doi:10.3969/j.issn.1000-2162.2019.06.007

基于医在回路的医疗健康知识图谱系统架构的研究

盛　明,张　勇,邢春晓
(清华大学 信息技术研究院 WEB与软件技术研究中心,北京 100084)

摘　要:知识图谱提供的服务质量在很大程度上取决于知识图谱构建的质量.自动构建知识图谱的方法已被广泛应用于许多领域,但知识图谱在医学领域的应用却面临着很多困难,原因有:医学概念/关系/事件的复杂和模糊性;数据标准不一致,源数据质量差;医疗数据异构多元化严重,如电子医学病例(electronic medical record,简称 EMR)等.在构建过程中,需要来自医学专家的大量先验知识和人工辅助.引入一个系统架构,该架构明确了在何时何处引入医学专家的相关工作,从而提高医疗健康知识图谱构建的质量和效率.

关键词:医疗知识图谱构建;医在回路;电子医学病例
中图分类号:TP391　　**文献标志码**:A　　**文章编号**:1000-2162(2019)06-0048-07

A knowledge graph framework for health based on doctor-in-the-loop

SHENG Ming, ZHANG Yong, XING Chunxiao
(Web and Software R&D Center of Research Institute of Information Technology, Tsinghua University, Beijing 100084, China)

Abstract: The quality of service (QoS) that a knowledge graph can provide largely depends on the quality of the knowledge. Automatic methods have been widely used in many domains to construct the knowledge graphs. However, it is more complex and difficult in the medical domain. There are three reasons: the complex and obscure nature of medical concepts and relations, inconsistent standards and heterogeneous multi-source medical data with low quality like EMRs. Therefore, the quality of knowledge graph requires a lot of manual efforts from experts in the process. In this paper, we introduced an overall framework that provided insights on where and when to import manual efforts in the process to construct a health knowledge graph. In this framework, four tools were provided to facilitate the doctors' contribution, i.e. matching synonym, discovering and editing new concepts, annotating concepts and relations, together with establishing rule base. The application for cardiovascular diseases demonstrated that this framework could improve the accuracy and efficiency of medical knowledge graph construction.

Keywords: medical knowledge graph construction; doctor-in-the-loop; EMR (electronic medical record)

收稿日期:2019-02-19
基金项目:国家自然科学基金资助项目(91646202);国家重点研发计划基金资助项目(2018YFB1404400,2018YFB1402700)
作者简介:盛明(1985—),男,河南郑州人,清华大学工程师,E-mail: shengming@tsinghua.edu.cn.

知识图谱可以将来自不同来源的信息和知识融合在一起. 在过去几年中，许多知识图谱，无论是面向通用的还是特定领域的，都已经被构建出来并且成为相关领域的宝贵资源. 概念医学知识图谱，如 UMLS(unified medical language system)、Gene Ontology 等，仅包含医学领域的概念. 事实医学知识图谱，如 Google Health Knowledge Graph，Knowlife，PDD Graph(patients，diseases and drugs graph)等，包括了概念和实例.

在医学领域，知识图谱是一个非常有用的工具，可以支持疾病预测、药物推荐[1]等服务. 很多知识图谱的构建[2-3]采用全自动化方法，没有任何人工参与，这些知识图谱的数据主要来自互联网. 尽管这些全自动化方法可以节省医学专家的时间和精力，但是当涉及特定医疗健康领域知识图谱的构建时，由于医学领域的概念/关系/事件是复杂而模糊的、医学领域的源数据质量差[4]、医学领域的数据标准不统一等原因，它们的表现不尽如人意.

因此，用于完全自动构建知识图谱的通用方法不能直接应用于医疗健康领域. 为了提高图谱质量，在构建过程中引入一些医学专家的先验知识是非常必要的. 另一方面，如果构建过程中涉及太多医学专家的工作，则需要花费大量的时间和精力，整个构建的效率将大大降低[5]. 更糟糕的是，整个系统将不具备可扩展性，无法适应和扩展到其他新的医学主题[6]. 因此，需要在合适的位置引入医学专家的相关工作. 医学专家的工作和自动化方法之间的平衡是非常重要的，需要谨慎对待.

