大数据环境下高血压知识库构建与系统集成方法研究*

刘忠宝 张志剑

（云计算与物联网技术福建省高等学校重点实验室（泉州信息工程学院） 泉州 362000） （中北大学软件学院 太原 030051）

赵文娟

（云计算与物联网技术福建省高等学校重点实验室（泉州信息工程学院） 泉州 362000）

〔摘要〕 针对多源异构数据提出一种数据驱动、自底向上、启发式的知识库构建方法。阐述其数据采集、本体库构建、知识抽取、知识融合、知识存储以及知识更新等方面，为知识图谱相关研究提供有益参考。

〔关键词〕 知识库；高血压；深度学习；大数据环境

〔中图分类号〕R－056 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673－6036.2020.10.006

Study on the Building of Hypertension Knowledge Base and System Integration Method under the Big Data Environment

LIU Zhongbao, *Key Laboratory of Cloud Computing and Internet－of－Things Technology*, *Fujian Province University* (*Quanzhou University of Information Engineering*), *Quanzhou 362000*, *China*; *ZHANG Zhijian*, *School of Software*, *North University of China*, *Taiyuan 030051*, *China*; *ZHAO Wenjuan*, *Key Laboratory of Cloud Computing and Internet－of－Things Technology*, *Fujian Province University* (*Quanzhou University of Information Engineering*), *Quanzhou 362000*, *China*

〔**Abstract**〕 The paper proposes a data－driven, bottom－up, heuristic approach to build the knowledge base for multi－source heterogeneous data. It expounds the data acquisition, building of the ontology base, knowledge extraction, knowledge fusion, knowledge storage and knowledge update, etc., provides useful references for studies related to knowledge graph.

〔**Keywords**〕 knowledge base; hypertension; deep learning; big data environment

1 引言

近年来随着我国经济飞速发展，工作压力增大、生活节奏加快、不健康生活方式等原因导致高血压呈现出“井喷”发展态势[1－2]。高血压起病隐匿、病程长且致死、致残率高，严重影响患者劳动能力和生活质量，极大增加社会成本和家庭经济负担[3]。高血

〔修回日期〕 2019－10－16

〔作者简介〕 刘忠宝，博士，教授，发表论文 50 余篇。

〔基金项目〕 福建省社会科学规划项目“大数据环境下面向图书馆资源的跨媒体知识服务研究”（项目编号：FJ2019B052）。

压疾病名称较多，特征和关系复杂，若能结合高血压疾病专家经验，建立高血压知识库，会给高血压诊断和治疗带来很大便利。鉴于此，本文利用深度学习算法对大数据环境下的高血压知识库构建和系统集成方法进行研究，以期为高血压诊断和治疗提供技术支持，为知识库相关研究提供新思路。

2 高血压知识库构建框架

2.1 知识图谱

高血压知识库包括高血压本体库和知识图谱。常见的知识图谱构建主要有自顶向下和自底向上两种方式[4]。其中自顶向下是使用高质量数据人工或自动提取本体和模式信息，继而构建知识图谱；自底向上则是借助一定的技术手段从大数据中提取出知识信息，创建知识图谱后再构建本体库。

2.2 本体库

高血压知识库的构建往往缺少成熟的本体库。传统本体库依赖领域专家构建，然而随着数据规模不断增大，人工构建方式越发困难，亟需引入本体库自动构建技术。鉴于此，提出一种数据驱动、自底向上、启发式的知识库构建方法。该方法的基本流程是：从不同数据源采集数据，对数据进行一定预处理；自动构建高血压本体库，其主要作用是指导生成高血压知识库；基于高血压本体库，根据数据存储类型进行知识抽取，将知识进行融合，进而生成知识图谱；利用主题模型对知识图谱进行主题抽取，生成新的本体，进而更新高血压本体库；基于更新后的高血压知识库再次进行知识抽取，生成更为精细的知识图谱，进一步完善高血压知识库，如此迭代，直至得到满足要求的高血压知识库和本体库。具体流程，见图 1。

