软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(10):3167−3183 [doi: 10.13328/j.cnki.jos.005825]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于领域语义知识库的疾病辅助诊断方法*

陈德彦[1,2,4], 赵 宏[1,2,4], 张 霞[1,2,3,4]

[1](东北大学 计算机科学与工程学院,辽宁 沈阳 110169)

[2](计算机软件国家工程研究中心(东北大学),辽宁 沈阳 110179)

[3](沈阳东软智能医疗科技研究院有限公司,辽宁 沈阳 110179)

[4](东软集团股份有限公司,辽宁 沈阳 110179)

通讯作者: 陈德彦, E-mail: deyan_chen@126.com

摘 要: 健康医疗领域是一个知识密集型的领域,临床诊断的质量主要依赖于医生所掌握的健康医疗知识以及临床经验.然而,单个医生的能力仍然非常有限,所以目前临床诊断的质量并不高.为此,提出一种基于领域语义知识库的疾病辅助诊断方法,基于 Freebase 中 medicine 主题域的知识建立了领域语义知识库,提出计算知识库中症状于疾病诊断的权重、计算与患者输入症状集相关的疾病的相关度和基于患者输入症状集推荐相关症状的算法.最后,基于随机选取的 6 种常见疾病的临床病历数据对所提出的方法与现有方法进行了对比评价,评价结果一方面表明了所提方法对已有方法存在的问题和不足的改进效果,另一方面也表明所提方法可以避免"冷启动"问题,可以快速支撑对大量常见疾病的辅助诊断.基于所提方法,有望为基层全科医生提供大量常见疾病的辅助诊断服务,或者为患者提供疾病自诊服务.

关键词: 本体;领域语义知识库;疾病辅助诊断;症状权重;疾病相关度;相关症状

中图法分类号: TP182

中文引用格式: 陈德彦,赵宏,张霞.基于领域语义知识库的疾病辅助诊断方法.软件学报,2020,31(10):3167−3183. http://www.jos.org.cn/1000-9825/5825.htm

英文引用格式: Chen DY, Zhao H, Zhang X. Aided diagnosis method for diseases based on the domain semantic knowledge base. Ruan Jian Xue Bao/Journal of Software, 2020,31(10):3167−3183 (in Chinese). http://www.jos.org.cn/1000-9825/5825.htm

Aided Diagnosis Method for Diseases Based on the Domain Semantic Knowledge Base

CHEN De-Yan[1,2,4], ZHAO Hong[1,2,4], ZHANG Xia[1,2,3,4]

[1](School of Computer Science and Engineering, Northeastern University, Shenyang 110169, China)

[2](National Engineering Research Center for Computer Software (Northeastern University), Shenyang 110179, China)

[3](Neusoft Research of Intelligent Healthcare Technology Co. Ltd., Shenyang 110179, China)

[4](Neusoft Corporation, Shenyang 110179, China)

Abstract: The health care domain is a knowledge-intensive domain. The quality of clinical diagnosis depends mainly on the knowledge of health care and clinical experience held by doctors. However, the ability of a single doctor is very limited, so the quality of clinical diagnosis is not high. To this end, this study proposes an aided diagnosis method based on the domain semantic knowledge base. Firstly, based on the knowledge of the medicine subject matter domain in Freebase, a domain semantic knowledge base is established. Then, based on the semantic knowledge base, the algorithms for calculating the weights of the symptoms in the knowledge base, the relevancy

* 基金项目: 国家自然科学基金(61232015); 国家高技术研究发展计划(2015AA020103); 国家重点研发计划(2016YFC1303000); 沈阳东软智能医疗科技研究院有限公司开放课题基金(NRIHTOP1802)

Foundation item: National Natural Science Foundation of China (61232015); National High-Tech Research and Development Plan of China (863) (2015AA020103); National Key Research and Development Program of China (2016YFC1303000); Open Program of Neusoft Research of Intelligent Healthcare Technology, Co. Ltd. (NRIHTOP1802)

收稿时间: 2018-01-25; 修改时间: 2018-08-12; 采用时间: 2019-02-21

of the diseases related to the input symptom set from a patient, and the related symptom set related to the input symptom set from the patient are proposed. Finally, based on the clinical data of 6 kinds of common diseases randomly selected, the method proposed in this study is compared with the existing methods. On the one hand, the evaluation results show that the method of this paper improves the problems and deficiencies of the existing methods. On the other hand, it shows that the method can avoid the “cold start” problem and can quickly support the aided diagnosis of a large number of common diseases. Using the method presented in this paper, it is expected to provide a comprehensive diagnosis service for a large number of common diseases for the general practitioners at the grassroots level, or provide patients with self-diagnosis services for diseases.

Key words: ontology; domain semantic knowledge base; aided diagnosis of disease; symptom weight; disease relevancy; related symptom

疾病诊断是医疗活动中重要的环节之一,它为患者的治疗和预后提供了一个坚实的基础[1].疾病诊断的质量主要依赖医生所掌握的医疗知识以及医疗经验,但单个医生所掌握的医疗知识以及积累的医疗经验仍然有限,如何提升医生(尤其是没有经验的医生)的临床诊疗水平并减轻医生的工作负荷,是一个亟待解决的问题.这方面的研究在早期主要有专家系统,专家系统的思想是,将专家的经验知识进行形式化,想以此代替专家进行诊断.从专家那里提取经验知识是一种劳动密集性的工作,由于专家的诊断往往具有“直觉性”,所以很多时候专家都无法提供这种具有直接因果关系的经验知识.

随着医学科学的发展和医院信息化水平的提高,临床上积累了大量的诊疗知识和电子化的病历数据,医生的诊疗经验也蕴藏在这些病历数据中.相应地,云计算、大数据、人工智能(artificial intelligence,简称 AI)等技术的出现和发展为这些数据的挖掘和利用提供了有力支撑.在这种有利条件下,基于数据挖掘和机器学习[2−5]算法的疾病辅助诊断和预测研究如雨后春笋般大量出现.但这些研究大多都针对某个单病种或者专科疾病,其得出的疾病辅助诊断模型无法为大量基层全科医生提供常见疾病的辅助诊断服务,也无法为患者提供大量常见疾病的自诊服务.目前出现的智能导诊服务机器人可以代替院内的导诊护士向患者提供更加高效和精准的分诊服务,但它们也依赖对大量常见疾病的辅助诊断服务.

本体(ontology)是共享概念模型的明确的、形式化的规范说明[6],其提供了一种结构化地表示领域知识的形式化方法,并提供了推理能力,构造本体可以实现某种程度的知识共享和重用.由于本体具有的强大的知识表示和推理能力,已经在很多领域得到了广泛的应用,例如语义 Web[7]、知识工程、自然语言处理、信息获取、信息集成、生物医学等领域,用于领域问题求解、异构信息源之间的交互、辅助组织中人与人之间的沟通等.在生物医学领域,已经出现了大量基于本体构建的知识库[8],例如基因本体[9]、人类表型本体[10]、疾病本体[11]等.相应地,在健康医疗领域也出现了大量基于本体知识库的疾病辅助诊断研究[1,12−14]和其他应用研究[15−17].基于本体知识库可以快速支撑大量常见疾病的辅助诊断.

疾病诊断是一个不断迭代的复杂过程,包括前瞻性诊断(prospective diagnoses)和回顾性诊断(retrospective diagnoses)[1].前瞻性诊断即诊断过程,在这个过程中医生不断收集关于患者的症状、检查结果和病史等详细信息,以缩小疾病诊断的范围.在这个过程的某个点上,医生可能积累了足够多的信息,这时可以给出一个或几个最可能的最终诊断.前瞻性诊断是一个基于收集到的患者信息的前向推理(forward reasoning)过程.在得出最终诊断以后,医生还需要通过回顾性诊断来验证最终诊断的正确性.一方面,验证最终诊断中疾病关联的体征、症状、异常指标等是否与医生收集到的患者的信息相一致;另一方面,最终诊断中的疾病可能还表现有其他一些医生未收集到的信息,医生需要基于此作进一步的收集和确认,这个过程是一个反向推理(backward reasoning)的过程.

在健康医疗领域的本体知识库中,围绕疾病建立了其与体征、症状、检查、病因、药物、手术等之间的静态关联关系.一种疾病可以表现出多种症状,相同的症状也可能出现在多种疾病中,不同的疾病可能表现出一个或者多个相同的症状.医生在基于收集到的患者信息对疾病进行筛查时,需要对筛查出的疾病进行可能性大小排名.常规的思路是,假定收集到患者的 5 个症状,本体知识库中某个疾病 d_1 匹配了其中的 4 个症状,而另一个疾病 d_2 匹配了其中的 3 个症状,那么可以认为患者患疾病 d_1 的概率比 d_2 要大,为此,在筛查出的疾病结果排序中,d_1 排在 d_2 的前面.以上思路的假设是,所有症状于疾病诊断的重要性是相同的,而实际情况并非如此.虽然由于存

在个体差异,相同疾病于不同的患者可能表现出不同的症状,但通常某种疾病于大多数人会表现出一些相同的典型症状,比如感冒的典型症状有咳嗽、流鼻涕、流眼泪、发烧、食欲不振等,糖尿病的典型症状有“三多一少(多饮、多食、多尿、体重减轻)”等.所以,虽然不同的疾病可能表现出一些相同的症状,但这些疾病表现出的典型症状不一定相同.或者说,相同症状于不同疾病的判断权重可能是不一样的.

