分析,可以看出它是建立在多学科领域研究成果的基础之上的一门实用技术,是人工智能、信息检索、自然语言处理、万维网等交叉领域的理论研究热点和应用技术的集合.就医学领域而言,由于医学知识和规则的专业性、规范性、术语有限性等特点,可以从标准化医学词典、医学数据库等来源中获取高质量数据并构建医学知识图谱.虽然目前有很多项目进行医学知识图谱技术的研究,但医学知识图谱构建的关键环节还面临着一些巨大的困难和挑战.

1) 文本抽取难度大.在医学知识抽取环节,面向开放域的知识抽取方法研究还处于起步阶段.部

分研究成果虽然在特定数据集上取得了较好的结果,但普遍存在算法准确性低、限制条件多、扩展性不好等问题.尤其是医学电子病历抽取过程中涉及到的纯文本信息抽取是当前面临的重要挑战.

2)实体对应不准确.医学知识融合阶段的主要挑战就是实现准确的实体链接.虽然关于实体消歧、共指消解技术的研究已经有很长时间的历史,但是由于医学知识来源的多样性导致医学实体在不同的数据源中存在严重的多源指代问题,而迄今为止所取得的研究成果距离医学领域的实际应用还有很大的距离.如何在上下文信息受限(跨语境、跨文本等)的条件下准确地将文本中抽取的实体正确连接到医学知识库中是当前学术界普遍关注的问题.

3)知识图谱存储方式.目前医学知识图谱主要采用图数据库进行存储,在受益于图数据库带来的查询效率的同时,也会失去关系型数据库的优点,例如图数据库不能支持 SQL 语言查询、查询效率较低等.将自然语言的查询语句翻译为知识图谱可以理解的查询表达式及等价表达式也是医学知识图谱应用需要解决的关键问题.

5 结束语

在医疗领域中,随着医学信息化水平的逐步深入,积累了大量医学数据,医疗数据的有效使用对精准医疗、疾病防控、研发新药、医疗费用控制、攻克顽疾、健康管理等工作都有着重要的意义.构建医疗领域的知识图谱提供了一种从海量医学文本和图像中抽取结构化知识的手段,具有广阔的应用前景.本文从医学知识图谱构建的视角出发,对医学知识图谱的架构、医学知识图谱构建关键技术以及研究应用发展现状进行了全面调研和深入分析,并对医学知识图谱构建工作所面临的重要挑战和关键问题进行了总结.

知识图谱在医疗领域的意义不仅在于它是一个全局医学知识库,也是支撑例如辅助诊疗、智能搜索等医疗智能应用的基础,而且在于它是一把打开人类知识宝库的钥匙,它能够推进医学数据自动化和智能化处理,为医疗行业带来新的发展契机.

参 考 文 献

[1] Chen Min, Mao Shiwen, Liu Yunhao. Big data: A survey [J]. Mobile Networks & Applications, 2014, 19(2): 171-209

[2] Mervis J. Agencies rally to tackle big data [J]. Science, 2012, 336(6077): 22-29

[3] Bello-Orgaz G, Jung J J, Camacho D. Social big data: Recent achievements and new challenges [J]. Information Fusion, 2016, 28: 45-59

[4] Mayerschönberger V, Cukier K. Big data: A revolution that will transform how we live, work, and think [J]. Mathematics & Computer Education, 2014, 47(17): 181-183

[5] Xu Zenglin, Sheng Yongpan, He Lirong, et al. Review on knowledge graph techniques [J]. Journal of University of Electronic Science and Technology of China, 2016, 45(4): 589-606 (in Chinese)
(徐增林, 盛泳潘, 贺丽荣, 等. 知识图谱技术综述[J]. 电子科技大学学报, 2016, 45(4): 589-606)

[6] Pujara J, Miao H, Getoor L, et al. Knowledge Graph Identification [M]. Berlin: Springer, 2013

[7] Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods [J]. Semantic Web, 2017, 8(3): 489-508

