《智能信息处理》课程作业

基于知识图谱的肝病问答系统研究

田延凯

作业	分数[20]
得分	

2021 年 12 月 13 日

基于知识图谱的肝病问答系统研究

田延凯

(大连海事大学计算机科学与技术辽宁省大连市中国 116026)

摘　要　肝病患者对自身病情认识不足以及早期检查方法匮乏导致在治疗时期病情已经较为严重。为改变现状，文中通过网络爬虫获取寻医问药网中的数据，构建肝病知识图谱，并在此基础上开发智能问答系统。其技术核心和关键是计算机智能辅助诊疗，利用该肝病基础知识的知识图谱，可以迅速针对患者提出的问题给出相当准确的辅助诊疗决策。该智能问答系统采用规则匹配方式完成，它将用户输入的问题进行语义分析，问题分类并进行问句相似度计算，在已建构的肝病知识图谱中查询到对应问题的答案，根据系统返回数据组装问句进行回答。基于知识图谱的肝病智能问答系统可以有效解决肝病相关的问题，节省医生和患者大量时间，具有较高的准确率和使用价值。

关键词　肝病；知识图谱；问答系统；语义解析；句子相似度

Research on the Question and Answer System of Liver Diseases Based on the Knowledge Graph

Tian Yankai

(Dalian Maritime University, Computer Science and Technology, Dalian, Liaoning Province, China 116026)

Abstract　Patients with liver disease have insufficient knowledge of their own condition and lack of early examination methods, which has led to a more serious condition during the treatment period. In order to change the status quo, this article uses a web crawler to obtain data from Xunyiwenyao.com, build a knowledge map of liver disease, and develop an intelligent question-and-answer system on this basis. The core and key of its technology is computer-assisted diagnosis and treatment. Using the knowledge map of the basic knowledge of liver disease, it can quickly make quite accurate auxiliary diagnosis and treatment decisions in response to the questions raised by the patient. The intelligent question answering system is completed by rule matching. It analyzes the questions input by the user, classifies the questions, and calculates the similarity of the question sentences. The answers to the corresponding questions are queried in the constructed liver disease knowledge graph and assembled according to the data returned by the system. Questions to answer. The liver disease intelligent question answering system based on the knowledge map can effectively solve liver disease-related problems, save a lot of time for doctors and patients, and has high accuracy and use value.

Key words　liver disease; knowledge map; question answering system; semantic parsing; sentence similarity

1 引言

肝病有多种，常见的肝病有病毒性肝炎、酒精性肝炎、非酒精性脂肪性肝病、原发性肝癌等，其中病毒性肝炎最常见的是乙肝。乙肝病毒传染性极强，遍布全球，其中母乳传播比较明显[1]。虽然中国现在脱掉了乙肝大国的帽子，乙肝治疗技术和药物研发也在不断改进，2019 年国内乙型肝炎发病率仍达 71. 77 /10 万人，死亡率达 0. 032 /10 万人。脂肪肝这一疾病患者人数也在不断攀增，因此对于健康人和肝病患者来说，及早熟悉肝病知识、匹配疾病与自身症状关系是非常重要的。

在肝病诊断时，通常需要医生与患者一对一诊疗和仪器检查的配合。如果在问医之前患者就可以了解自己大概病症，能够快速地告知医生自己的身体状况，这无疑可以减少医生的工作量。并且医生少数情况下也会受到主观因素的影响导致诊断不准确，很多患者也因相关肝病知识的匮乏导致就医不及时，病情恶化。

目前越来越多的在线医疗问答小程序出现，如“阿里医生在线咨询”、“平安好医生”、“39 健康网”等，患者可以及时在线咨询相关疾病信息，方便快捷，但是这种问答系统需要大量医生入驻小程序并上线才能响应患者的问题。针对以上问题，文中开发了基于知识图谱的肝病智能问答系统，基于垂直类医疗网站，囊括肝病绝大多数数据，为病患提供肝病自助查询服务。该问答系统及时解决患者问题，不用再等待，实现了医疗行业的智能化。