论文介绍了一个系统架构，该架构表明在医疗健康领域知识图谱的构建过程中有哪些环节、在什么时机需要引入医学专家的工作等. 其目的是：基于自动化的方法可以帮助医生节省时间和精力，基于医生的先验知识可以弥补相关通用或自动化图谱构建方法的不足. 通过这种方式，提高了知识图谱构建的效率和质量.

1　相关工作

1.1　知识图谱构建工具

现在已经有许多自动知识图谱构建的工具，这些工具可以处理海量数据并无须人工参与构建知识图谱. 在医疗领域中，典型的知识图谱构建工具有 RDR(ripple-down rules)[7]、cTAKES(clinical text analysis and knowledge extraction system)[8]、pMineR(process mining R library)[9]、I-KAT[10]、myDIG、semTK (the semantics toolkit). 表 1 是这些工具的对比.

表 1　知识图谱构建工具

名称	领域	数据源	实体识别	关系抽取	实体对齐	数据模型映射	人工参与
RDR	medical	—	×	×	×	×	√
cTAKES	medical	UMLS	√	√	√	×	×
pMineR	medical	EMR (electronic medical record)	×	×	×	×	×
I-KAT	medical	SNOMED-CT(systematized nomenclature of medicine-clinical term)	×	×	×	√	√
myDIG	general	csv，JSON	√	√	×	×	×
semTK	general	csv	×	×	×	√	×

如表 1 所示，主流知识图谱构建工具包括 RDR、cTAKES、pMineR、I-KAT 等. 可以看出只有不到一半的工具涉及图谱构建过程中的人工参与. 它们中任何一个工具都没有完全包含 5 个常用功能：实体识别、关系抽取、实体对齐、数据模型映射(从 ER(entity relation)模型到 RDF(resource description framework)模型))、人工参与. 因此，使用这些工具构建医学知识图谱的效果较差.

1.2 医生在医学知识信息化过程中的角色

如何组织医学知识一直是一个重要问题. 文献[11]根据医生的先验知识和修订意见建立了生物医学知识库,并使用贝叶斯网络进行疾病预测. 案例基础推理CBR(case-based reasoning)可以组织文本医学知识并将其整合到案例中. 在CBR系统中,数据需要通过特征提取、特征选择和加权进行预处理,这些步骤通常在医生的帮助下进行. 首先,临床医生可能会向系统提供一些初步经验或知识,然后将这些知识经验用于解决新病例. 在此过程中,医生可能会对他们以前的知识进行一些调整. 案例解决后,这些知识集得以更新. 此外,有很多大型生物医学本体库,如基因本体库、疾病本体库或其他关联生命数据本体库等,为人们提供更加全面的结果.

1.3 医在回路(doctor-in-the-loop)

在医学领域,基于机器学习的自动方法在许多方面取得了显著成果,如疾病预测和临床记录分类. 尽管医学领域的自动机器学习(automatic machine learning,简称aML)吸引了许多研究人员的兴趣并且一直发展迅速,但这些方法缺点在于其无法解释性[12]. 机器学习模型通常被视为"黑箱",内部结构和原则超出了人们的理解范围[13]. 更重要的是,aML需要具有大量训练集才能获得较好的结果,但在医学领域,数据集是有限的,研究人员可能会遇见一些特例事件,这将导致aML受到训练数据集不足的影响. 因此,需要能够与医学专家交互并且可以通过这些交互来优化其学习行为的算法. 通过这种互动,可以启发式地选择训练样本,并且可以大大减少研究时间. 涉及人工交互的算法可以被定义为人在回路[14]. 人在回路实际上已经被应用于人工智能的许多方面,如命名实体识别[15]和规则学习. 在医学领域,大都是尝试结合医在回路机制来改善性能,特别是在知识图谱构建方面.

2 架构和工作流

2.1 架构

图1为应用医在回路的医疗知识图谱构建的架构.