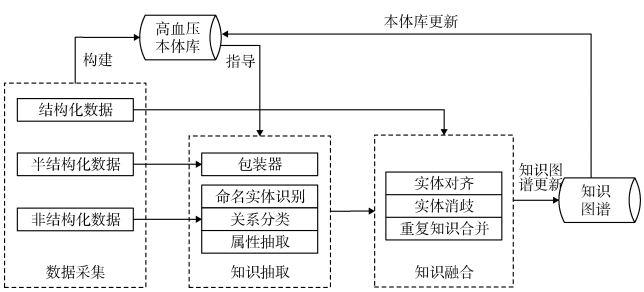

图 1 高血压知识库构建流程

3 数据采集与预处理

3.1 数据来源

高血压知识库数据来源主要包括非结构化数据：Pubmed 论文摘要、成人高血压诊断与治疗指南、Studentdoctor 论坛数据；半结构化数据：维基百科网站和成人高血压诊断与治疗指南的图表；结构化数据：中文通用知识图谱 CN－DBpedia。由于知识图谱以 3 元组形式组织，实体对齐后即可存入知识图谱。

3.2 词向量构建

词向量是词的一种特征表示，使用向量表示每个词[5]，是很多自然语言处理任务的基础步骤。通过词向量可以计算空间距离，以表征文本语义空间上的相似度。根据语义相似度可以实现实体对齐。利用 Word2vec 方法对文本信息进行低维稠密的向量表达。停用词通常是一些出现频率高但无实际语义信息、无区分度的词[6]。由于英文文本中含有大量的停用词（如 able，about，above），去除停用词将有助于提高检索速度、减少存储空间，且不影响检索结果的准确性。采集到的数据中包含众多领域词（如 ace inhibitors，beta blocker，carotid artery），通过构建领域词表指导分词，可以确保分词过程中领域词的完整性。

3.3 分句

句子是构建高血压本体库和知识抽取的基本单位，中文能够直接以标点符号进行分句，而英文中的标点符号分为无歧义和有歧义标点符号两种。无歧义标点符号包括分号、感叹号、问号等；有歧义标点符号主要为：“.”。其在英文中不仅表示句号，还表示小数点、简写符号等。利用正则匹配对非结构化数据分句，例如无歧义标点符号表示句的结束，则分句；若“.”两侧均为数字，判定为一个浮点数，不进行分句；若“.”左侧为"Mr"或"Ms"，判定为简写符号，不进行分句。

3.4 词形规范化

有两种形式：词干提取和词形还原[7]。词干提取采用缩减策略，抽取词语词干部分，无法保证词语完整性及语义一致性。例如"airliner" 经词干提取得到"airlin"。词形还原采用还原策略，将词语转化为原形形式，得到的词语具有良好的完整性。例如"driving" 经词形还原得到"drive"。利用自然语言处理工具 NLTK 可以实现词形还原，从而提高知识库构建和查询效率。

4 构建高血压本体库

4.1 方法

领域本体包含领域概念、语义关系、公理以及推理规则，通过本体库不但可以对知识抽取进行有效监督，还可以通过逻辑推理对深层知识进行挖掘，是知识组织的有效方式，也是构建知识库的重要环节[8]。领域本体库构建方法主要分为两类：人工和自动构建。随着知识更新频率日益加快，领域专家知识存在盲区，传统人工构建知识库方法耗时耗力且效率低下。自动构建本体库是利用机器学习和统计学方法对海量数据进行处理，进而得到领域本体库，该本体库含有一定噪声，本体质量难以保证。因此提出一种数据驱动的高血压本体库构建方法。

4.2 步骤

对数据进行清洗，利用 Stanford NLP 工具生成初始知识图谱，继而利用分层狄利克雷过程（Hierarchical Dirichlet Processing，HDP）模型提取相关主题，在领域专家参与下生成层级本体，将层级本体相连构成本体库。基本步骤如下：令“高血压”为知识图谱根结点以及本体库顶层结点；从采集到的数据集中查找与根节点步长为 1 的结点所在的句子，得到句子集合；利用 HDP 主题模型对句子集合中的句子进行主题抽取，得到第 2 层主题集合；在领域专家参与下对第 2 层主题集合进行筛选和归纳，进而得到第 2 层本体；将第 2 层本体存入本体库；循环步骤 2－5，直至覆盖知识图谱中的所有结点。

5 知识抽取

5.1 概述

知识抽取是知识库构建的关键步骤。本文将知识表示为｛实体，关系 & 属性，实体｝的 3 元组形式。结构化数据可直接处理为 3 元组形式，从 CN－DBpedia 抽取出与高血压相关词条，调用 Google Translate API 将中文词条翻译为英文。非结构化和半结构化数据需要单独处理。

5.2 非结构化数据

5.2.1 概述 非结构化数据知识抽取分为命名实体识别、关系分类及属性抽取 3 部分。利用双向长短时记忆神经网络（Bi－directional Long Short－Term Memory，Bi－LSTM）对非结构化数据进行处理，其原因是 Bi－LSTM 能够很好地捕捉文本中前向和后向语义特征。