2 相关工作

2.1 问答系统发展及应用

问答系统概念提出的时间并不长，但发展迅速，已经形成了一些相对完整的体系。国内复旦大学开发的原型系统(FDUQA) 已经具有初步效果，同时哈尔滨工业大学(金山客服) 和中国科学院计算技术研究所也正在进行这一领域的研究。国外发展比较成熟，全球第一个基于互联网的问答系统，即 START 系统，使用知识仓库+信息搜索混合模型，知识库包括“START + KB”、“Internet+Public+Library”[2]。而华盛顿大学开发的第一个自动问答系统———MULDER 系统更进一步，它没有知识库，而是充分获取互联网的数据进行分析并给出一组候选答案。每个候选答案都会被赋予一个置信度，可以被用户当作参考的条件[3]。

国内的智能问答技术发展较晚，这是由于中文的语法和语义十分复杂，现在是以人工模板和智能检索技术为主，典型代表有华为小 E、小米小爱等。国际上目前的主要智能问答技术为计算机检索、知识网络、深度学习这三大技术，苹果的 Siri、微软的 Cortana 和谷歌的 GoogleNow 均十分具有代表性。与此同时，知识图谱的快速发展，为智能问答系统的实现提供了优质的知识来源，大大加速了问答系统在医疗领域的发展[4]。这项技术可以使专业人员更好地帮助用户去学习、沉浸使用真实世界中各类实体概念之间的联系。

2.2 知识图谱和知识库

知识图谱(knowledge map) 又称科学知识图谱，它在 2012 年由 Google 最先提出，并发表了基于 Freebase[5]知识库和维基百科的大规模知识图，为世界以及领域知识的构建提供了一个可借鉴的手段[6]。知识图谱的结构与图是一样的，都是由节点和边组成的，图中的节点对应知识图谱中的实体，边表示实体与实体之间的关系[7]。

知识库则比知识图谱容纳更多的知识信息，知识库中的知识有很多种不同的形式，例如本体知识、关联性知识、规则库、案例知识。知识库问答(knowledge base question answer) 任务是指利用知识库中的一个或多个知识三元组〈Sub，Re，Obj〉来回答自然语言问题。

比如提出一个自然语言问题“Where is Beijing?”可以运用＜Beijing，In China＞这一事实来回答。相比两者的概念，知识图谱更加侧重于关联性关系的构建和可视化，可借助知识推理(如规则等) 快速进行知识挖掘和推理获取新知识，发现实体或概念之间的新关联。因此，文中基于知识图谱进行肝病问答系统开发，这对于解决国内优质医疗资源供给不足和医疗服务需求持续增加的矛盾将产生重要的作用[8]。

3 构建问答系统的方法

3.1 语义分析

智能问答系统开发中所需的语义分析技术，通过使用不同的方法，了解和学习一段文字所表示的

语义内容，对语言的各种理解可以被归类为语义分析。一段文字一般由单词、句子和段落组成，根据语言单元理解对象，语义分析可以分为词汇级语义分析、句子级语义分析和章级语义分析[9-11]。

语义分析是编译过程中不可或缺的逻辑阶段，它对源程序上下文的相关性进行分析检查，保证代码结构必须正确。语义分析阶段就是检查源程序有无语义错误，为代码生成阶段收集信息。本系统主要运用句子级语义分析，主要包括浅层语义分析和深层语义分析。

3.1.1 浅层语义分析

浅层语义分析，又被称作语义单元表示(semantic unit representation) 。这一现象将深层语义分析简单化，在文章中明示与文字研究相关的语义单元。比如施事者、受事者、时间和地点状语等[12]。目前 NLP 技术和人工智能技术的成熟，可以令浅层语义分析真正在实践中发挥作用，同时对于智能问答、智能翻译等功能系统产生巨大帮助。

3.1.2 深层语义分析

深层语义分析，又可直接称作语义分析(semantic parsing) 。它会将问句翻译为一种特定形式化体现出来，不再以谓词为中心。其在中英文转换中经常使用，由于汉语和英语的外层构造之间联系密切，对应关系复杂，如果想要发现其中的规律是十分困难的，在进行研究和使用过程中必须发现两类不同语言之间共有的更深一层关系，即深层语义结构。

本系统通过语义分析有监督学习。也就是说，通过现有的训练样本(即已经知道的数据和对应的出)进行训练，获得最好的模型(该模型属于特定函数的集合，最优模型表示在特定的评价准则中是最好的)。然后利用该模型将所有输出进行相应的映射，得到结果输出，通过对该结果进行简单的判断来实现分类目的，同时也学习到了对未知数据进行分类的能力。描述了知识图谱对于用户提出的问句文本进行语义理解过程，如图 1 所示。