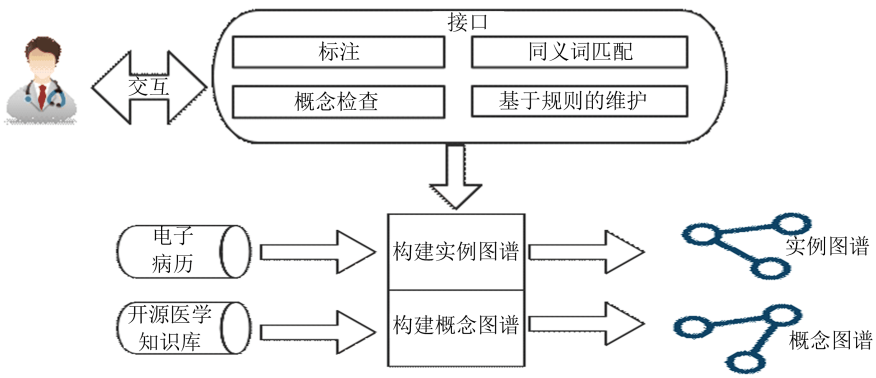

图1 应用医在回路的医疗知识图谱构建的架构

如图1所示,构建疾病特定医学知识图谱的系统应包括疾病专科医生,这是整个构建过程中最重要的部分;其他部分主要包括人机交互接口、数据来源、图谱构建工具、概念图谱和实例图谱.

在图谱的构建过程中,医生应能够与系统互动. 因此,应该为医生提供一套接口. 通过这种方式,医生可以将他们的经验和知识应用到构建系统中. 知识图谱构建的完全手动方法不仅耗时而且容易出错,需要一套用于构建医学知识图谱的自动化工具. 因此,通过提供接口,系统能够设法将医生的知识与自动构建方法结合起来.

2.2 工作流

图2详细展示了应用人在回路的医疗知识图谱构建的工作流程. 在该系统中,医生的参与主要体现在4个方面:同义词匹配融合和概念对齐;新词发现和新概念;标注实体和电子病历(EMR)的关系提取;建立规则库,包含实体和关系提取的映射规则和模式.

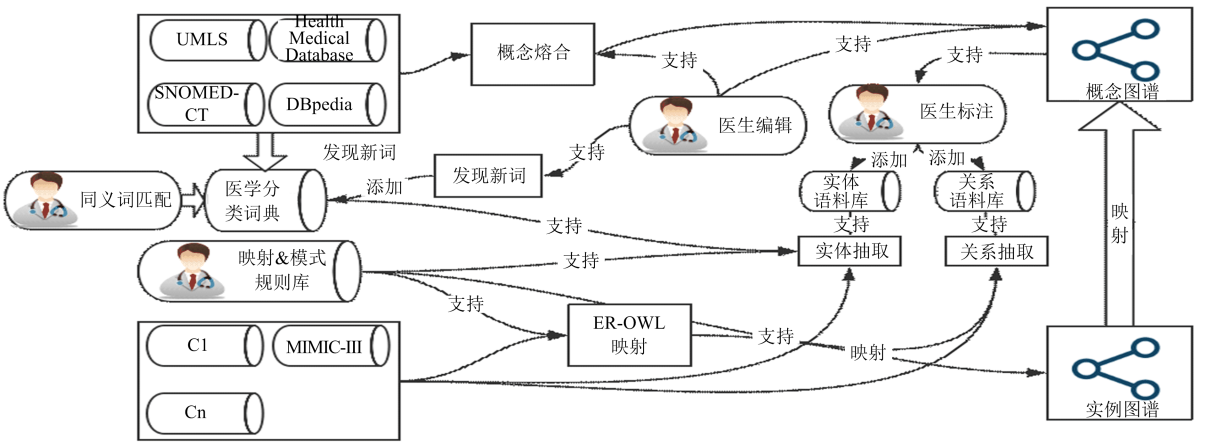

图 2　应用人在回路的医疗知识图谱构建的工作流程

3　构建过程中的医在回路

3.1　同义词对齐模块

现有的医学知识库是知识图谱的重要来源. 为了充分利用信息，具有相同含义的不同概念和关系必须适当对齐并融合在一起. 为了提高自动匹配方法的准确性和手动对齐方法的效率，论文提出了一个同义词模块，这个模块可以整合医生和自动匹配器的结果. 这个模块中有两个阶段：匹配阶段，聚合阶段.

