文章编号：2095-2163(2020)12-0204-05 中图分类号：TP391.1 文献标志码：A

基于带权知识图谱的智能预问诊系统的研究

徐霄玲，郑建立，邵奕琛，李浩东
（上海理工大学 医疗器械与食品学院，上海 200093）

摘 要：人们日益增长的就医需求和医疗资源的短缺不仅造成医生工作强度大，还极易引起医患沟通不充分，延误病情等问题。随着人工智能技术不断发展，知识图谱以三元组形式组织知识的结构化框架不断被认可，亦被应用到各个领域的知识库中。本文研究设计了基于带权知识图谱的智能预问诊系统，以期能提高医患沟通效率。首先，构建以疾病为中心的知识图谱模式层，来组织疾病与症状、检查和治疗多方面的知识；其次，从专业书籍中根据模式层提取知识，然后基于 XGBoost 训练分类器和概率统计获取各个关系之间 ω_{ds} 和 ω_{sd} 权重，并将所有信息存储至 neo4j 图数据库；最后，根据 ω_{ds} 、ω_{sd} 和节点入度计算得到总权重 W，设计问诊模型，完成预问诊系统。

关键词：知识图谱；XGBoost；neo4j；预问诊

Research of intelligent pre-consultation system based on weighted knowledge graph

XU Xiaoling, ZHENG Jianli, SHAO Yichen, LI Haodong
(School of Medical Instrument and Food, University of Shanghai for Science and Technology, Shanghai 200093, China)

【Abstract】 The increasing demand for medical treatment and the shortage of medical resources not only result in the heavy work intensity of doctors, but also easily lead to the problems of inadequate communication between doctors and patients and delay of illness. With the continuous development of artificial intelligence technology, the structured framework of knowledge graph, which organizes knowledge in the form of triples, has been recognized and applied to the knowledge base in various fields. In this paper, an intelligent pre-inquiry system based on weighted knowledge graph is designed to improve the efficiency of doctor-patient communication. Firstly, the knowledge map pattern layer centered on disease is constructed to organize the knowledge of disease and symptoms, examination and treatment. Secondly, knowledge was extracted from professional books according to the pattern layer, and then based on XGBoost training classifier and probability statistics, ω_{ds} and ω_{sd} weights of each relationship were obtained, and all the information was stored in the Neo4j graph database.

Finally, the total weight W was calculated according to ω_{ds} 、ω_{sd} and node entry degree, and the inquiry model was designed to complete the pre-inquiry system.

【Key words】 Knowledge graph; XGBoost; Neo4j; Pre-consultation

0 引 言

问诊是指医生采用对话的方式，向患者了解其疾病发生发展经过、现有症状和既往健康状况等基本情况的一种病史采集手段，问诊结果的准确性和完整性对后续的检查与诊断都具有重大意义。现阶段医疗资源缺乏，就医需求不断提高，导致单个病人的问诊时间不断缩短，可能出现医患沟通不充分、病情了解不全面等问题，既延误病情，又可能加剧医患矛盾。为了改变问诊现状，依托知识图谱等智能技术手段，实现预问诊，以优化就医流程。预问诊实际就是诊前问诊，利用患者候诊的时间，通过图谱中的知识，模拟医患问诊流程，收集主诉、检查、治疗等信息，形成初步的电子病历，以供医生了解病情。

知识图谱由语义网发展而来，是结构化的语义知识库，2012 年谷歌正式提出后受到了学界广泛关注[1]。知识图谱通过三元组对现实世界的事物及其关系进行形式化描述，已经被广泛当作问答系统的知识来源，其中三元组的形式包括实体-关系-实体、实体-属性-属性值。目前，世界范围内高质量大规模开放领域知识图谱主要有 YAGO、DBpedia、Freebase 等。中文方面则在 OpenKG 联盟的推动下，出现了以 Zhishi.me、CN-DBPedia 为代表的开放知识图谱[2]。但在垂直领域如医疗，相关的知识图谱资源不多，且多以医学文献和网络数据为主[3]。