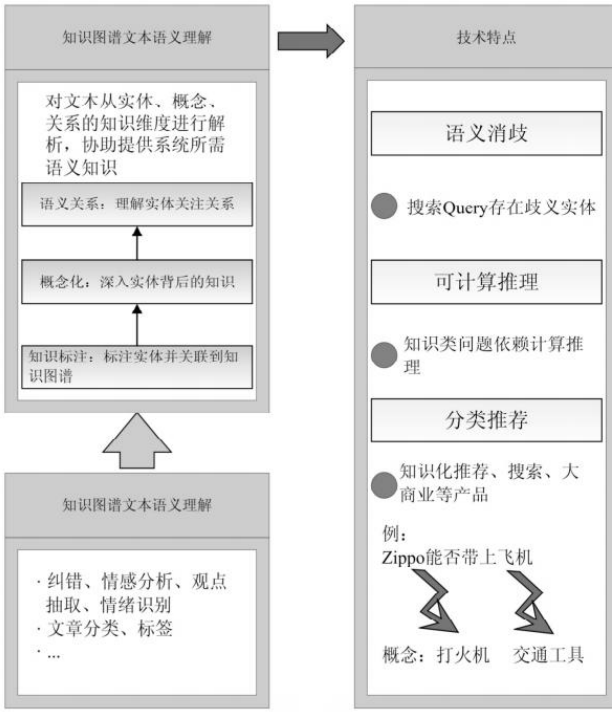

图 1 知识图谱文本语义理解

3.2 问题分类

智能问答系统在用户输入问题后，将问题分为两类，求知性问题和求证性问题。求知性提问就是在知识图谱中为用户获得他们的未知知识，填充用户问题，比如“什么是乙肝”; 求证性提问则是在用户对一个专业性问题模棱两可时，向智能问答系统寻求帮助。进行提问之前用户已经了解关于问题的部分信息，主要目的是通过该系统对不完整知识或模糊信息进行弥补、证明或肯定，比如“长期大量酗酒、过度暴饮暴食是脂肪肝的病因吗”。

从问句的形式上来说，分为疑问句、反问句、设问句和是非问句、特指问句[10-11]; 从问题目的上分为查找数据、验证事实、收集信息等; 从性质上分为封闭型和开放型问句; 按照问题复杂程度分为七大类别: 语境性问题、是非问题、推理性问题、选择问题、特指性问题、专家性问题、概括性问题。测试结果问句如表 1 所示。

对于专家性问题，目前分类不是很明确。因为该类问题应该既包括推理性问题，也会含有概括性问题，特指性也较强。问句进行分类之后，对各个类型布置特定的答案抽取规则，能够更加快速地在该阶段利用规则来获得问题最终的结果。现在市面上大多智能问答系统都是已经提前将问题归好类，但是分类方法还不是很科学，不确定因素太多，并且太繁杂，不能从根本上满足用户的需求。

因此本系统通过构建 Aho －Corasick 字符串匹配算法(AC 算法) 匹配问句中的领域词(医学本身就是一个封闭领域[11]) ，同时收集问句中提到的实体类型，确认其中是否包含该实体类型中的特定特征词，从而判断问句类型属于哪一种。AC 算法主要是解决多字符串匹配问题，采用多模式串建立

一个树形有限状态机，即先以多模式串(短字符串)为基础创建自动机[12]。以主串(长字符串)作为该自动机的输入，也是就把长字符串在状态机里面跑一遍，使状态机进行各种状态之间的转换，当到自动机达某些特定的状态时，用状态表示字符是否匹配，即某些字符串是否匹配成功。

3.3 问句相似度的计算

问答系统中，除了语义分析和问题分类十分重要外，不可或缺的还有问句相似度的计算。系统需要对用户提出的问题进行解析，并进行问句相似度计算，得出每个相似问句的最终答案。

本系统采用新的问句相似度计算方法，主要包括词形相似度、问句长度相似度、词序相似度和距离相似度等[17]。词形相似度计算方法主要是计算两个问句等词的数量，利用公式计算，两个相同的疑问句中同一词的数量越多，两个疑问句就越相似。在分析问句的语义时，仅仅关注其中的词语是远不够的，还需要考虑词语间的结构[13]。因此，在词性相似度计算方法的基础上，引入了句法特征，从全新的角度考虑对应句法成分间的词语相似度，进而衡量句子间的相似度。词性相似度计算如公式（1）所示:

$$\mathrm{WordSim}(q_1,q_2)=\frac{\mathrm{SameWc}(q_1,q_2)}{\max(\mathrm{Len}(q_1),\mathrm{Len}(q_2))} \quad (1)$$