5.2.2 命名实体识别 目的是识别出文本中的人名、地名、组织机构名、时间、日期等。基本流程是：将词向量作为输入，通过 Bi－LSTM 抽取文本中的语义特征，利用条件随机场（Conditional Random Field，CRF）对 Bi－LSTM 所得特征进行约束，最终得到全局最优标签序列。网络结构，见图 2。其中数据标签采用 IOB 格式，I 表示内部实体，O 表示外部实体，B 表示实体的开始词汇。实体标签共分为 5 类｛人名：PER，地名：LOC，组织机构：ORG，其他类别实体：MISC，非实体：O｝。

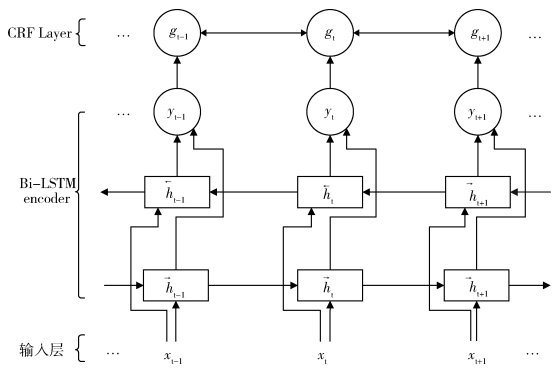

图 2 Bi－LSTM＋CRF 网络结构

5.2.3 关系分类及属性抽取 用于识别两个命名

实体之间的关系及属性，其分类效果直接影响上层应用准确性。属性可视为实体与属性值之间一种名词性关系，因此可将属性抽取任务转化为关系抽取任务。利用引入注意力机制的 Bi－LSTM（Att－BiLSTM）模型进行关系分类及属性抽取。基本流程是：将预训练的词向量作为输入，利用 Bi－LSTM 抽取文本中的高层语义特征，注意力层通过引入权重向量将词级特征合并为句级特征，以此捕捉句子的深层语义特征，利用特征分类器得到两个实体之间的关系。关系分类及属性抽取基本流程，见图 3。

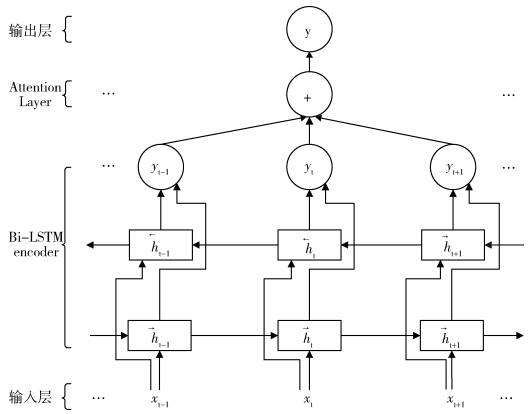

图 3 关系分类及属性抽取基本流程

5.3 半结构化数据

面向半结构化数据的知识抽取利用包装器。包装器是一种基于规则的文本信息抽取模型，其规则集合易于建立且抽取精度高，适用于半结构化数据的知识抽取。包装器基本工作流程是：首先根据输入数据从规则库中选择对应的规则，将规则传入规则执行模块；然后将规则执行模块中的规则应用于输入数据，抽取出有用信息；最后将上述信息传入信息转换模块中，将传入的信息转换为特定格式的知识。包装器工作流程，见图 4。

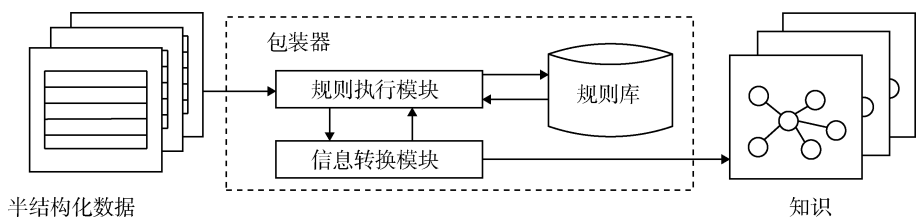

图 4 包装器工作流程

6 知识融合与知识存储

6.1 知识融合

6.1.1 概述 知识融合是知识图谱构建过程中重要步骤。通过知识融合可将知识图谱内部实体进行精简，使知识图谱运转更加有效。知识融合包括实体对齐、实体消歧和重复知识合并 3 方面内容。