本文研究的主要内容是基于带权知识图谱的智能预问诊系统。首先，使用正式出版的专业医学书籍作为数据源，构建以疾病为中心的内科疾病带权知识图谱，并将其存入图数据库 neo4j，以供预问诊使用；其次，建立疾病预问诊系统，充分将自然语言处理技术与知识图谱技术结合，尽可能全面地了解

作者简介：徐霄玲(1994-)，女，硕士研究生，主要研究方向：自然语言处理；郑建立(1965-)，男，博士，副教授，主要研究方向：医学信息集成；邵奕琛(1999-)，男，本科生，主要研究方向：自然语言处理；李浩东(1999-)，男，本科生，主要研究方向：自然语言处理。

收稿日期：2019-12-26

患者当前情况，生成可供医生查阅的电子病历。

1　知识图谱构建

知识图谱自上而下可分为模式层和数据层。模式层是图谱的核心，利用抽象概念对知识数据进行约束，由实体和实体关系组成。通常情况下使用本体库管理模式层，故也称为本体。数据层则指的是具体知识，与模式层相互映射。知识图谱构建方式有自顶向下和自底向上[4]。自顶向下指的是从现有高质量结构化数据（如百度百科等）直接提取模式层，并将其加入图谱；自底向上则指的是对非结构数据运用技术手段归纳总结后，最终得到模式层的方法。本次研究是二者方式的结合，先参考 Schema.org 构建，通过自底向上的方式填充知识图谱。

1.1　模式层设计

此次构建的知识图谱是为医疗问诊服务的，所以需要参考实际问诊过程中医生期望得到的信息来组织图谱。根据医生经验反馈，问诊过程中，问诊人往往需要引导患者说出自身当前症状，以及针对当前症状做过的检查、用过的药，曾经所患疾病、过敏史等。因此将疾病作为图谱的中心，通过建立疾病与症状、检查、治疗等多方面实体的事实关系，完成图谱构建。

医学实体类别共分为 5 大类，分别为疾病、症状、检查、治疗和病因。检查大类下包含体格检查、实验室检查、影像学检查和辅助检查 4 个子类；治疗大类下则细分了手术、药物和饮食 3 个小类。同时为实体增加了别名、时相、多发人群、定性值、多发地区、发病部位、否定词、性质、热型、时间等十大属性。

知识图谱通过医学实体关系将医学实体联系起来，表达出真实的医学知识。根据医学领域专家的经验，并结合医学书上的专业知识，总结出以下 12 种实体关系和 1 种语义关系属性：

（1）E_1 has_cause E_2 关系：表示实体 E_2 是实体 E_1 的病因；

（2）E_1 has_subclass E_2 关系：表示实体 E_2 是实体 E_1 一个子概念；

（3）E_1 typing_staging_of E_2 关系：表示实体 E_2 是实体 E_1 的分型分期；

（4）E_1 differential_disease_from E_2 关系：表示诊断时注意鉴别实体 E_2 与实体 E_1，用于区分临床上容易混淆的疾病；

（5）E_1 secondary_to E_2 关系：表示实体 E_1 继发于实体 E_2；

（6）E_1 has_symptom E_2 关系：表示实体 E_2 是实体 E_1 的症状；

（7）E_1 has_complication E_2 关系：表示实体 E_2 是实体 E_1 的并发症；

（8）E_1 has_examination E_2 关系：表示实体 E_2 是针对实体 E_1 的检查；

（9）E_1 has_treatment E_2 关系：表示实体 E_2 是针对实体 E_1 的治疗；

（10）E_1 past_history E_2 关系：表示实体 E_2 是针对实体 E_1 的既往病史；

（11）E_1 family_history E_2 关系：表示实体 E_2 是针对实体 E_1 的家族史；

（12）E_1 belong_to E_2 关系：表示实体 E_1 属于实体 E_2，此关系是为了避免多跳过程中出现歧义而设置的。

根据以上定义，在模式层设计中对不同关系所连接的实体类别做了限制，如鉴别诊断关系（differential _ disease _ from）、既往史关系（past _ history）、家族史关系（family _ history）、并发症关系（has_complication）必然连接两个疾病实体，症状关系（has_symptom）连接疾病和症状实体等。结果关系（rel_result）存在于上述（8）和（9）中，用以记录与疾病相关检查和治疗的结果。完整的内科疾病知识图谱模式层结构，如图 1 所示。