[8] Yuan Kaiqi, Deng Yang, Chen Daoyuan, et al. Construction techniques and research development of medical knowledge graph [J]. Application Research of Computers, 2018, 8(7): 1929-1936 (in Chinese)
(袁凯琦, 邓扬, 陈道源, 等. 医学知识图谱构建技术与研究进展[J]. 计算机应用研究, 2018, 8(7): 1929-1936)

[9] Ruan Tong, Sun Chenglin, Wang Haofen, et al. Construction of traditional Chinese medicine knowledge graph and its application [J]. Journal of Medical Informatics, 2016, 37(4): 8-13 (in Chinese)
(阮彤, 孙程琳, 王昊奋, 等. 中医药知识图谱构建与应用[J]. 医学信息学杂志, 2016, 37(4): 8-13)

[10] Murdoch T B, Detsky A S. The inevitable application of big data to health care [J]. The Journal of the American Medical Association, 2013, 309(13): 1351-1352

[11] Wikipedia. Knowledge graph [OL]. [2016-05-09]. https://en.wikipedia.org/wiki/Knowledge_Graph

[12] Nickel M, Murphy K, Tresp V, et al. A review of relational machine learning for knowledge graphs [J]. Proceedings of the IEEE, 2015, 104(1): 11-33

[13] Wang Quan, Mao Zhendong, Wang Bin, et al. Knowledge graph embedding: A survey of approaches and applications [J]. IEEE Transactions on Knowledge & Data Engineering, 2017, 29(12): 2724-2743

[14] Wang Zhen, Zhang Jianwen, Feng Jianlin, et al. Knowledge graph and text jointly embedding [C] //Proc of Conf on Empirical Methods in Natural Language Processing. Doha: Association for Computational Linguistics, 2014: 1591-1601

[15] Gui Qilin, Gao Huan, Wu Tianxing. The research advances of knowledge graph [J]. Technology Intelligence Engineering, 2017, 3(1): 4-25 (in Chinese)
(漆桂林, 高桓, 吴天星. 知识图谱研究进展[J]. 情报工程, 2017, 3(1): 4-25)

[16] Maclennan A. The artificial life route to artificial intelligence: Building embodied, situated agents [J]. Journal of the American Society for Information Science & Technology, 2010, 47(6): 482-483

[17] Martinez-Gil J. Automated knowledge base management: A survey [J]. Computer Science Review, 2015, 18: 1-9

[18] Ramesh A N, Kambhampati C, Monson J R, et al. Artificial intelligence in medicine [J]. Artificial Intelligence in Medicine, 2004, 7(5): 334-338

[19] Yang Xiaohui, Wan Rui, Zhang Haibin, et al. Semanticalsymbol mapping embedding learning algorithm for knowledge graph [J]. Journal of Computer Research and Development, 2018, 55(8): 1773-1784 (in Chinese)
(**杨晓慧, 万睿, 张海滨, 等. 基于符号语义映射的知识图谱表示学习算法**[J]. **计算机研究与发展**, 2018, 55(8): 1773-1784)

[20] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2013: 2787-2795

[21] Socher R, Chen D, Manning C D, et al. Reasoning with neural tensor networks for knowledge base completion [C] //Proc of Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2013: 926-934

[22] Jenatton R, Roux N L, Bordes A, et al. A latent factor model for highly multi-relational data [C] //Proc of Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2012: 3167-3175

[23] Sutskever I, Salakhutdinov R, Tenenbaum J B. Modelling relational data using bayesian clustered tensor factorization [J]. Advances in Neural Information Processing Systems, 2009, 22(3): 1821-1828

[24] Nickel M, Tresp V, Kriegel H P. A three-way model for collective learning on multi-relational data [C] //Proc of the 28th Int Conf on Machine Learning. New York: ACM, 2011: 809-816