针对以上问题,基于症状来诊断疾病的关键点是给出每一个症状于疾病诊断的重要性.对此,文献[1]进行了研究,提出这个重要性将基于症状被包含的疾病数量来给出.比如,肌无力(muscle weakness)是一个在许多疾病中都出现的症状,因此它于疾病诊断的贡献很小;而另一个症状,心动过缓(bradycardia)是另一小簇疾病特有的症状,如果患者提供的症状中包含了这个症状,那么该患者患有的疾病很有可能落在这一小簇疾病中.基于此推理,文献[1]提出了用于计算本体知识库中症状 s 的权重(weight)w_s 的算法,并基于 w_s 提出了计算与患者输入症状集 S 相关的疾病 d_i 的相关度(relevancy)w_i 的算法,然后基于 w_i 对筛查出的相关疾病进行排序.但文献[1]中计算 w_i 的算法存在以下 3 点明显问题.

(1) w_i 的取值大于 1,且最小值为 1.当疾病 d_i 关联了患者输入症状集 S 中的所有症状时,w_i 的取值为 1;否则,w_i 的值大于 1.由于疾病 d_i 与患者输入症状集 S 的相关度是一个概率,从概率的含义上讲,w_i 的取值不可能大于 1;如果 w_i 的取值为 1,表示必然事件,即疾病 d_i 一定与该患者相关.所以,文献[1]中计算 w_i 的算法存在明显的错误.

(2) 根据计算 w_i 的算法,当疾病 d_i 关联的患者输入症状集 S 中的症状的 $\sum w_s$ 越小时,w_i 的取值反而越大,这明显与上面文献[1]中的症状重要性推理结论相悖.

(3) 在计算 w_i 的算法中,分子的取值始终为 $\sum_{s\in S} w_s$,症状集 S 中与疾病 d_i 没有关联的症状于疾病 d_i 的相关度计算是没有作用的,而知识库中与疾病 d_i 关联的所有症状对疾病 d_i 的诊断都是有贡献作用的,这与静态知识库的假设基础有关,详见下面对文献[1]不足点的分析.

除此以外,文献[1]中计算 w_i 的算法还存在以下两点不足.

(1) 未考虑知识库中与疾病 d_i 关联的其他症状(不在患者输入症状集 S 中的症状)的影响作用.本体知识库中的每种疾病都关联了一定数量的症状,这种静态关联的假设基础是,疾病关联的所有症状共同作用于该种疾病.也即,如果一位患者表现出的症状集 S 完全覆盖了某种疾病关联的所有症状,没有多余症状,也没有缺失的症状,那么可以认为患者就是得了这种疾病.在这个假设条件下,如果两种疾病都包含了患者输入的所有症状或相同症状,并不能认为这两种疾病与患者的相关度是相同的,还需要考虑这两种疾病关联的其他症状的影响作用.

(2) 未对筛查出疾病集中的疾病进行回顾性验证,即未根据筛查出的疾病集合评估和推荐与患者输入症状集 S 中的症状最相关的其他症状,供医生或患者确认;进而基于初始收集到的患者症状集 S 和患者再次确认的症状结果对疾病筛查结果集进行调整,并重新计算调整后的疾病集中的疾病的相关度.

针对以上文献[1]中疾病辅助诊断算法存在的明显问题和不足,本文进行了改进和完善.本文的贡献如下.

(1) 提出了一种基于领域语义知识库的疾病辅助诊断方法,包括前瞻性诊断和回顾性诊断.该方法分别给出了计算领域语义知识库中症状 s 的权重 w_s 的算法、与收集到的患者症状集 S 中一个或者多个症状相关联的疾病 d_i 的相关度 w_i 的算法、与症状集 S 中的症状最相关的症状集 S_{rel} 的算法.

(2) 选取了 6 种常见疾病的临床病历数据对本文提出的方法进行了评价.对于每一份病历,从患者主诉和现病史中抽取症状信息作为患者输入的症状集.基于该症状集和本文提出的方法获得相关疾病列表及疾病相关度排名.从疾病相关度排名中,选取 Top-1(首诊断)和 Top-3(前 3 诊断)分别与病历数据中医生给出的诊断进行比较.同时,与文献[1]中的方法和基于统计的方法就诊断命中率进行了比较.

本文第 1 节给出领域语义知识库的相关定义.第 2 节介绍本文的疾病辅助诊断方法所依赖的领域语义知识库的构建方法.第 3 节给出本文的疾病辅助诊断方法.第 4 节对本文提出的疾病辅助诊断方法进行评价.最后对本文进行总结,并指出下一步的研究工作.

1 相关定义

定义 1(领域本体定义(domain ontology schema)). 领域本体定义通过捕捉某个领域中共同认可的概念、概念的属性、概念间的语义关系(包括分类关系和非分类关系)及相关语义约束来描述该领域的知识.记 O_{domain} 表示领域本体定义,其定义如下:

$$O_{domain}=\langle C,A,R,X,I\rangle.$$

C 表示概念(concept)集,概念又称为类(class),用于表达具有某类相似特征的个体(individual)的集合,个体又称为实例(instance).例如,概念 person 代表所有个体人的集合.A 表示所有概念的属性(attribute)集,概念的属性又称为数据类型属性(data type property),它描述概念所包含的实例本身的特征,例如个体“张三”的姓名、性别、身高、体重等属性.R 表示语义关系(semantic relation)集,语义关系又称为对象属性(object property),用于描述概念之间、数据类型属性之间和对象属性之间的分类关系(taxonomic relation)或者实例之间的非分类关系(non-taxonomic relation).例如,概念 person 可以进一步分为两个子概念 man 和 woman,个体“张三”和“李四”之间可能具有“好友”关系.X 表示公理集,公理(axiom)用于定义概念、属性和关系之上的语义约束.例如,约束概念 person 的生日属性只能有一个值,或者其在生物学意义上的父亲和母亲都具有唯一的值.I 表示实例数据集,用于描述领域中共同认可的常识知识.例如,描述“2 型糖尿病”的表现症状有“多饮”“多食”“多尿”“体重减轻”等.本体定义中一般不包含实例数据,除非用于表达一般性的领域常识知识,即本体定义中的实例描述不针对任何特定的应用场景,是领域内共同认可的知识.仅包含实例数据的 RDF(resource description framework)[18]描述不能称为本体定义[19].W3C 推荐的本体标准描述语言有 RDF、RDFS(RDF schema)[20]和 OWL(Web ontology language)[21].

定义 2(领域实例数据(domain instance data)). 领域实例数据为基于领域本体定义中的语义组件来描述的领域中的个体的知识.例如,可以定义一个 people 本来用于描述个体“张三”,或者定义一个疾病本体用于描述“2 型糖尿病”.记 I_{domain} 表示领域实例数据,其定义如下:

$$I_{domain}=\{(s,p,o)\mid s\in I,p\in A\cup R,o\in I\bigcup V\}.$$

(s,p,o)表示描述实例数据的陈述或称为三元组(triple),s 为某个实例对象,I 表示实例对象集,p 表示用于描述实例对象的属性或者语义关系,A 和 R 分别表示描述 I 中实例对象所用到的属性集和语义关系集,o 表示属性或语义关系的取值,或者为实例对象,或者为字面值(literals),V 表示字面值的集合.

定义 3(领域语义规则集(domain semantic rule set)). 领域语义规则同时服务于领域本体的定义和领域实例数据的描述.一方面,用于描述领域中领域专家所获得的启发式经验知识;另一方面,用于补充本体描述语言的语义描述能力.记 F_{domain} 表示领域语义规则集,其定义如下:

$$F_{domain}=\{r_1,r_2,...,r_i,...,r_n\},n\geqslant 0.$$

r_i 表示其中一条语义规则.语义规则是典型的条件语句:if-then 子句,只有当特定陈述(statement)集合为真时,才会添加新的知识.例如,使用语义规则描述疾病诊断的知识,当收集到的某位患者的信息满足某条疾病诊断规则的条件时,可以基于规则推理得出疾病诊断结果并建立该患者与该疾病的语义关系.

语义 Web 层次结构[22]提供了多种知识表示形式,包括从 RDF 到最新版本的 OWL 等多种格式,每一层都对表达能力进行了进一步的扩展,并且允许用户根据语义程序具体所需的语义量来采用相应的表示方式.但本体描述语言在表达能力和灵活性方面仍然存在一些不足,语义规则用于扩展本体描述语言的描述能力以及灵活性.W3C 建议的语义规则描述语言为 SWRL(semantic Web rule language)[23].

定义 4(领域语义知识库(domain semantic knowledge base)). 领域本体定义、领域实例数据和领域语义规则集一起构成了领域语义知识库.记 SKB_{domain} 表示领域语义知识库,其定义如下:

$$SKB_{domain}=\langle O_{domain},I_{domain},F_{domain}\rangle.$$

领域本体定义的结束,便是领域语义知识库构建的开始[24].基于领域本体定义和语义规则来描述具体的实例,形成实例数据和语义规则集,以领域本体定义作为领域背景知识,以领域实例数据和语义规则集作为具体的知识,它们一起形成了面向领域特定应用需求的语义知识库.例如,基于健康医疗领域本体定义构建面向慢性病患者的健康风险评估、疾病辅助诊断、疾病干预方案(药物、运动、饮食、心理、睡眠等)、远程监护服务、

健康知识问答服务、健康教育/咨询服务等需求的语义知识库.