6.1.2 实体对齐 也称为共指消解、实体匹配、实体同义，用于解决多个指称对应同一实体对象问题。利用实体对齐可将多个指称项关联到统一实体对象，以便将语义网络中的分散实体互联起来。本文通过计算 Word2vec 模型的词间空间距离，以此代表词间语义相似度，设定相似度阈值来划分本体间的关系，以此得到待对齐实体。

6.1.3 实体消歧 可以消除同名实体产生的歧义。由于目标实体概念集合不确定，采用基于聚类的命名实体消歧。将指向目标实体的指称项取出并聚在同一个类别下。每个类别包含某一个命名实体的所有可能指向的指称项。根据命名实体间的特征相似度，利用聚类算法决定实体对应的类别。

6.1.4 重复知识合并 多种来源数据虽然保证知识全面性，但也导致较大概率出现知识重叠。重复知识不仅会增加系统运行负担，还会使查询时间变长，效率降低。在知识存储前需要将重复知识合并，以此降低系统冗余，提高系统运行效率。

6.2 知识图谱存储

经知识融合后，高血压知识图谱构建基本完成。下一步需要将知识图谱进行存储。和传统数据库相比，图数据库在海量节点存储、管理、可视化、推理等方面具有很高灵活性、敏捷性和扩展

性。常用的 3 种图数据库相关指标对比，见表 1。其中 Neo4j 图数据库较其他两类数据库性能更优。利用 Neo4j 图数据库来存储高血压知识图谱。将所有知识存入 Neo4j 图数据库。利用 Javascript 可视化库 ECharts 得到结果，见图 5。

表 1　常用图数据库对比

对比项	Neo4j	OrientDB	JanusGraph（Titan）
ACID	支持	支持	不支持
扩展性	中	低	高
性能	高	中	高
支持编程语言数	13	12	3
多标签	支持	不支持	不支持
启动及维护	简单	简单	复杂

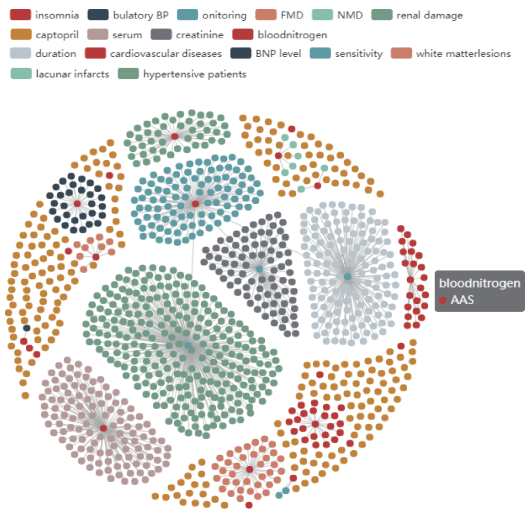

图 5　高血压知识图谱

7　知识图谱更新

7.1　迭代策略（图 6）

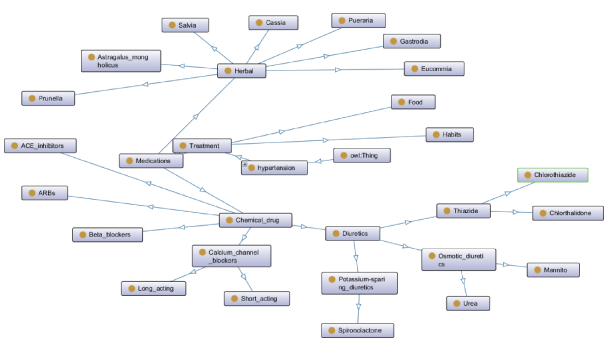

图 6　局部高血压本体库

知识抽取过程依赖本体库监督，但本体库并不完备，依据本体库生成的知识图谱规范性和完整性较差，关系及属性种类较少，无法满足实际需求。因此引入迭代策略以进一步完善知识图谱与本体库。迭代策略分为两种：整体和层次迭代。整体迭代策略是利用本体库对知识图谱进行完善，更新后的知识图谱再对本体库进行更新。层级迭代策略则是利用主题模型对数据关系及属性进行更深层、更细致地挖掘。从数据角度看，整体迭代策略基于知识图谱层次结构，按步数累加直至覆盖所有结点，进而实现本体库的更新。利用更新后的本体库进行知识抽取为知识图谱得到更丰富的关系及属性。层级迭代策略从数据本身对关系及属性进行深层次细分，得到更多关系及属性。利用 Protégé 工具对本体库进行可视化及管理，与高血压治疗有关的局部本体库。