[25] Fang Yang, Zhao Xiang, Tan Zhen, et al. A revised translation-based method for knowledge graph representation [J]. Journal of Computer Research and Development, 2018, 55(1): 139-150 (in Chinese)
(**方阳, 赵翔, 谭真, 等. 一种改进的基于翻译的知识图谱表示方法**[J]. **计算机研究与发展**, 2018, 55(1): 139-150)

[26] Bordes A, Usunier N, Garcia-Duran A, et al. Translating embeddings for modeling multi-relational data [C] //Proc of Int Conf on Neural Information Processing Systems. New York: Curran Associates Inc, 2013: 2787-2795

[27] Saxena S, Saraceno B. The ICD-10 Classification of Mental and Behavioural Disorders [M]. Geneva: World Health Organization, 1993

[28] Elkin P L, Brown S H, Husser C S, et al. Evaluation of the content coverage of SNOMED-CT: Ability of SNOMED clinical terms to represent clinical problem lists [J]. Mayo Clinic Proceedings, 2006, 81(6): 741-748

[29] Wu S T, Liu Hongfa, Li Dingcheng, et al. Unified medical language system term occurrences in clinical notes: A large-scale corpus analysis [J]. Journal of the American Medical Informatics Association Jamia, 2012, 19(1): 149-156

[30] Kazama J, Makino T, Ohta Y, et al. Tuning support vector machines for biomedical named entity recognition [C] //Proc of the Workshop on Natural Language Processing in the Biomedical Domain. New York: ACM, 2001, 5(2): 1-8

[31] Lin Yifeng, Tsai T H, Chou W C, et al. A maximum entropy approach to biomedical named entity recognition [C] //Proc of Int Conf on Data Mining in Bioinformatics. Berlin: Springer, 2004: 56-61

[32] Lecun Y, Bengio Y, Hinton G. Deep learning [J]. Nature, 2015, 521(7553): 436

[33] Jagannatha A N, Yu Hong. Structured prediction models for RNN based sequence labeling in clinical text [C] //Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. New York: ACL 2016: 856-865

[34] Zhang Chenwei, Li Yaliang, Du Nan, et al. On the generative discovery of structured medical knowledge [C] //Proc of the ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 23-37

[35] Surdeanu M, Tibshirani J, Nallapati R, et al. Multi-instance multi-label learning for relation extraction [C] //Proc of Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. New York: ACM, 2012: 455-465

[36] Lin Yankai, Liu Zhiyuan, Sun Maosong. Neural relation extraction with multi-lingual attention [C] //Proc of Meeting of the Association for Computational Linguistics. New York: ACL, 2017: 34-43

[37] Lei Kai, Chen Daoyuan, Li Yaliang, et al. Cooperative denoising for distantly supervised relation extraction [C] //Proc of the 27th Int Conf on Computational Linguistics. New Mexico: Santa Fe Community Convention Center. 2018: 20-26

[38] Lin Hailun, Wang Yuanzhuo, Jia Yantao, et al. Network big data oriented knowledge fusion methods: A survey [J]. Chinese Journal of Computers, 2017, 23(1): 1-27 (in Chinese)
(**林海伦, 王元卓, 贾岩涛, 等. 面向网络大数据的知识融合方法综述**[J]. **计算机学报**, 2017, 23(1): 1-27)

[39] 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

[40] Bilenko M, Mooney R J. Adaptive duplicate detection using learnable string similarity measures [C] //Proc of ACM Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2003: 39-48

[41] Chen R-C, Bau C T, Yeh C J. Merging domain ontologies based on the WordNet system and fuzzy formal concept analysis techniques [J]. Applied Soft Computing, 2011, 11(2): 1908-1923

[42] Li Yang, Wang Chi, Han Fangqiu, et al. Mining evidences for named entity disambiguation [C] //Proc of ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2013: 1070-1078

[43] Bhattacharya I, Getoor L. Collective entity resolution in relational data [J]. ACM Transactions on Knowledge Discovery from Data, 2007, 1(1): 299-304

[44] Cucerzan S. Large-scale named entity disambiguation based on wikipedia data [C] //Proc of the 2007 Joint Conf on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. New York: ACL, 2007: 708-716