W3C 推荐的针对语义知识库的语义层(即 RDF 层)查询标准语言为 SPARQL[25],这种查询语言不仅理解 RDF 的语法,而且理解 RDF 的数据模型和 RDF 词汇的语义,几乎所有的 RDF 查询工具都提供了对 SPARQL 查询语义的支持[22].

2 领域语义知识库的构建

由于领域知识的专业性,目前公认领域本体的开发需要领域专家的参与,并由知识工程师将领域专家提供的领域知识建模并形式化为可被计算机处理和共享利用的领域本体知识.但又由于领域知识体系的复杂性,完全由人工从头构建几乎是不可能的,并且在时间上也是不可接受的.所以本体工程[22]以及几乎所有本体建模方法[26−31]都强调在基于本体构建领域语义知识库之前,考虑集成和复用已经存在的领域本体知识库.例如,通过本体集成(ontology integration)[32]和本体映射(ontology mapping)[33,34]的方法来快速构建所需要的领域本体知识库;或者采用本体学习(ontology learning)[35]技术自动或半自动地从领域数据源中获取领域知识,并基于本体进行描述,领域数据源包括领域中的结构化、半结构化和非结构化数据;或者从其他开放语义知识库中抽取可复用的领域知识.例如,Freebase[36−38]、DBpedia[39]、YAGO[40]等.本文拟采用从开放语义知识库中抽取可复用的领域知识,考虑到本文的研究目的,需要获取关于疾病、症状及其语义关系相关的知识.为此,针对这部分知识的正确性和完整性,和医学专家一起,对 Freebase、DBpedia、YAGO 等开放语义知识库进行了对比分析,最终选择了 Freebase 作为抽取的数据源.

Freebase 是一个实用的、可伸缩的、图形化的、结构化的一般人类知识的数据库,用户可以在一个开放的平台上协作创建、结构化和维护其内容.Freebase 数据被表示为三元组(又称为事实)的格式,可以被可视化表示为一个有向图.Freebase 数据来自大量高质量的开放数据源,例如 Wikipedia[41]、MusicBrainz[42]、WordNet[43]等.Freebase 已经成为 LOD(linked open data)[44]项目的一个重要的数据源.Freebase 以每周为单位,将其数据在 CC-BY[45]许可下发布为 *N*-Triples[46]RDF 格式的 dump 文件.该文件采用 gzip 进行压缩,解压后为一个单一的文本文件.例如,在 2014 年 8 月下载的 gzip 压缩包的大小为 22GB,解压后大小为 250GB,共包含约 19 亿个三元组.Freebase RDF dump 包包含了 11 个实现域(implementation domain)、5 个 OWL 域和 89 个主题域(subject matter domain)[47].Freebase 的 medicine 主题域描述了健康医疗领域的本体定义和领域实例数据(即领域常识知识).例如,medicine 主题域的本体定义中描述了疾病、症状、病因、风险因素、药物等相关概念和属性,并基于该本体定义描述了健康医疗领域的一些常识知识,即由相关概念标注的实例数据及其语义关系.

本文的研究选择直接从 Freebase RDF dump 包中抽取 medicine 主题域的知识.在完全理解与 Freebase 相关的概念、Freebase 的知识表示模型和 Freebase RDF dump 包的结构特征的情况下,结合 Linux 下的命令行工具(例如 gzip、sed、grep、cat、wc、head、tail 等)、shell 脚本、Apache Jena[48]提供的相关工具和 SPARQL 查询,设计和实现了一种从 Freebase RDF dump 中快速、准确、完整地抽取某个或者某几个领域的知识的方法[49].该方法包括 6 个主要步骤.

1) 数据预处理.数据预处理的目的有两个:(a) 缩减数据包的规模,以有利于对其进行存储和处理.该方法采用 Linux 下的命令行工具和 shell 脚本对数据包进行预处理.一方面,将 *N*-Triples 格式转换为 Turtle[50]格式,这将使得数据包的规模减少约一半;另一方面,删除实现域中的事实(这些事实用于内部授权管理或者链接到外部资源,对于领域知识库用户来说并不需要或者并不关心),这可以使数据包的规模再减少约一半.(b) 删除和处理非法格式的三元组,这些非法格式的三元组会导致在将数据包加载到存储中时发生中断.

2) 数据装载.使用 Jena 提供的 TDB[51]作为 RDF Store(又称为 triple store),使用 Jena 提供的 tdbloader 命令行工具将预处理后的数据包加载到 TDB 中.

3) 存储发布.使用 Jena 提供的 Fuseki SPARQL Server[52]对 TDB 进行公开,以利用 Fuseki 提供的 SPARQL 查询服务端点(endpoint)实现对 TDB 存储的查询.

4) 知识抽取.通过构造 SPARQL 查询分别抽取 medicine 领域的本体定义 $O_{medicine}$(包括类定义和属性定义)

和实例数据 $I_{medicine}$.Freebase RDF dump 包中没有领域语义规则集的定义.

5) 数据后处理.包括两个方面,一方面,在 Freebase RDF dump 中,不区分数据类型属性和对象属性,为此需要对其进行处理;另一方面,在 Freebase RDF dump 中,所有主题域的名称空间相同,为此,需要结合领域需求对其进行替换.

6) 数据集成和处理.包括 3 个方面:(a) 将后处理后的 $O_{medicine}$ 和 $I_{medicine}$ 整合为 $SKB_{medicine}$;(b) 将采用 MID(machine identifier)标识的类和属性表示为人类可读的 ID,在 Freebase 中,所有类和属性既有唯一的 MID,也有唯一的人类可读的 ID,而所有实例只有唯一的 MID,其人类可识别的标识通过标注属性 rdfs:label 进行描述;(c) 将 $SKB_{medicine}$ 转换为本体标准语言描述的格式.Freebase RDF dump 包虽然采用 RDF 来进行描述,但其中的某些语义描述组件并未采用本体标准描述语言定义的语义组件,而是采用了 Freebase 实现域中的一些语义组件.例如,字面值的类型采用了 Freebase type 域中的类型定义,还有一些标注属性使用了 Freebase common 域中的属性定义等.所以还需要将抽取的结果转换为本体标准语言描述的形式,以利用通用的语义 Web 工具对其进行处理.

由于本文讨论的疾病辅助诊断只依赖 medicine 主题域中的疾病、症状实例及其语义关系,所以最后由医学专家对这部分知识内容进行了校对和进一步的完善.图 1 所示为使用本体编辑工具 Protégé[53]打开的、最终得到的 medicine 领域语义知识库 $SKB_{medicine}$,其规模说明如下.

1) 以 Turtle 格式表示的 medicine 领域语义知识库的文件(扩展名为.ttl)大小为 1.6GB,装载到 TDB 中以后,占用的文件系统存储空间大小为 1.3GB(比原始文件还要小).

2) 知识库中包含 70 个概念、63 个数据类型属性、156 个对象属性、886 272 个实例和 7 073 580 个三元组.

3) 疾病实例有 7 367 个(不含同义实例),其中,3 590 个疾病实例包含了合计 3 802 个同义实例,这些同义实例被归一化到(通过 owl:sameAs 语义组件)其对应的标准疾病实例(下文所指疾病实例均指标准疾病实例)中.

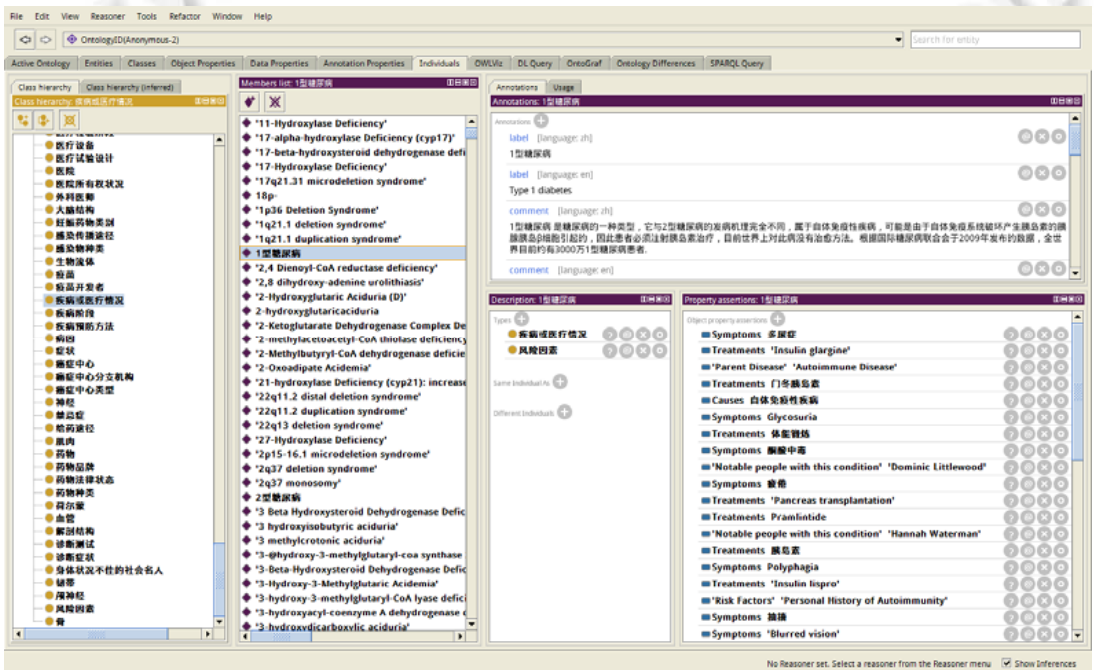

Fig.1 The domain semantic knowledge base in medicine

图 1 Medicine 领域的语义知识库

4) 症状实例有 1 444 个(不含同义实例),其中,1 112 个症状实例包含了合计 1 352 个同义症状,它们也被归

一化到其对应的标准症状实例(下文所指症状实例均指标准症状实例)中.