1.2　权重确定

基于知识图谱的问诊过程，实际上就是根据患者的回答对图谱中的知识不断查询并不断排除的过程。为了减少不相干信息的提问，要求问诊系统具备能根据患者情况动态定位关键知识并提问的能力。为了实现上述功能，为实体关系赋予了权重。考虑到不同症状和既往疾病等对某一疾病重要程度不同，同一症状和既往疾病等对鉴别不同疾病也存在强弱关系，因此实体关系需要拥有两个权重，分别为 ω_{ds} 和 ω_{sd}。本次权重研究以来源于上海市某三甲医院的内分泌科住院病历数据为例，使用 XGBoost 构造 I 型 II 型糖尿病分类器，并结合临床病历数据中的统计概率最终得到上述两个权重。

XGBoost 是在 GBDT 算法上的改进。由于其效率高，效果显著且对输入要求不高，被广泛应用于各大比赛的分类任务中。从住院病历中提取出性别、年龄、BMI 值、体温（T）、脉搏（P）、呼吸（R）、收缩压（SBP）、舒张压（DBP）、既往史、家族史和主诉信息，并使用填充缺失值、one-hot 编码、标签映射等方法进行数据预处理，最终得到 234 份包含 165 维特征

的数据集。使用处理好的数据训练 XGBoost，调参后得到最优模型。模型中特征属性对模型的重要性得分，见表 1。表中的特征分别为糖尿病（家族史）、性别、BMI、年龄、体温、收缩压、血糖升高（主诉）、口干（主诉）、高血压（既往史）、舒张压和呼吸频率。未包含在表 1 中的特征代表值为 0。

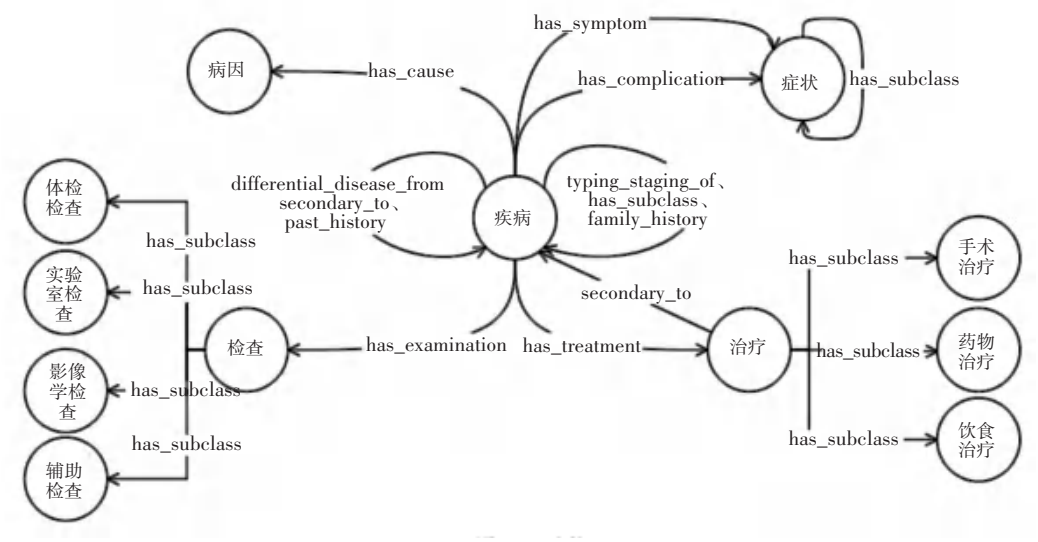

图 1 模式层结构图

Fig. 1 Structure diagram of model layer

表 1 XGBoost 特征重要性得分

Tab. 1 Characteristic importance score from XGBoost

Feature	Score
Family_ diabetes	0.219
Sex	0.152
BMI	0.123
Age	0.112
T	0.096
SBP	0.082
P	0.080
Complaint_ hyperglycemia	0.047
Complaint_ xerostomia	0.038
Past_ hypertension	0.031
DBP	0.011
R	0.007 2