其中，q1 和 q2 分别表示两个问句，Len( ) 表示问句的长度，SameWc( ) 表示两个问句中相同词的个数。

若问句的长度相似，足以反映两个问句形态之间的相似性，两个问句的长度相似的话，两个疑问句相似可能性很高。问句长度相似度计算如公式(2):

$$\mathrm{LenSim}(q_1,q_2)=1-\frac{\mathrm{abs}(\mathrm{Len}(q_1)-\mathrm{Len}(q_2))}{\mathrm{Len}(q_1)+\mathrm{Len}(q_2)} \quad (2)$$

其中，q1 和 q2 分别表示两个问句，Len( ) 表示问句的长度。

关键词汇的顺序也可以反映出两个问句的相似度，因为两个问句中如果相似或者相同意思的词汇在同一位置上，足可以证明问句较为相似。在一个词汇或者短语整体被大幅移动时，使用词序相似度判断可以快速准确地断定是否与原来问句相似。编辑距离是计算将一个句子换成另一个句子时所需的最少操作数，其中句子以文字为最小单元。编辑工作是指插入、删除、替换，但是以字为单位的编辑操作代价太大，便用词汇代替一个字的单词，词汇作为最基本的编辑单元参与计算，对于不同的编辑操作赋予不同的权重以代表相应的重要程度。

4 结束语

本系统基于肝病专业医疗知识建立了肝病知识图谱，并在该知识图谱的基础上，通过问句分类、问句解析、查询结果三个步骤，调用大量函数构建问答框架。

基于规则匹配的问答系统没有复杂的算法，一般采用模板匹配的方式寻找最优答案，响应结果依赖于判断问句类型是否准确、知识库覆盖是否全面。用户抛出一个知识图谱中涵盖的问题，可以及时准确地给出全面的答案。但是整个问答系统的优劣过度依赖于知识图谱中知识的数量与质量，本系统还存在一定缺陷，如知识图谱覆盖面较窄数据提取和推理演绎较为困难，问句分类的算法不够效、在自然语言处理方面存在不足等，未来还需进一步完善。

参考文献

［1］ 许红梅，刘作义．乙型肝炎病毒母婴传播及其阻断［J］．实用儿科临床杂志，2005，20(9) : 835－837.

［2］ 李舟军，李水华．基于 Web 的问答系统综述［J］．计算机科 学，2017，44(6) : 1－7.

［3］ 冯 升．聊天机器人问答系统现状与发展［J］．机器人技术 与应用，2016(4) : 34－36.

［4］ 侯梦薇，卫 荣，陆 亮，等．知识图谱研究综述及其在医疗领域的应用［J］．计算机研究与发展，2018，55(12) : 2585－ 2599.

［5］ BOLLACKEK，EVANSC，PITOSHP，et al．Freebase: acollaboratively created graph database for structuring human knowledge［C］/ /Proceedings of the ACM SIGMOD international conference on management of data．Vancouver，Canada: ACM，2008: 1247－1250.

［6］ SINGHAL A． Introducing the knowledge graph: things，not strings［EB /OL］． 2012． https:

/ / googleblog． blogspot． com / 2012 /05/introducingknowledge－graph－things not．html．
［7］ NICKEL M，MU R PHY K，TESPV，et al．A review of relational machine learning for knowledge graphs［J］． Proceedings of the IEEE，2015，104(1) : 11－33．
［8］ 夏宇航，高大启，阮 彤，等．基于知识图谱的医疗病历数据存储研究[J].计算机工程，2019，45(1) : 9－16．
［9］ 陈功文．人工智能中的语义分析技术及其应用［J］．电子技术与软件工程，2019(21) : 239－240．
［10］ 褚晓敏，朱巧明，周国栋．自然语言处理中的篇章主次关系研究［J］．计算机学报，2017，40(4) : 842－860．
［11］ 李亚梦，张国鹏，刘 浏，等．智能外呼系统研究及设计［J］． 邮电设计技术，2018(12) : 77－82．
［12］ 谷兴龙，谢 珺，靳红伟，等．基于词特征与语义特征的评价对象识别［J］．计算机工程，2019，45(11) : 218－224．
［13］ 魏楚元．开放域问答系统问题理解关键技术研究［D］．北 京: 北京理工大学，2016．