该模块在语料库级别上工作，并且能够跨不同的数据源操作. 医生可以在模块中输入新单词或短语，然后将输入文本传递到匹配器库（一组不同的匹配器）上进行处理. 匹配库将输入文本的可能同义词的候选列表返回给医生. 候选名单大约包含 10 个语料，这大大缩小了医生的搜索范围. 在这之后，医生可以自己决定列表中的语料是否是输入文本的同义词. 如果列表中有与输入文本同义的语料，则医生可以将其与他们认为最匹配的现有语料对齐. 如果没有，医生可以创建新节点并且把新输入的文本集成到语料库中，存储在同义词库中的词可以用来支持实体提取. 图 3 为同义词匹配模块的工作流，图 4 为概念结构的层级.

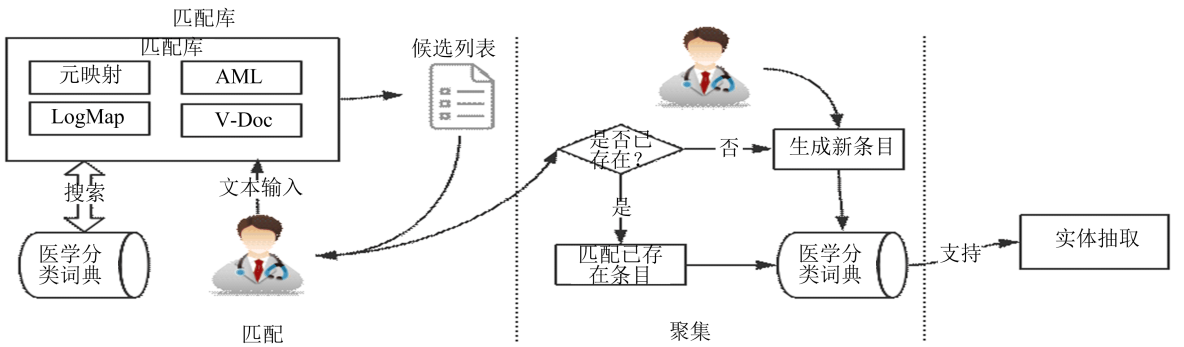

图 3　同义词匹配模块的工作流

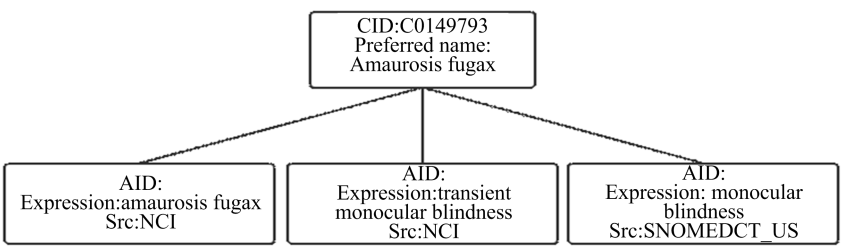

图 4　概念结构的层级

该模块的关键部分是如何组织具有不同拼写、来自不同数据源但含义相同的单词和短语. 为了解决这个问题,引入了一个层次结构表达方式,给每个不同的概念(具有独特含义的词/短语)分配一个唯一的概念身份(concept identity,简称CID). 概念可能有许多表达形式,但只有一个表达形式是首选,此首选表达式是该概念的默认表示形式. 对于具有相同含义但拼写不同或数据源不同的表达形式,为每个表达形式分配唯一的原子标识(atom identity,简称AID),AID是相应CID的子节点.

3.2 概念编辑模块

与在语料库级别上工作的同义词匹配模块不同,该模块在概念级别上工作,并且可以向医生提供在概念图上操作的接口. 该概念审核模块主要为医生提供两种功能:概念选择和对齐,新词发现.

知识图谱构建的质量在很大程度上取决于图谱包含的概念. 然而,由于医学术语的模糊性和专业性,医学词库中的概念必须由医生仔细检查. 在建立分层次存储概念的医学词库之后,医生应该能够根据他们自己的要求和掌握的知识来审查概念,并决定将哪些概念放入最终的概念图谱中.