5) 包含从标准疾病实例指向标准症状实例的语义关系 6 028 个.

在 $SKB_{medicine}$ 中,疾病实例通过属性 med:medicine.disease.icd_9 和 med:medicine.disease.icd_10 分别指向了 ICD-9 和 ICD-10 国际疾病分类[54].同样,症状实例也通过属性 med:medicine.symptom.icd_9 和 med:medicine.symptom.icd_10 分别指向了 ICD-9 和 ICD-10 国际疾病分类.在 ICD-10 中,以 R 开头的编码是关于症状和体征的分类.这里,med 为 medicine 主题域的名称空间前缀.

在 Freebase 中,所有主题域中的概念只有顶层概念.例如,在 medicine 主题域中,疾病和症状只有一个顶层概念,疾病之间的分类关系通过属性 med:medicine.disease.parent_disease 和 med:medicine.disease.includes_diseases 来实现.即对于疾病来说,除了顶层概念,都是实例,实例间通过属性来表达分类关系.同样,症状实例之间的分类关系通过属性 med:medicine.symptom.parent_symptom 和 med:medicine.symptom.includes_symptoms 来实现.这与 SNOMED CT[55]不同,在 SNOMED CT 本体(最新版本为 ontology-2018-12-26_09-14-08.owl)中,所有疾病都为概念,概念之间通过 owl:subClassOf 属性建立分类关系.在 Freebase 中,所有实例只有 MID.例如,ns:m.0c58k 为糖尿病实例的 MID,*ns* 为 Freebase 的默认名称空间.在 SNOMED CT 本体中,所有概念 ID 采用 SNOMED CT 的约定编码.对于语义上相同的疾病,其在 Freebase 和 SNOMED CT 本体中的 rdfs:label 标注属性值并不完全相同.所以,Freebase 的 medicine 主题域中的知识和 SNOMED CT 本体之间并未建立联系.

3 疾病辅助诊断方法

3.1 症状权重的计算

$SKB_{medicine}$ 中症状 s 于疾病诊断权重 w_s 的计算,依赖于当前知识库中包含此症状 s 的疾病的数量(即与症状 s 具有语义关系的疾病的数量)和知识库中当前的疾病总数,一旦这两个数量中的任何一个发生变化,即应对 w_s 进行重新计算.

假定 $SKB_{medicine}$ 中包含的疾病总数为 N,对于每一个症状 s,我们定义 N_s 为与症状 s 具有语义关系的疾病的数量,w_s 是症状于疾病诊断的权重.那么 w_s 的计算方法如下:

$$w_s=\left(\frac{N-N_s}{N-1}\right)^2 \tag{1}$$

其中, $N_s\geqslant 1, w_s\leqslant 1$. 从公式(1)可见,与症状 s 具有语义关联的疾病的数量 N_s 越大,那么症状 s 于疾病的诊断权重越小.对方程取平方的目的是为了强调随着关联疾病数量的增加,所表现出来的症状权重之间的差异性.分母取 N–1 是因为当与症状 s 具有语义关联的疾病的数量 N_s 为 1 时,确保 w_s 为 1,即基于该症状,s 可唯一确定一种疾病.w_s 与 N_s 的关系如图 2 所示.

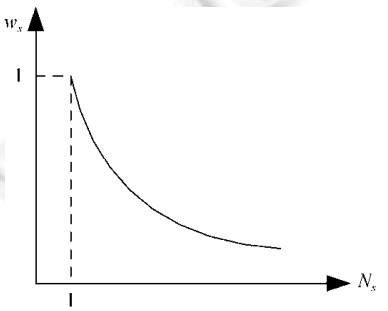

Fig.2 The relation between the weight of a symptom and the number of the diseases related to the symptom

图 2 症状于疾病诊断的权重和与该症状具有语义关联的疾病数量的关系

基于公式(1)可以为知识库中的每个症状 s 计算出一个 w_s,如图 3 所示.

Fig.3 Weights of symptoms in the domain semantic knowledge base for disease diagnosis

图 3 领域语义知识库中症状于疾病诊断的权重

3.2 疾病相关度的计算

基于知识库中每个症状 s 的 w_s,可以计算知识库中疾病 d_i 与患者输入症状集 S 的相关度,从而为医生或患者推荐可能的疾病列表及其相关度排名.

假定知识库中的症状数为 M,疾病数为 N;患者输入的症状集为 $S=\{s_1,s_2,...,s_j\},1\leqslant j\leqslant M$;知识库中与该症状集 S 中的一个或者多个症状具有语义关联的疾病构成的集合为 $D=\{d_1,d_2,...,d_i\},1\leqslant i\leqslant N$;其中,疾病 d_i 在知识库中关联的症状集合为 S_i, $S_i'=S_i\cap S$;疾病 d_i 与患者输入症状集 S 的相关度为 w_i,那么 w_i 的计算方法如下:

$$w_i=\frac{\sum_{s\in S_i'}w_s}{\sum_{s\in S_i}w_s} \tag{2}$$

公式(2)的分子仅考虑了症状集 S 中和疾病 d_i 相关联的症状,因为其他症状于 d_i 的相关度计算没有作用.而分母考虑到了知识库中与疾病 d_i 关联的所有症状,因为它们共同作用于疾病 d_i 的诊断.

假定患者输入“咽部异物感”症状,基于公式(2)可以计算出知识库中与此症状关联的疾病的相关度,如后文的图 4(a)所示.

3.3 相关症状推荐

在筛查出与患者输入症状集 S 相关联的疾病集合 D 以后,还需要对集合 D 中的疾病进行回顾性验证,即根据筛查出的疾病集合 D 评估和推荐与患者输入症状集 S 中的症状最相关的其他症状,供医生或患者确认;进而基于初始收集到的患者症状集 S 和患者再次确认的症状结果对疾病筛查结果集进行调整,并重新计算调整后疾病集 D 中疾病的相关度.

与患者输入症状集 S 中的症状最相关的症状集 S_{rel} 的推荐算法描述如下.

输入:患者初次输入的症状集 S.

输出:与患者输入症状集 S 中的症状最相关的 Top-6(这里仅保留 6 个,可以根据需要调整)个症状组成的症状集 S_{rel}.

S-1:首先从记录的历史输入症状组合中进行推荐.系统自动对不同患者输入和选择的历史症状组合进行记录,记录方式为 $\{s_1,s_2,...,s_i\}_f$,症状组合中的症状不分先后顺序,f 表示该症状组合发生的频率.如果患者输入的症状集 S 落入某一个或者多个历史症状组合中,则按 f 值从高至低,从某一个或者多个历史症状组合中选取除集合 S 中的症状以外的 Top-6 个症状作为 S_{rel},转到步骤 S-6;如果不够 6 个症状,则以实际可选择的症状数目作为 S_{rel},转到步骤 S-6;如果没有多余的症状或者 S 未落入任何一个历史症状组合,则转到步骤 S-2.注意,这个步骤选取症状时,不考虑症状的权重,而是按症状出现的先后顺序依次选取.

S-2:从知识库中查询与症状集 S 中的一个或者多个症状具有语义关联的疾病集合 $D=\{d_1,d_2,...,d_i\}$, $1\leqslant i\leqslant N$, N 表示领域语义知识库中的疾病数量.

S-3:设疾病 d_i 在领域语义知识库中关联的症状集合为 S_i,求 $S'=S_1\cup S_2\cup...\cup S_i$.

S-4:对集合 S'中的症状按照其 w_s 的大小进行降序排列得到 S^*.

S-5:从 S^*中选取前 Top-6 个症状作为与患者输入症状集 S 最相关的症状集合 S_{rel};如果 Top-6 个症状中包含症状集合 S 中的症状,则跳过,往后依次选取.

S-6:输出 S_{rel}.

如图 4(a)所示,根据患者输入的症状“咽部异物感”推荐了 6 个最相关的症状;当从推荐的 6 个症状中进一步选取了 3 个症状:“咽喉疼痛”“声嘶”“咽部充血”以后,将基于患者先后两次输入的共 4 个症状重新计算疾病的相关度,并调整排名,如图 4(b)所示.同时,也会基于这 4 个症状重新推荐最相关的症状集合 S_{rel}.