ω_{ds} 表示其他相关实体与某一疾病的权重，初始值为 0。对于非数值型特征，其 ω_{ds} 为表 1 中的 Score 值。数值型数据和性别，需要分段处理，例如性别有男女之分。首先需要统计计算出病历中患 I 型糖尿病中男、女的概率分别为 P_1、P_2，则 I 型糖尿病与男性实体间的 ω_{ds} 为 P_1 和 Score 的乘积，女性同理。ω_{sd} 表示某一相关实体与不同疾病间的权重，通过计算 I 型 II 型糖尿病患者中存在该实体的概率得到，其初始值也为 0。

1.3 数据层获取与存储

数据层获取就是通过一定的方法从文本中抽取出与模式层相匹配的医学知识。医学书籍中，知识均以非结构化形式展示，采用人工标注的方式从书中提取出各类实体和实体关系，并将其转化成三元组的形式。最终，得到了由 6 350 个三元组，包含 232 种常见内科疾病的知识图谱。

实体、实体关系和关系权重全部确定后，使用 Neo4j 存储。Neo4j 是开源的图数据库，有别于关系型数据库，将数据以节点、属性和边的形式进行存储，提供 Cypher 作为图查询语言，并且支持各种图查询算法，在存储和检索知识图谱方面有显著优势。以“慢性坏死性肺曲霉病亦称半侵袭性肺曲霉病，患者有长期呼吸道症状如咳嗽、咳痰等，也多有发热，常用 X 线检查”为例，图 2 中展示了该知识在 Neo4j 中最终存储结果。

2 问诊系统设计

2.1 系统架构

本次研究的预问诊系统，以微信小程序作为系统与患者的交互载体，使得预问诊系统能以和患者对话的形式进行。基于微信小程序确保系统能为各种手机机型的患者提供预问诊服务，系统架构如图 3 所示。带权知识图谱和疾病问诊模型为整个系统核心。知识图谱提供知识，权重有利于快速定位关键信息，有效缩短问诊时长，优化问诊流程。疾病问诊由输入处理模块和知识检索模块构成，前者负责将患者描述转化成知识图谱内的标准数据，后者则负责对患者情况做全面采集，最终将问诊得到的信