此功能类似于同义词匹配模块提供的服务,但在概念知识图谱级别上运行. 如果医生想要将医学词库中的新概念添加到概念图谱中,可以输入文本,然后输入的字符串将被传入医学词库和概念图谱上的搜索引擎上. 医学词库上的搜索引擎将返回与医生输入字符串对应的概念列表,概念图谱上的搜索引擎将从图谱中返回与输入字符串相似的概念列表. 医生只需要快速扫描搜索引擎提供的概念列表,而不必手动搜索整个词库中的大量概念. 医生可以自己决定与输入字符串相对应的概念是否是新概念:如果是,医生可以从与输入字符串对应的概念列表中选择一个并添加到概念图谱中;如果医学词库上的搜索引擎没有返回与输入字符串对应的结果,则进入新词发现模块并更新医学词库. 图5为概念选择对齐模块的工作流程.

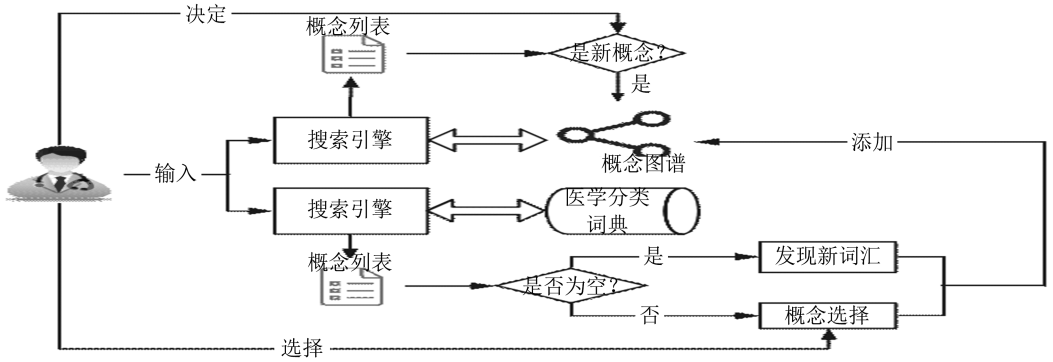

图5 概念选择对齐模块的工作流程

虽然医学知识库中存储的信息十分丰富,但是临床实践中仍有医学术语尚未纳入词库,这些医学术语可能来自患者的EMR,或仅仅来自医生的先验知识. 新单词发现功能为医生提供了一组接口,医生可以通过这些接口以定制添加不在医学词库中的术语和概念.

(1) 数据驱动方法

该方法可以从患者的EMR中获取信息. 患者的EMR记录了患者的实际情况,可以作为构建高质量医学知识图谱的数据来源. 但是,EMR的某些特征未存储在概念图谱中. 表2为病人EMR的一部分.

表2 病人EMR的一部分

项目	心尖搏动	心音A2	心包摩擦
结果	加剧	分裂	正常

表2显示了心尖部搏动心音A2和心包摩擦都是心脏疾病诊断的重要特征. 但是,这几个特征中没有一个可以与概念图谱中的概念对齐. 在这种情况下,医生可以使用该模块提供的接口将这个新概念添加到图谱中.

(2) 需求驱动方法

除了根据患者EMR中的特征定义概念之外,医生可以根据自己的经验定义一些概念和关系. 有

时，EMR 中的信息过于复杂，并且涉及很多方面. 有些特征过于分散，而医生只想专注于特定的几个特征. 在这种需求驱动的方法中，医生可以先抛开 EMR，并在更高层次上定义概念和关系. 图 6 展示了由医生定义的图谱示例.