(a) 基于初始输入症状得到的诊断结果 (b) 基于初始输入症状和选择的推荐症状得到的诊断结果

Fig.4 Calculation of disease relevancy and recommendation of related symptoms

图 4 疾病相关度计算及相关症状推荐

整个疾病辅助诊断过程是在医生和患者参与下的一个循环迭代的过程.对于基层全科医生来说,可以从推荐的疾病列表中查看疾病的详细信息(包括疾病介绍、就诊科室、高发群体、有无传染性、症状、检查、诊断和鉴别、治疗、饮食宜忌、预防等)以判断是否需要做进一步的检查以及做哪些检查;由于推荐的疾病列表中的疾病在症状表现上具有一些相似性,所以可以通过诊断和鉴别信息进行鉴别诊断,如图 5(b)所示.如果用于患者自诊,患者可以通过选择推荐疾病列表中的某种疾病获得所在区域的医疗资源,包括所在区域的医院、科室以及医生的推荐.推荐过程还可以考虑对区域内医疗机构的客观评价(例如,诊疗水平、诊疗费用、诊疗效率等指标)以及患者的历史就诊行为偏好等,如图 5(a)所示;或者查看疾病的详细信息,并与自身表现进行对照.

(a) 医疗资源推荐 (b) 疾病详情

Fig.5 Medical resource recommendation and disease details

图 5 医疗资源推荐及疾病详细信息

4 方法评价

为了对本文建设的领域语义知识库和提出的疾病辅助诊断方法进行评价,本文从某个地市的健康医疗大数据中心随机选取了 6 种常见疾病的临床病历数据,这些病历数据来自多个不同的三甲医院,每份病历数据包括患者的性别、年龄、主诉、现病史、既往史、个人史、家族史、过敏史、查体、辅助检查、诊断等信息,这里只用到了主诉、现病史和诊断信息.但病历数据的质量并不高,比如,很多病历数据的主诉字段的值为空,或者为“未填写”,有些病历数据的主诉内容为“急性咽炎复查”“咨询”“要求彩超”等.这里在选取病历数据时,首先过滤掉了这几种情况的无效病历数据,然后组织医学专家对选取的病历数据做了进一步的筛查,过滤掉了一些质量不高的病历数据,即基于这些病历数据中的主诉和现病史中的症状描述很难得出相应的诊断.

从病历数据的主诉和现病史中抽取相关症状,作为患者输入的症状集 S,以病历数据中的诊断作为参考诊断,基于本文建设的领域语义知识库、公式(1)和公式(2)获取相关疾病列表及其相关度排名,分别选取 Top-1 诊断(第 1 诊断)和 Top-3 诊断(前 3 诊断)与参考诊断进行比较,如果 Top-1 诊断和参考诊断一致,表明 Top-1 命中;否则,如果 Top-3 中的某个诊断和参考诊断相一致,则表明 Top-3 命中.以病种为单位,分别统计 Top-1 和 Top-3 的命中率.然后从如下两个方面对本文建设的领域语义知识库和提出的疾病辅助诊断方法进行评价.

1) 与文献[1]中的方法就诊断命中率进行比较

采用本文建设的领域语义知识库和选取的 6 种常见疾病的临床病历数据,分别基于本文的方法和文献[1]中的方法统计 Top-1 和 Top-3 命中率,比较在不同病历规模情况下的命中率及其变化趋势,如图 6 所示.

从图 6 可见,针对随机选取的 6 种常见疾病的 Top-1 和 Top-3 命中率,本文的方法均高于文献[1]中的方法.在本文开始部分已对文献[1]中方法存在的问题进行了详细分析,采用文献[1]中的方法,主诉和现病史中的症状与知识库中对应参考诊断关联的症状越不相关(当然,必须至少有一个症状相关),得到的疾病相关度 w_i 反而越大.由于本文在

选取病历数据时有医学专家的参与,所以病历质量比较好,使得采用文献[1]中的方法仍有少量病历数据被命中.

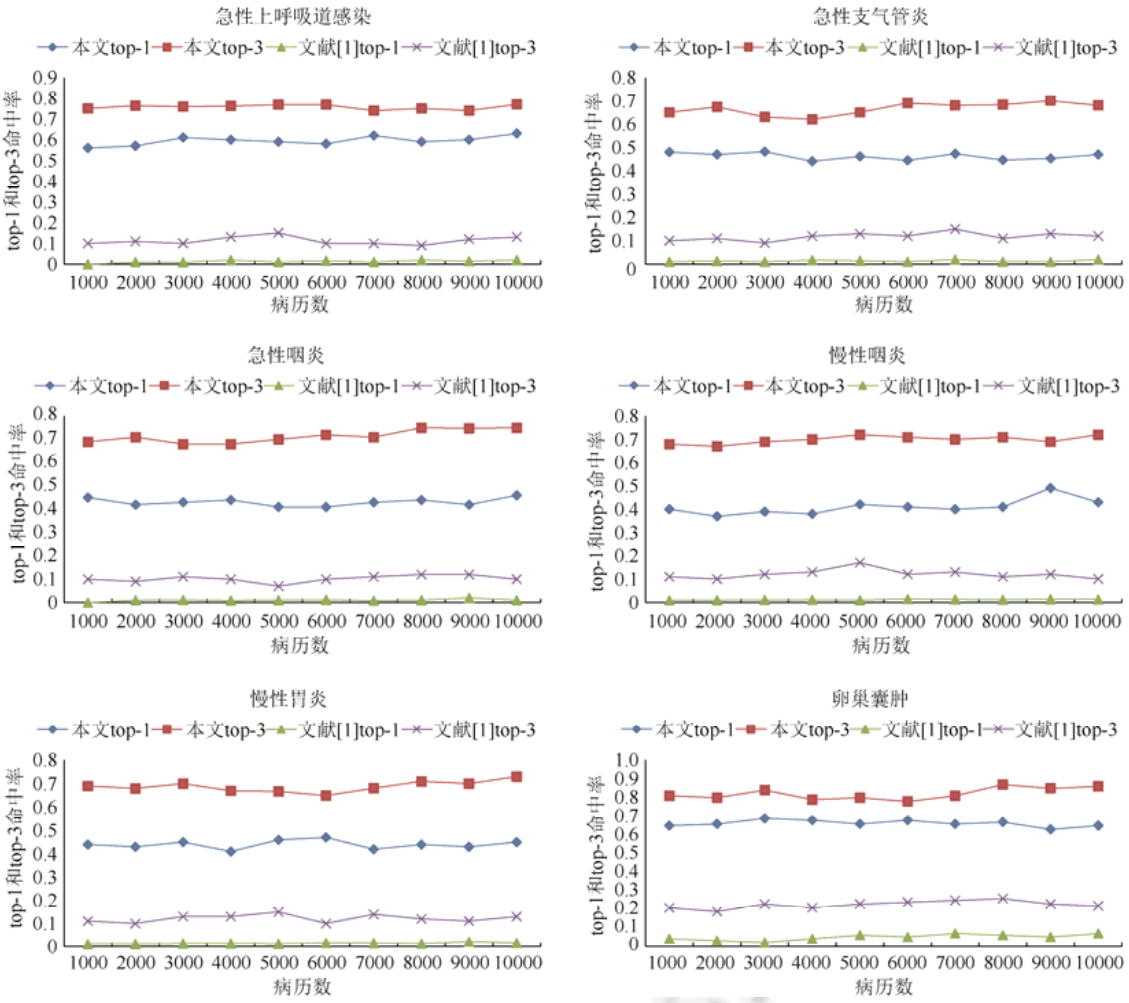

Fig.6 Comparison of the diagnostic hit ratio between the method in this paper and the one in Ref.[1]

图 6 本文方法和文献[1]中方法的诊断命中率比较

根据对病历数据来源地市的门/急诊诊断的统计,急性上呼吸道感染(简称上感,又称感冒)是发病率最高的疾病,广义的上感不是一个疾病诊断,而是一组疾病,包括普通感冒、病毒性咽炎、喉炎、疱疹性咽峡炎、咽结膜热、细菌性咽-扁桃体炎[56].急性上呼吸道感染的临床表现症状几乎涵盖了急性支气管炎、急性咽炎、慢性咽炎等疾病的症状,或者说这几种疾病可能经常被诊断为急性上呼吸道感染,所以急性上呼吸道感染的命中率相较急性支气管炎、急性咽炎和慢性咽炎稍微高一些,急性支气管炎、急性咽炎和慢性咽炎的命中率基本接近.与胃肠相关的疾病也是非常常见的疾病,具有很多共性的临床表现症状.卵巢囊肿是妇科疾病中非常常见的疾病,但其临床表现相对比较聚焦,所以卵巢囊肿的命中率相较其他几种疾病都要高.

随着病历数的增加,虽然 Top-1 和 Top-3 命中率有少量波动,但变化幅度不大.由于领域语义知识库保持不变,影响命中率的关键要素是来自主诉和现病史中的症状,又由于病历数据是经过医学专家筛选后所得,同种疾病,每千份病历数据之间的质量相差并不大,所以累计的 Top-1 和 Top-3 命中率变化幅度不大.