7.2　迭代流程

首先基于更新后的本体库对知识抽取进行监督，得到更多关系及属性；其次对新知识进行知识融合；再次更新知识图谱；最后更新本体库。层级迭代流程：令“高血压”为顶层关系及属性，标记为 R_1；对采集到的数据集进行主题抽取，在领域专家的参与下得到第 2 层关系及属性集合 R_2；重新标注 R_2 中的关系及属性，训练 Att－BiLSTM 模型，面向所有数据集进一步提取其中的关系及属性；找到与 R_2 中每个关系及属性对应的句子，对其进行主题抽取，在领域专家的参与下得到第 3 层关系及属性；重新标注第 3 层关系及属性，训练 Att－BiLSTM 模型，进一步提取该层关系及属性；循环步骤 4－5，直至覆盖 R_2 中所有关系及属性，得到最终的第 3 层关系及属性集合 R_3；循环步骤 4－6，直至生成满足实际需求的知识图谱。

7.3　更新机制

7.3.1　概述　高血压知识图谱并非一成不变，随着时间推移会有新知识产生、旧知识消亡、错误知识更正等。因此有必要建立知识库的动态感知和更新机制。根据更新周期可将更新机制分为局部和全

局两类。

7.3.2 局部更新 对近期产生的新数据采用局部更新策略。将这些经过预处理的新数据输入到训练好的模型或定义好的规则内进行知识抽取。所抽取知识经过知识融合后存入知识图谱，完成一次局部更新。还可以按照新闻热搜词进行更新，当新闻热搜词出现和高血压相关度高的新闻时，可直接跳过周期限制，以该词汇在数据源中进行查询匹配，将所得数据进行一次局部更新。局部更新响应快，灵活性高，资源消耗少，是知识库更新的主要手段。

7.3.3 全局更新 对一段时间以来产生的数据采用全局更新策略。该策略以采集数据为基础重新对模型进行训练，对规则进行重新定义。将数据传入更新后的模型和规则进行知识抽取和融合，生成知识图谱并存入图数据库，完成一次全局更新。可以更新知识、降低冗余、提高查询效率，还可以标注新的实体标签、关系及属性，为上层应用提供更丰富的数据支持，是知识库更新的重要手段。综上所述，在实际应用中根据需要将局部和全局更新机制混合使用，能够有效提高知识库更新效率。

8 结语

高血压严重威胁人们健康，构建高血压知识库可为其诊断和治疗提供重要依据。面向多源异构数据提出一种数据驱动、自底向上、启发式知识库构建方法。重点探讨数据采集、本体库构建、知识抽取、知识融合、知识存储以及知识更新等问题。但研究还面临一些挑战，如深度学习模型虽然在知识抽取中表现优良，但其性能依赖于人工标注数据规模，如何利用自动化手段实现标注值得关注；医疗实践对高血压知识库正确性有极高要求，通过自动化技术生成的高血压知识库准确性不足，如何进一步提高知识库构建效率和精度值得深入探讨。

参考文献

1 《国际医药卫生导报》编辑部. 慢性病的治与防［J］. 国际医药卫生导报, 2018, 24（15）: 2219.

2 中国医疗保健国际交流促进会高血压分会, 中国医师协会高血压专业委员会, 中国老年医学学会高血压分会. 中国高血压防治指南（2018 年修订版）［J］. 中国心血管杂志, 2019, 24（1）: 1－44.

3 刘力生. 中国高血压防治指南 2010［J］. 中国医学前沿杂志（电子版）, 2011, 3（5）: 42－93.

4 吴运兵, 阴爱英, 林开标. 基于多数据源的知识图谱构建方法研究［J］. 福州大学学报（自然科学版）, 2017（3）: 329－335.

5 李枫林, 柯佳. 词向量语义表示研究进展［J］. 情报学报, 2019, 37（5）: 155－165.

6 俞琰, 赵乃瑄. 专利文本主题建模中领域停用词自动选取研究［J］. 图书情报工作, 2018, 62（11）: 120－126.

7 吴思竹, 钱庆, 胡铁军. 词形还原方法及实现工具比较分析［J］. 数据分析与知识发现, 2012, 28（3）: 27－34.

8 张文秀, 朱庆华. 领域本体的构建方法研究［J］. 图书与情报, 2011, 40（1）: 16－19.