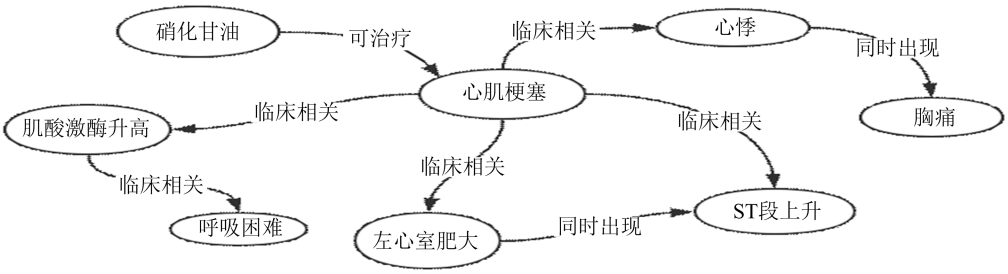

图 6　由医生定义的概念图谱

3.3　实体 & 关系标注模块

为了从患者的 EMR 中获取信息，需要提取实体和关系，提取的质量在很大程度上取决于标注. 然而，在医学领域，有许多实体类型不符合传统定义的 4 类范式：人名、地名、机构名、混杂类型. 例如，在临床说明中，有疾病和症状、临床发现、测试结果等类型的实体. 如果忽略这些特定领域的标签，那么基于深度学习的提取质量将下降. 因此，标注模块为医生提供了标注患者的 EMR 界面.

该界面能够加载患者 EMR 并向医生呈现这些临床记录. 在界面的左侧列出一些预定义的实体和关系标签. 实体标签包括疾病诱因、患病期、疾病名称、胸痛部位、伴随症状、放射部位、药物名称等. 除了这些预定义的标签，医生还可以根据自己的需要定制自己的标签. 通过预先准备好这些标签，医生可以在文本中选择单词或短语，并为其分配适当的标签. 医生还可以从 EMR 中选择实体对，并为该实体对分配关系标签，然后可以将实体和关系标注的结果分别添加到实体和关系库中以支持实体和关系提取.

为了节省医生的时间和精力，该模块应与实体提取和关系提取模块配合. 数据工程师可使用机器学习模型，如 CRF（conditional random field）和 CNN-LSTM（convolutional neural network-long short-term memory），从临床记录中自动提取信息. 医生可以专注于模型的结果，并为模型生成训练材料.

3.4　规则库模块

为了支持图谱构建过程，需要医生生成两种类型的规则：一种是从 ER 模型映射到 RDF 模型的规则，另一种是提取规则.

实例图谱是基于 RDF/OWLS（web ontology language semantics）模型进行的. 但是，目前 EMR（无论是公共数据集还是私有数据集）都以 ER 模型存储在关系数据库中. ER 模型不适合对图结构进行表示，需要被转换为 RDF/OWLS 模型.

如图 7 所示，左侧是来自一名患者的 EMR 的一部分，有 6 种类型的心音，医生在患者的症状后面做标记. 右上表示直接的 ER 到 RDF 映射结果，直接将此 ER 模型映射到 RDF/OWLS 可能会导致 RDF/OWLS 极其复杂. 然而，利用医生定义的映射规则，映射结果（右下）可以变得更加简单且更有意义. 所有 6 种类型的心音被分配给一个称为“心音类型”的属性，6 种类型的心音成为这一属性的值.