2) 与基于统计的方法就诊断命中率进行比较

以主诉和现病史中的症状作为特征属性,由医学专家参与对特征属性进行了选取,以症状的有无作为特征

属性的取值划分,以病历中的参考诊断作为分类标记,从筛选的 6 种疾病的病历数据中分别先后选取 1 000、4 000 和 8 000 份病历数据作为训练样本,即合计的训练样本数分别为 6 000、24 000 和 48 000,并使用朴素贝叶斯分类[57]和决策树分类[58]算法进行训练,然后使用不同训练样本规模下训练的模型分别对 6 种疾病的各 2 000 份病历(作为测试样本)进行 Top-1 和 Top-3 命中率的统计,结果见表 1 和表 2.

Table 1 Hit rate statistics based on naive Bayesian classification

表 1 基于朴素贝叶斯分类的命中率统计

命中率(%)		训练样本数		
		6 000	24 000	48 000
急性上呼吸道感染	Top-1	37.6	53.5	61.7
	Top-3	63.1	73.2	76.4
急性支气管炎	Top-1	22.3	38.5	48.3
	Top-3	53.7	61.8	64.3
急性咽炎	Top-1	36.3	41.2	45.7
	Top-3	55.9	66.7	75.1
慢性咽炎	Top-1	35.6	40.8	43.2
	Top-3	53.7	65.2	70.8
慢性胃炎	Top-1	35.7	41.3	44.6
	Top-3	54.8	63.2	73.7
卵巢囊肿	Top-1	42.6	68.3	65.8
	Top-3	68.6	78.4	85.7

Table 2 Hit rate statistics based on decision tree classification

表 2 基于决策树分类的命中率统计

命中率(%)		训练样本数		
		6 000	24 000	48 000
急性上呼吸道感染	Top-1	35.3	51.9	60.3
	Top-3	62.6	70.7	75.6
急性支气管炎	Top-1	22.5	37.3	45.8
	Top-3	52.3	60.1	63.2
急性咽炎	Top-1	36.1	40.6	44.7
	Top-3	54.5	67.3	73.8
慢性咽炎	Top-1	31.9	39.5	41.7
	Top-3	52.7	64.5	69.9
慢性胃炎	Top-1	31.3	41.7	43.4
	Top-3	53.8	61.9	72.4
卵巢囊肿	Top-1	41.8	57.5	65.3
	Top-3	67.8	77.7	84.9

基于表 1 和表 2 的统计结果可见,朴素贝叶斯分类和决策树分类的命中率结果基本接近,整体上,朴素贝叶斯分类的命中率稍高,但针对其中卵巢囊肿疾病的命中率,两者基本相当,主要原因仍是卵巢囊肿疾病的临床表现症状更加聚焦.在训练样本数为 6 000 时,Top-1 和 Top-3 的命中率不及本文的方法;在训练样本数为 24 000 时,Top-1 和 Top-3 的命中率有了显著提升,但仍不及本文的方法;在训练样本数为 48 000 时,Top-1 和 Top-3 的命中率接近本文的方法,其中,基于朴素贝叶斯分类的方法,急性咽炎和慢性胃炎的 Top-3 命中率稍高于本文的方法 0.1%和 0.7%,卵巢囊肿的 Top-1 命中率高于本文的方法 0.8%;基于决策树分类的方法,卵巢囊肿的 Top-1 命中率高于本文的方法 0.3%.

综上,文献[1]中的方法存在明显的问题和不足,基于朴素贝叶斯分类和决策树分类的方法,在训练样本较少时,明显不及本文的方法,在训练样本足够大时,与本文的方法接近甚至高于本文的方法.当然,除了样本的规模以外,朴素贝叶斯分类和决策树分类方法的准确性依赖于特征属性选取的有效性和样本的质量.本文方法的准确性也依赖于两个关键因素:领域语义知识库的规模和质量,以及公式(1)和公式(2)的有效性,但是通过回顾性验证可以提升本文方法的准确性.与朴素贝叶斯分类和决策树分类方法相比,本文的方法具有如下优势.

1) 避免“冷启动”问题:即在没有或没有足够的训练样本时,基于本文的方法可以达到更好的效果.

2) 可以快速支撑大量常见疾病的辅助诊断:采用朴素贝叶斯分类和决策树分类的方法,需要针对每种疾病选取有效的特征属性,并准备足够多的训练样本,结果的准确性对特征属性的有效性和训练样本的质量比较敏

感,很难在短期内提供对大量常见疾病的辅助诊断支持.而采用本文的方法则可以避开这些问题,快速支撑对大量常见疾病的辅助诊断.

当然,本文的方法也存在一些不足.一方面,领域语义知识库的建立和维护是一项知识密集型的工作,需要领域专家的参与;另一方面,在训练样本足够大时,本文的方法在准确性上不及朴素贝叶斯分类和决策树分类的方法.

5 相关工作

疾病诊断的质量主要依赖于医疗专家所掌握的医疗知识以及医疗经验,早期的专家系统(expert system,简称 ES)[59]试图通过对人类专家的问题求解能力的建模,采用 AI 中的知识表示和知识推理技术来模拟通常由专家才能解决的复杂问题,从而达到或超过专家解决问题的能力水平.ES 中的知识通常为采用规则描述的专家拥有的启发式经验知识,并采用基于规则推理(rule-based reasoning,简称 RBR)的方法提供领域问题求解服务.从专家那里获取知识的过程是一个时间密集型的过程,知识获取困难,而且依赖专家的意见,专家的意见有时具有主观性.文献[60,61]结合本体知识库和语义 Web 规则实现了一个高血压疾病的诊断模型,基于 RBR 提供高血压疾病的诊断推理.

在健康医疗领域,专家的经验知识往往蕴含于其诊治过的患者的病历数据中,如果能够直接利用蕴含于这些病历数据中的专家的经验知识,则将避开从专家那里直接获取经验知识的瓶颈,因为知识获取只不过是获得过去发生过的案例(case).基于案例的推理(case-based reasoning,简称 CBR)[62]分类法的研究即借鉴了这样的思想.CBR 的底层思想是基于这样一个假定:相似问题具有相似的解.例如,在医疗健康领域,通过收集和存储医疗专家诊治过的患者的病历和治疗方案,作为源案例库,并基于源案例库求解目标案例,即用来帮助诊断和治疗新的患者.由于传统 AI 技术存在的知识获取、记忆、维护等方面的问题,文献[63]讨论了实现智能医疗诊断系统的 CBR 方法学、研究问题和技术方面.Ain Shams 大学的医疗信息研究组(medical informatics research group)基于 CBR 技术开发了一个用于癌症和心脏病诊断的系统,文献[63]也对此进行了讨论.文献[64]讨论了 CBR 在医疗领域的适宜性,指出了存在的问题、局限性和部分克服这些问题、局限性的可能性.在健康医疗领域,专家的知识包括理论知识和经验知识,针对典型、复杂的诊疗案例,专家会基于理论知识,经验知识,特定的空间、时间和患者个体情况做出综合的诊疗建议推理.历史诊疗案例虽然可能蕴含了一部分专家的理论知识和经验知识,但仍然脱离了大量专家知识的支撑,所以新旧案例适配(adaptation)的合理性是 CBR 面临的主要问题.

健康医疗大数据的出现也为利用数据挖掘和机器学习技术直接从大量的历史病历数据中获取知识提供了可能性.文献[2]基于决策树和朴素贝叶斯算法提出了一种用于心脏疾病诊断的新方法,可以减少诊断需要输入的属性数量,从而减少诊断过程中需要对患者进行的实验数量,以提升诊断的效率.文献[3]使用减法聚类算法(subtractive clustering algorithm)开发了一个模糊推理系统(fuzzy inference system),并运用该系统对患者的 MRI 影像进行分类,以识别轻度认知障碍(mild cognitive impairment)、阿尔茨海默病(Alzheimer’s disease)和正常对照组(normal control).文献[4]使用 BP(backpropagation)学习算法训练了一个多层感知机(multi layer perceptron),用于诊断和预测新生儿疾病(neonatal diseases).文献[5]对有监督机器学习算法在临床上用于辅助诊断帕金森病(Parkinson’s disease)和进行性核上性麻痹(progressive supranuclear palsy)的可行性进行了评估.这些研究探讨了数据挖掘和机器学习算法在某些单病种疾病辅助诊断和预测方面的应用效果.利用数据挖掘和机器学习技术构建的疾病诊断模型,可以提供比人工手段更高的疾病识别率和检出效率.但针对不同病种,需要分别构建单独的疾病诊断模型,所以短时间内还无法提供对大量常见疾病的辅助诊断服务.