息转换成预问诊电子病历方便医生查阅。

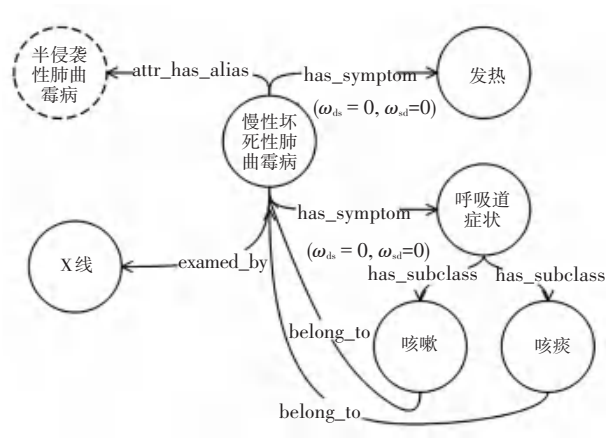

图 2　neo4j 存储结果示例

Fig. 2　The example in neo4j

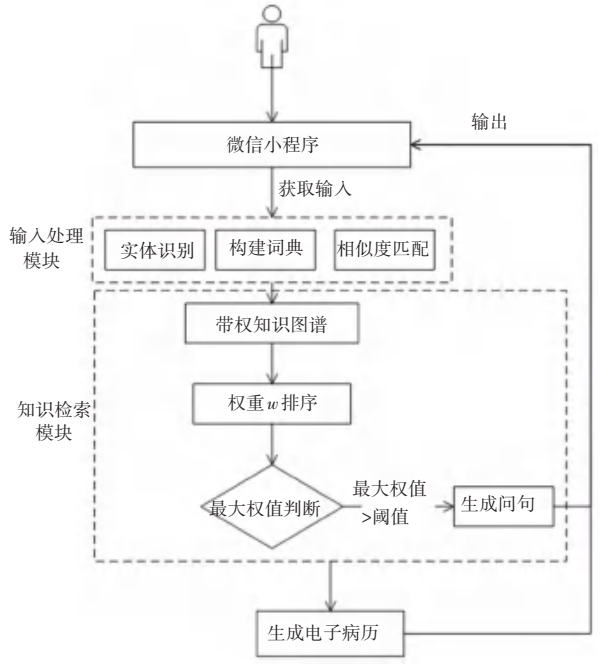

图 3　系统架构图

Fig. 3　System architecture diagram

2.2　输入处理模块

输入处理模块是为了理解患者的描述信息，从中得到其所述的症状、检查和治疗实体。首先，需要对患者描述文本进行命名实体识别。命名实体识别可以看作是一个序列标注问题，即给每个字一个合适的标签[5]。常见的命名实体识别方法有最大熵模型(ME)、隐马尔科夫模型(HMM)和条件随机场 CRF。近年来，随着深度学习技术不断发展，BiLSTM- CRF 已经成为解决该任务的通用结构。随后根据知识图谱存储的三类实体构建字典，利用 word2vec 得到字典和识别出的实体的短文本向量表示，并利用余弦相似度做相似度匹配，得到字典中与上述向量相似度>0.8 的实体或者是短文本。

2.3　知识检索模块

在知识检索和生成问诊问题的过程中，通过权值 W 确定实体的最终重要程度。实体 v 的权值 W 由 ω_{ds}、ω_{sd} 和该实体入度 $\delta(v)$ 三者共同决定，这样有利于兼顾问诊过程中与疾病相关实体的重要性和广泛性，具体如公式(1)所示。其中 n 表示与该实体 v 存在有向边的疾病数量：

$$W(v) = \sum_{i=1}^{n} (\omega_{d_i s_i} + \omega_{s_i d_i}) + 0.5 * \delta(v). \qquad (1)$$

将查询实体作为条件查询图谱，得到与该实体相关的疾病。将与这些疾病相关的实体取交集，并根据实体权值 W 对交集降序排序。根据 W 的排序，先取大，后取小，依次提问。若 $W > 0.7$，将 W 所对应的知识填充至问题模板中，对患者进行提问。若患者回答“有”，则重新确定约束条件进行下一轮图谱查询；否则，对下一个 W 值代表的实体提问；若 $W \leqslant 0.7$，则退出问诊。

问诊过程中，系统记录患者对问题的所有回答。问诊结束后，结合问题和回答，生成预问诊电子病历。

2.4　实验结果

本文选取了 20 份内分泌科电子病历，提取出其年龄、性别、BMI、主诉症状、既往史疾病和家族史疾病作为测试数据。随机选择主诉中的一个症状作为模型输入，模拟患者描述。统计系统与每个患者的对话次数，求得平均对话次数。问诊结束后，根据询问过程中问题是否完全包含病历中患者病情信息，计算问诊完整率。实验结果见表 2。

表 2　实验结果对比

Tab. 2　Comparison of experimental results

方法	平均对话次数	问诊完整率
带权知识图谱	8.5	96.4%
普通知识图谱	14.8	100%

从表 2 的结果可以看出，本文基于带权知识图谱的问诊完整率略低于普通知识图谱，但问诊平均对话次数明显减少。因此，基于带权知识图谱的问诊，能有效提高预问诊效率，提升患者使用体验。

3　结束语

本文设计的基于带权知识图谱的预问诊系统能利用患者候诊时间完成早期病情信息收集工作，尽可能提高医患沟通的效率。其中，带权图谱构建和问诊流程设计是此次研究的关键部分。知识图谱以图的方式存储知识，不仅查询效率高，而且有利于后

期知识的增加。预问诊过程中,依靠图谱中的权重,能在尽可能完善地收集患者患病情况的同时,减少无关信息提问,具有一定使用价值。但现阶段知识图谱只包含了部分常见内科疾病,内容还有待完善,如何将多个数据源的知识整合进图谱将会是下一步研究方向。

参考文献

[1] AMIT S. Introducing the knowledge graph[R]. America: Official Blog of Google, 2012.
[2] 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.
[3] 崔洁,陈德华,乐嘉锦. 基于 EMR 的乳腺肿瘤知识图谱构建研究[J]. 计算机应用与软件,2017,34(12):122-126.
[4] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述[J]. 计算机研究与发展, 2016, 53(3):582-600.
[5] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3):329-340.