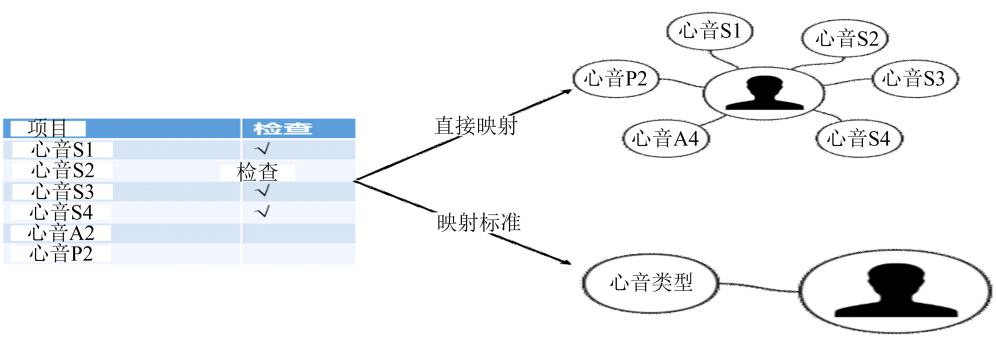

图 7　ER 到 RDF 的映射过程

实体提取有两种方法:一种是基于序列注释方法,另一种是基于规则和模式的方法.基于机器学习的序列注释方法在实体和关系提取中取得了不错的结果.然而,基于规则和模式的提取因其灵活性而成为序列注释方法的一个补充.因为医生的要求经常发生变化,这种灵活性在医学领域尤为重要.通过为医生提供定制规则和模式的界面,可以使医生将精力更多地集中在更有意义的工作上.如果医生想要更多地关注患者的症状,可以定制一些表达式,如"表现出 * 的症状","*"用作通配符以匹配表示症状的单词/短语;具有匹配和提取功能的 NLP(natural language processing)工具(如 spaCy,jieba)也可被应用到临床记录上.

4 结束语

论文介绍了一个关于医疗健康知识图谱构建的系统.构建过程的关键是将医生先验知识和相关工作与自动化方法相结合,以实现准确性和效率之间的平衡.将来,作者希望能够构建包含事件节点的知识图谱.构造事件节点的过程类似论文中提到的实例节点和概念节点的构造,有助于提高事件图谱构造的质量.

参考文献:

[1] WU C C, YEH W C, HSU W D, et al. Prediction of fatty liver disease using machine learning algorithms [J]. Computer Methods and Programs in Biomedicine, 2019, 170 (1): 23-29.

[2] MARTÍNEZ-RODRÍGUEZ J, LÓPEZ-ARÉVALO I, RIOS-ALVARADO A B. Open IE-based approach for knowledge graph construction from text[J]. Expert Syst Appl, 2018, 113 (1): 339-355.

[3] WANG C, MA X, CHEN J. Information extraction and knowledge graph construction from geoscience literature[J]. Computers & Geosciences, 2018, 112 (1): 112-120.

[4] 王华,胡学钢. 基于关联规则的数据挖掘在临床上的应用[J]. 安徽大学学报(自然科学版), 2006 (2): 21-25.

[5] ROTMENSCH M, HALPERN Y, TLIMAT A, et al. Learning a health knowledge graph from electronic medical records[J]. Scientific Reports, 2017, 7 (1): 59-94.

[6] CHEN P, LU Y, ZHENG V W, et al. Know edu: a system to construct knowledge graph for education[J]. IEEE Access, 2018, 6 (1): 31553-31563.

[7] HYEON J, OH K, KIM Y J, et al. Constructing an initial knowledge base for medical domain expert system using induct RDR[C]//International Conference on Big Data and Smart Computing, 2016: 408-410.

[8] SAVOVA G K, MASANZ J J, OGREN P V, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications[J]. Journal of the American Medical Informatics Association, 2010, 17 (5): 507-513.

[9] GATTA R, VALLATI M, LENKOWICZ J, et al. Generating and comparing knowledge graphs of medical processes using pMineR[C]//Proceedings of the Knowledge Capture Conference, 2017: 36-40.

[10] AFZAL M, HUSSAIN M, KHAN W A, et al. Knowledge button: an evidence adaptive tool for CDSS and clinical research[C]//IEEE International Symposium on Innovations in Intelligent Systems and Applications (INISTA), 2014: 273-280.

[11] TYAGI S, BHARADWAJ K K. A hybrid knowledge-based approach to collaborative filtering for improved recommendations[J]. KES Journal, 2014, 18 (2): 121-133.

[12] AMARAL A D. Rule-based named entity extraction for ontology population[C]//Recent Advances in Natural Language Processing, 2013: 58-62.

[13] YANG Y, KANDOGAN E, LI Y, et al. A study on interaction in human-in-the-loop machine learning for text analytics[C]//Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces, 2019: 147-153.

[14] HOLZINGER A. Interactive machine learning for health informatics: when do we need the human-in-the-loop? [J]. Brain Informatics, 2016, 3 (2): 119-131.

[15] SILVA T, MAGALH R P, MACÊDO J, et al. Improving named entity recognition using deep learning with human in the loop[C]//Advances in Database Technology-22nd International Conference on Extending Database Technology, 2019: 594-597.

(责任编辑 朱夜明)