在健康医疗领域,已经积累了大量结构化的知识.例如,基于本体模型构建的领域语义知识库.基于领域语义知识库的方法可以快速提供大量常见疾病的诊断服务.文献[12]探讨了一些用于癌症疾病的基于本体的医疗系统的技术问题,也提出了一种基于本体的用于癌症疾病辅助诊断的方法学.该方法学能够被应用于帮助患者、学生和医生判断癌症的类型、癌症所处的分期以及如何治疗.文献[13]提出了一种新的遗传疾病鉴别诊断的数学模型,而不是传统的基因突变分析方法.它通过本体描述了“基因型-表型”关联关系.患者新出现的基因突

变被映射到人类表型本体(human phenotype ontology)中的标准化词汇表上.然后用这些术语进行鉴别诊断.将信息理论与模糊关系理论相结合,通过度量基于本体的语义相似度来实现鉴别诊断.该系统能够诊断 5 种复杂疾病的发生概率,即淋巴水肿产生综合症(lymphedema-distichiasis syndrome)、狄兰吉氏症候群(Cornelia de Lange syndrome)、科恩综合症(Cohen syndrome)和 Smith-Lemli-Opitz 症候群.文献[14]基于同现(co-occurrence)和信息内容,提出了度量本体中术语间相似性以及使用本体中的术语标注实体间的语义相似性的方法.新的相似性度量方法被证明比现有的使用生物学途径(biological pathway)的方法更好.该相似性度量方法使用与疾病相关的生物过程(biological processe)来评估疾病间的相似性,并使用已知疾病相似性的人工策划的数据集对该方法进行了评价.此外,使用本体来对疾病、药物和生物过程进行编码,并演示了一种方法,该方法使用基于网络的算法将有关疾病的生物学数据与药物信息结合起来,从而为现有药物找到新的用途.通过与现有的药物相关临床实验进行对比,验证了该方法的有效性.文献[1]的研究工作是欧洲项目 K4CARE[65]的一部分,该项目的目标是将医疗保健和一些西方和东欧国家的信息和通信技术(ICT)经验结合起来,以建立、实施和验证一个基于知识的医疗保健模式,以向居家老年患者提供专业援助.该项目聚焦于 9 种慢性疾病、2 种综合征和 5 个社会问题,使用 CPO(case profile ontology)本体描述了与它们相关的知识,并使用 SDA(state-decision-action)图描述了相关的干预计划.文献[1]展示了该项目中开发的用于疾病诊断和本体个性化的方法和工具,其疾病诊断方法存在一些明显的问题和不足,在本文的介绍部分已给出了详细说明.

表 3 对已有的疾病辅助诊断方法的研究进行了总结,分析了各自的优势和劣势.这些方法除了在技术原理上不同以外,它们利用的数据也不一样.比如,文献[1]和文献[12]的方法利用了患者的症状和体征数据,文献[13]的方法利用了患者的基因数据,文献[3]的方法利用了患者的 MRI 影像数据等.有的数据(例如,症状和体征)比较好获取,而有的数据(例如,基因数据)的获取就比较困难.

为了面向基层全科医生提供大量常见疾病的辅助诊断服务以及面向患者提供疾病自诊服务,本文的研究采用基于领域语义知识库的疾病辅助诊断方法.针对现有研究的不足,本文进行了校正和完善.

Table 3 Comparison of the aided diagnosis methods for diseases

表 3 疾病辅助诊断方法比较

方法	描述	优势	劣势
基于 RBR 的方法[60,61]	基于规则描述疾病诊疗的知识或者专家拥有的启发式经验知识,并基于 RBR 提供领域问题的求解服务	准确性高、效率高	规则知识获取和维护困难;当规则比较少时,将无法提供问题解;当规则数量比较大时,规则推理效率低
基于 CBR 的方法[63,64]	基于“相似问题具有相似解”的思想,直接存储专家诊治过的历史病历作为源案例库,并基于目标案例和源案例的相似性比较,为目标案例直接提供问题解或者修正后的解	不需要明确的领域模型,避开了知识获取的瓶颈;可以很快地产生问题的解;问题解容易理解,具有直接的案例证据;即使具有少量的案例,CBR 也可以运行	由于案例涉及患者隐私,案例获取困难;案例脱离了部分理论知识和经验知识的支撑,案例适配的合理性是一个主要的问题;CBR 推理过程不具有重用性等
基于统计分析的方法[2−5]	利用数据挖掘和机器学习技术从健康医疗大数据中获取疾病诊疗的模型知识,并提供疾病诊疗服务	可以提供比人工手段更高的疾病识别率和检出效率	针对不同病种,需要分别构建单独的疾病诊断模型,短时间内无法提供对大量常见疾病的辅助诊断服务
基于领域语义知识库的方法[1,12−14]	采用结构化的领域语义知识库直接建立用于疾病诊疗的相关知识,并基于知识查询和知识推理提供疾病诊疗服务	可以快速提供对大量常见疾病的辅助诊断服务	准确性不高;知识构建和维护是一项知识和时间密集型的工作

6 总结和进一步的工作

本文从 Freebase RDF dump 包中抽取了 medicine 主题域的知识,并基于本体构建了 medicine 领域的语义知识库,由医学专家对疾病、症状及其语义关系进行了校验和完善.在此基础上,提出了基于领域语义知识库的疾病辅助诊断方法,包括计算知识库中症状于疾病诊断的权重、计算与患者输入症状集相关的疾病列表及其相关度排名、推荐与患者输入症状集中症状最相关的症状集供医生或患者进一步确认.整个疾病辅助诊断过程是在医生和患者参与下的一个循环迭代的过程,包括前瞻性诊断和回顾性诊断.最后基于医学专家筛选的真实病历

数据对本文建设的领域语义知识库和提出的疾病辅助诊断方法进行了评价,包括与文献[1]中方法的对比、与基于统计的方法的对比.对比结果表明,本文方法解决了文献[1]中方法存在的问题和不足.同时,与基于统计的方法对比,本文方法可以避免“冷启动”问题,可以快速支撑大量常见疾病的辅助诊断.采用本文的方法,有望为基层全科医生提供常见疾病的辅助诊断服务,或者为患者提供疾病自诊服务.

但本文的方法仍然存在一些不足,这也是下一步需要研究的工作.

1) 引入更多的诊断要素:针对患者疾病自诊的场景,通过问卷的方式进一步获取年龄、性别、既往病史、家族病史等诊断要素;针对院内的辅助诊断,可以从患者的历史诊疗记录中获取既往病史、家族病史、历史检查检验结果等信息,也可能得到最新的查体、检查、检验等结果数据;

2) 对一些常见的疾病直接建立辅助诊断规则,先进行基于规则推理的疾病诊断;

3) 对本文的方法进行改进,一方面需要考虑多维诊断要素之间存在的关联关系和权重,另一方面需要考虑不同诊断要素对疾病诊断的贡献作用.

References:

[1] Romero-Tris C, Riaño D, Real F. Ontology-based retrospective and prospective diagnosis and medical knowledge personalization. In: Knowledge Representation for Health-care. Berlin: Springer-Verlag, 2010. 1–15. [doi: 10.1007/978-3-642-18050-7_1]

[2] Bhatla N, Jyoti K. A novel approach for heart disease diagnosis using data mining and fuzzy logic. Int’l Journal of Computer Applications, 2012,54(17):16–21. [doi: 10.5120/8658-2498]

[3] Krashenyi I, Popov A, Ramirez J, Gorriz JM. Application of fuzzy logic for Alzheimer’s disease diagnosis. In: Proc. of the Signal Processing Symp. IEEE, 2015. 85–88. [doi: 10.1109/SPS.2015.7168288]

[4] Chowdhury DR. An artificial neural network model for neonatal disease diagnosis. Int’l Journal of Artificial Intelligence & Expert Systems, 2011,2(3):96–106.

[5] Salvatore C, Cerasa A, Castiglioni I, *et al*. Machine learning on brain MRI data for differential diagnosis of Parkinson’s disease and progressive supranuclear palsy. Journal of Neuroscience Methods, 2014,222:230–237. [doi: 10.1016/j.jneumeth.2013.11.016]

[6] Studer R, Benjamins VR, Fensel D. Knowledge engineering: Principles and methods. Data and Knowledge Engineering, 1998, 25(1/2):161–197. [doi: 10.1016/S0169-023X(97)00056-6]

[7] Berners-Lee T, Hendler J, Lassila O. The semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 2001,284(5):34–43.

[8] Smith B, Ashburner M, Rosse C, *et al*. The OBO foundry: Coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology, 2007,25(11):1251–1255. [doi: 10.1038/nbt1346]

[9] Licata G. Employing fuzzy logic in the diagnosis of a clinical case. Health, 2010,2(3):211–224. [doi: 10.4236/health.2010.23031]

[10] Köhler S, Schulz MH, Krawitz P, *et al*. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. American Journal of Human Genetics, 2009,85(4):457–64. [doi: 10.1016/j.ajhg.2009.09.003]

[11] Bodenreider O. Disease ontology. In: Encyclopedia of Systems Biology. 2013. 578–581.

[12] Alfonse M, Aref MM, Salem ABM. An ontology-based cancer diseases diagnostic methodology. In: Recent Advances in Information Science. 2013. 95–99.

[13] Jayaratne L. Ontology based approach for diagnosis in personalized medicine. In: Proc. of the Int’l Conf. on Computer Games, Multimedia and Allied Technology. 2015.

[14] Mathur S. Ontology-based methods for disease similarity estimation and drug repositioning [Ph.D. Thesis]. Kansas City: Computer Science and Mathematics, University of Missouri-Kansas City, 2012.

[15] Izumi S, Dai K, Itabashi G, *et al*. An ontology-based advice system for health and exercise. In: Proc. of the 10th Iasted Int’l Conf. on Internet and Multimedia Systems and Applications. DBLP, 2006. 95–100.