(上接第 203 页)

本文算法明显。本文算法对暗光图像处理后,使得图像增强后细节信息更加完整,原始结构信息更加丰富,与人眼所能感受到的真实景像更加接近,图片的质量和色彩也变得更好。

2.2 客观评价:

为了更好的评价本文算法对低照度图像的增强性, 使用峰值信噪比(*PSNR*) 来对增强后的低照度图像质量进行衡量,*PSNR* 是由 *MSE* 计算而来,用来一个衡量图像失真或是抗噪声水平,其值越大失真越小,抗噪水平越高。*PSNR* 是目前常用和公认的一种图像客观评价指标。

关于 *PSNR* 的计算公式如下:

$$MSE = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} (X(i,j) - Y(i,j))^2, \quad (18)$$

$$PSNR = 10 \log_{10} \left(\frac{(2^n - 1)^2}{MSE} \right). \quad (19)$$

其中,*MSE* 表示均方误差;*MSE* 的计算见公式中的 *H* 和 *W* 代表输入图像中的长和宽;*PSNR* 是通过 *MSE* 得出的。

表 1 是原始图像经过 4 种方法增强处理后图像性能的客观评价标准。从表 1 的数据可以得知本文算法和自适应直方图均衡化算法所得到的 *PSNR* 比较大,但本文算法所处理过的图像更加结构清晰、色彩丰富、人眼的视觉感更强。所以总体来说,本文算法优于其它几类热门的图像增强方法。

表 1 不同增强算法的客观评价结果

Tab.1 Objective evaluation results of different enhancement algorithms

增强算法	测试指标	测试图像			
		图 3	图 4	图 5	图 4
直方图均衡化	PSNR	5.541	6.792	8.560	7.171
MSRCR	PSNR	4.137	5.768	7.496	7.0123
自适应直方图均衡化	PSNR	19.670	16.011	12.516	16.277
本文算法	PSNR	24.945	15.694	12.785	16.238

3 结束语

为了解决低照度图像的缺点,本文提出了一种以基于引导滤波的低照度图像增强算法,通过实验数据的对比分析可以得出本文算法能够很好的改善低照度图像的缺点,使低照度图像的对比度、色彩饱和度都有了显著提升,其细节和轮廓变得更加清晰,弥补了一些低照度图像增强算法对图像细节的模糊,又弥补了一些低照度图像增强算法对图像画面过度增强而引起的画面变白和色彩失真的现象,同时也使画面的色彩丰富自然,使之人眼的视觉效果和图像的质量提升。所以本文算法在增强低照度图像的实际应用中具有很深的研究价值。

参考文献

[1] 冈萨雷斯,伍兹. 数字图像处理[M]. 阮秋琦,阮宇智,译.3 版. 北京:电子工业出版社,2011.
[2] 行薇. 图像插值在图像处理中的应用[D]. 长春: 长春理工大学,2011.
[3] JIANXIN Y U, ZHANG W H, ZHIWEI Y U, et al. Single image defogging algorithm based on HSI color space[C]// International Workshop on Education Technology &Computer Science. IEEE Computer Society,2014. 909-913.
[4] 刘岚. 彩色图像增强算法的研究与实现[D]. 武汉: 武汉理工大学, 2012.12.
[5] 于天河,李昱祚,兰朝凤. 基于顶帽底帽变换的仿生图像增强算法[J]. 计算机应用,2020,40(5):1440-1445.
[6] HE Kaiming, SUN Jian, TANG Xiaoou. Guided image filtering.[J]. IEEE Trans on Pattern Analysis & Machine Intelligence 2013, 35(6): 1397-1409.
[7] 王殿伟,韩鹏飞,李大湘,等. 基于细节特征融合的低照度全景图像增强[J]. 控制与决策.2019, 34(12): 2673-2678.