[45] Giacomo G D, Lenzerini M. TBox and ABoxreasoning in expressive description logics [C] //Proc of the 1996 Int Workshop on Description Logics. New York: ACL, 1996: 37-48

[46] Buchanan B G, Shortliffe E H. Rule-Based Expert Systems: The MYCIN Experiments of The Stanford Heuristic Programming Project [M]. Reading, MA: Addison-Wesley, 1984: 67

[47] Bousquet C, Henegar C, Louët A L, et al. Implementation of automated signal generation in pharmacovigilance using a knowledge-based approach [J]. International Journal of Medical Informatics, 2005, 74(7): 563-571

[48] Khan J, Wei J S, Ringnér M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. [J]. Nature Medicine, 2001, 7(6): 673-679

[49] Goldberg D E. Genetic algorithm in search optimization and machine learning [J]. Reading, MA: Addison Wesley, 1989, 8(7): 2104-2116

[50] ángelGarcía-Crespo, Rodríguez A, Mencke M, et al. ODDIN: Ontology-driven differential diagnosis based on logical inference and probabilistic refinements [J]. Expert Systems with Applications, 2010, 37(3): 2621-2628

[51] Martínezromero M, Vázqueznaya J M, Pereira J, et al. The iOSC3 system: Using ontologies and SWRL rules for intelligent supervision and care of patients with acute cardiac disorders [J]. Computational and Mathematical Methods in Medicine, 2013, 2013(5904): 650-671

[52] Nie Lili, Li Chuanfu, Xu Xiaoqian, et al. Study on application of artificial intelligence in the building of medical diagnosis knowledge-graph [J]. Journal of Medical Informatics, 2018, 5(6): 7-12 (in Chinese)
(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, 5(6): 7-12)

[53] Díaz-Galiano M C, Martín-Valdivia M T, Ureña-López L A. Query expansion with a medical ontology to improve a multimodal information retrieval system [J]. Computers in Biology & Medicine, 2009, 39(4): 396-403

[54] Koopman B, Bruza P, Sitbon L, et al. Towards semantic search and inference in electronic medical records: An approach using concept—based information retrieval [J]. Australasian Medical Journal, 2012, 5(9): 482-488

[55] Mourão A, Martins F, Magalhães J. Multimodal medical information retrieval with unsupervised rank fusion [J]. Computerized Medical Imaging & Graphics, 2015, 39: 35-45

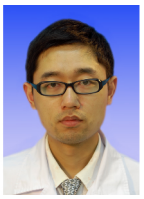
[56] Abacha A B, Zweigenbaum P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies [J]. Information Processing & Management, 2015, 51(5): 570-594

[57] Ruan Tong, Sun Chenglin, Wang Haofen, et al. Construction of traditional Chinese medicine knowledge graph and its application [J]. Journal of Medical Informatics, 2016, 37(4): 8-13 (in Chinese)
(阮彤, 孙程琳, 王昊奋, 等. 中医药知识图谱构建与应用[J]. 医学信息学杂志, 2016, 37(4): 8-13)

[58] Shen Ying, Deng Yang, Yang Min, et al. Knowledge-aware attentive neural network for ranking question answer pairs [C] //Proc of the 41st Int ACM SIGIR Conf on Research & Development in Information Retrieval. New York: ACM, 2018: 901-904

Hou Mengwei, born in 1988. Master. Her main research interests include information retrieval, machine learning in medical domain.

Wei Rong, born in 1980. Master. Deputy director in the Information Technology Department of the First Affiliated Hospital of Xi'an Jiaotong University. His main research interests include medical information technology and big data application.



Lu Liang, born in 1982. PhD. His main research interests include data mining and database administration.

Lan Xin, born in 1983. Master. Her main research interests include data mining, network behavior and machine learning.

Cai Hongwei, born in 1973. PhD. Associate chief technician. His main research interests include interactive Web response random allocation system, non-randomized clinical data analysis, clinical data standar-dization and clinical data mining.