[16] Cantais J, Dominguez D, Gigante V, *et al*. An example of food ontology for diabetes control. In: Proc. of the Int’l Semantic Web Conf. Workshop on Ontology Patterns for the Semantic Web. 2005.

[17] Kostopoulos K, Chouvarda I, Koutkias V, *et al*. An ontology-based framework aiming to support personalized exercise prescription: Application in cardiac rehabilitation. IEEE Engineering in Medicine and Biology Magazine, 2011,2011(4):1567–1570.

[18] Manola F, Miller E, McBride B. RDF 1.1 Primer. W3C, 20140225, 2014. https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140225/

[19] Hebeler J, Fisher M, Blace R, Perez-Lopez A. Semantic Web Programming. Indianapolis: Wiley Publishing, 2009.

[20] Brickley D, Guha RV. RDF Schema 1.1. W3C, 20140225, 2014. https://www.w3.org/TR/2014/REC-rdf-schema-20140225/

[21] Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph S. OWL 2 Web Ontology Language Primer. 2nd ed., W3C, 20141211, 2012. http://www.w3.org/TR/2012/REC-owl2-primer-20121211/

[22] Antoniou G, Harmelen FV. A Semantic Web Primer. 2nd ed., Cambridge: MIT Press, 2008.

[23] Horrocks I, Patel-Schneider PF, Boley H, Tabet S, Grosof B, Dean M. SWRL: A semantic Web rule language combining OWL and RuleML. W3C, 20040521, 2004. https://www.w3.org/Submission/SWRL/

[24] Noy NF, McGuinness DL. Ontology development 101: A guide to creating your first ontology. Technical Report, KSL-01-05, Stanford Knowledge Systems Laboratory, 2001. [doi: 10.1016/j.artmed.2004.01.014]

[25] Garlik SH, Seaborne A. SPARQL 1.1 query language. W3C, 20130321, 2013. https://www.w3.org/TR/sparql11-query/

[26] Kim IW, Lee KH. A model-driven approach for describing semantic Web services: From UML to OWL-S. IEEE Trans. on Systems, Man, and Cybernetics Part C: Applications and Reviews, 2009,39(6):637–646. [doi: 10.1109/TSMCC.2009.2023798]

[27] Iribarne L, Padilla N, Asensio JA, Criado J, Ayala R, Almendros J, Menenti M. Open-environmental ontology modeling. IEEE Trans. on Systems, Man and Humans, 2011,41(4):730–745. [doi: 10.1109/TSMCA.2011.2132706]

[28] Lee CS, Kao YF, Kuo YH, Wang MH. Automated ontology construction for unstructured text documents. Data & Knowledge Engineering, 2007,60:155–176. [doi: 10.1016/j.datak.2006.04.001]

[29] Raufi B, Ismaili F, Zenuni X. Modeling a complete ontology for adaptive Web based systems using a top-down five layer framework. In: Proc. of the ITI 31st Int'l Conf. on Information Technology Interfaces. 2009. 511–518. [doi: 10.1109/ITI.2009.5196136]

[30] Li J, Meng LS. Comparison of seven approaches in constructing ontology. New Technology of Library and Information Service, 2004,7:17–22 (in Chinese with English abstract).

[31] Subhashini R, Akilandeswari J. A survey on ontology construction methodologies. Int'l Journal of Enterprise Computing and Business Systems, 2011,1(1).

[32] Jadhav SB, Pardeshi SN. Ontology intergration with semantic similar entity classes amongst different ontologies for enhanced information retrieval. Int'l Journal of Recent Trends in Engineering, 2009,2(3):132–134. [doi: 10.1.1.381.7277]

[33] Zaiβ K, Schlüter T, Conrad S. Instance-based ontology matching using different kinds of formalisms. Int'l Journal of Computer, Electrical, Automation, Control and Information Engineering, 2009,3(7):1716–1724. [doi: 10.1.1.193.3112]

[34] Lambrix P, He T. Ontology alignment and merging. Computational Biology, 2008,6:133–150. [doi: 10.1007/978-1-84628-885-2_6]

[35] Du XY, Li M, Wang S. A survey on ontology learning research. Ruan Jian Xue Bao/Journal of Software, 2006,17(9):1837–1847 (in Chinese with English abstract). http:www.jos.org.cn/10000-9825/171837.htm [doi: 10.1360/jos171837]

[36] Freebase. 2018. https://en.wikipedia.org/wiki/Freebase

[37] Bollacker K, Cook R, Tufts P. Freebase: A shared database of structured general human knowledge. In: Proc. of the AAAI Conf. on Artificial Intelligence. Vancouver: DBLP, 2007. 1962–1963.

[38] Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In: Proc. of the SIGMOD 2008. 2008. 1247–1250.

[39] DBpedia. 2018. http://wiki.dbpedia.org/

[40] YAGO. 2018. https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[41] Wikipedia. 2018. https://www.wikipedia.org/

[42] MusicBrainz. 2018. https://musicbrainz.org

[43] WordNet. 2018. http://wordnet.princeton.edu/

[44] Linked Data. 2018. http://linkeddata.org/

[45] Attribution 2.5 Generic (CC BY 2.5). 2018. https://creativecommons.org/licenses/by/2.5/

[46] Beckett D. RDF 1.1 N-Triples. W3C, 20140225, 2014. https://www.w3.org/TR/n-triples/

[47] Niel C. Freebase-triples: A methodology for processing the freebase data dumps. eprint arXiv:1712.08707, 2017.

[48] Apache Jena. 2018. http://jena.apache.org/

[49] Chen D, Zhao H. Research on the method of extracting domain knowledge from the freebase RDF Dumps. IEEE Access, 2018,6:50306–50322. [doi: 10.1109/ACCESS.2018.2868516]

[50] Beckett D, Berners-Lee T. Turtle-Terse RDF triple language. W3C, 20110328, 2011. https://www.w3.org/TeamSubmission/turtle/

[51] Apache Jena—TDB. 2018. http://jena.apache.org/documentation/tdb/index.html

[52] Apache Jena Fuseki. 2018. http://jena.apache.org/documentation/fuseki2/index.html

[53] Protégé ontology editor. 2018. http://protege.stanford.edu/

[54] Int'l classification of diseases. 10th Revision, Clinical Modification (ICD-10-CM), 2018. https://www.cdc.gov/nchs/icd/icd10cm.htm

[55] SNOMED CT. 2018. http://www.snomed.org/

[56] Acute upper respiratory tract infection. 2018 (in Chinese). https://baike.baidu.com/item/急性上呼吸道感染

[57] Naive Bayes classifier. 2018. https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[58] Decision tree learning. 2018. https://en.wikipedia.org/wiki/Decision_tree_learning

[59] Expert system. 2018 (in Chinese). http://www.intsci.ac.cn/ai/es.html

[60] Gong M. Research on ontology based knowledge base of hypertension electronic medical records [MS. Thesis]. Xi'an: Xi'an Electronic and Science University, 2010 (in Chinese with English abstract).

[61] Gong M, Wen Y. Ontology based knowledge base for diagnosis of hypertension. Journal of Intelligence, 2010,29(b06):169–172 (in Chinese with English abstract).

[62] Kamber M, Han J, Pei J. Data Mining: Concepts and Techniques. 3rd ed., Elsevier, 2011.

[63] Salem ABM. Case based reasoning technology for medical diagnosis. In: Proc. of the World Academy of Science Engineering & Technolog. 2007.

[64] Schmidt R, Gierl L. Case-based reasoning for medical knowledge-based systems. Studies in Health Technology and Informatics, 2000,77(2-3):720–725.

[65] Campana F, Moreno A, *et al*. K4Care: Knowledge-based homecare e-services for an ageing Europe. In: Annicchiarico R, Cortés U, Urdiales C, eds. Agent Technology and e-Health (Whitestein Series in Software Agent Technologies and Autonomic Computing). Switzerland: Birkhäuser Verlag Basel, 2007. 95–115. [doi: 10.1007/978-3-7643-8547-7_6]

附中文参考文献:

[30] 李景,孟连生.构建知识本体方法体系的比较研究.现代图书情报技术,2004,7:17–22.

[35] 杜小勇,李曼,王珊.本体学习研究综述.软件学报,2006,17(9):1837–1847. http:www.jos.org.cn/10000-9825/171837.htm [doi: 10.1360/jos171837]

[56] 急性上呼吸道感染.2018.https://baike.baidu.com/item/急性上呼吸道感染

[59] 专家系统.2018.http://www.intsci.ac.cn/ai/es.html

[60] 巩沐歌.基于本体的高血压电子病历知识库研究[硕士学位论文].西安:西安电子科技大学,2010.

[61] 巩沐歌,温有奎.基于本体的高血压疾病诊断知识库.情报杂志,2010,29(b06):169–172.

陈德彦(1977－),男,博士,正高级工程师,主要研究领域为自然语言处理,语义 Web,知识工程,数据挖掘,机器学习,网络与信息安全.

赵宏(1954－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为下一代网络,网络与信息安全,网络管理,图像处理.

张霞(1965－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为软件架构,软件工程,数据库,大数据,人工智能.