

基于电子病历大数据的问答系统

张生泽 王庆阳 袁克虹

(清华大学深圳研究生院 深圳 518055)

〔摘要〕 基于电子病历大数据并结合信息检索与深度学习方法,设计一种辅助诊断问答系统。介绍系统的总体框架及电子病历数据库、价值网络、策略网络的设计,给出系统工作流程。该系统能帮助患者自查病情,也能为医生制定诊疗方案提供参考。

〔关键词〕 电子病历;信息检索;深度学习;辅助诊断

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2017.03.002

Question Answering System Based on the Big Data of Electronic Medical Records (EMR) ZHANG Sheng-ze, WANG Qing-yang, YUAN Ke-hong, Shenzhen Graduate School, Tsinghua University, Shenzhen 518055, China

〔Abstract〕 Based on the big data of Electronic Medical Records (EMR), information retrieval and the deep learning method, an auxiliary diagnosis Question Answering (QA) system is designed. The paper introduces the design of general framework, EMR database, value network and policy-based network of the system, provides the system operation process. This system can not only help patients to examine their conditions independently, but also provide reference for doctors when they make the diagnosis and treatment schemes.

〔Keywords〕 Electronic Medical Records (EMR); Information retrieval; Deep learning; Auxiliary diagnosis

1 引言

医院管理信息化和存储技术的高速发展,使电子病历成为病历管理的必然趋势。电子病历是医务人员在医疗过程中使用医疗信息系统生成的文字、图表、影像等数字化信息,能实现存储、管理、传输和重现的医疗记录^[1]。随着电子病历系统在医疗机构的迅速普及,大量医疗相关的重要信息以电子形式存储于医疗信息系统中,经过不断积累,各种形式的电子化系统产生了大量的医疗大数据。电子病历记录了患者就诊的所有信息,临床诊断的信息

检索、医疗数据的挖掘分析等相关研究都依赖于电子病历数据。

国外对电子病历的研究水平领先于国内并获得了政府的大力支持和资金投入。在一些国家如英国,电子病历在医院的应用已经超过 90%^[2]。在面向临床决策的电子病历系统研究方面,国外也取得了显著成果,例如 Rothman 等^[3]使用电子病历系统对患者的身体状况进行分级,辅助医生进行诊疗;Nucita 等^[4]在电子病历系统上构建 DERAM Software 系统,辅助医生完成艾滋病的预防和治疗。相对而言,国内对电子病历的研究起步晚,基础薄弱,缺乏足够的资金投入。目前,我国电子病历系统仅能实现部分区域的院内互联,缺乏共享性^[2]。但是,我国对电子病历的研究发展迅速,在临床决策与电子病历相结合方面,李国垒等^[5]通过对电子病历中的重要文本进行语义分析,提取临床辅助治疗的决

〔收稿日期〕 2017-03-28

〔作者简介〕 张生泽,硕士;通讯作者:袁克虹。

策方案,可实现电子病历的临床决策支持功能;李怡雪等^[6]从电子病历数据库、电子病历质量评价和文本检索 3 方面阐述了一个更全面的基于电子病历的临床辅助诊断系统。

电子病历记录患者的就诊过程,名医的电子病历更是代表着高效和准确的行医经验。建立电子病历数据库便是将医生的行医知识存储起来,合理利用数据库必然对辅助临床诊断、提高工作效率和医疗质量具有重要意义。由于电子病历中的数据包含了病历文本、化验结果和检查图像 3 类,所以只有结合这 3 类数据进行综合分析的系统才能得出更完整、更准确的医疗决策^[7]。基于以上分析,本文提出一种基于电子病历大数据的问答系统,采用 MATLAB 进行实现和实例分析,旨在综合、整理、共享电子病历,为经验不足的医生和患者提供诊疗参考。

2 系统介绍

2.1 概要

设计本系统的目的主要是构建电子病历数据库,利用病历文本、化验结果和医学图像的相关检索方法在数据库中查询相似度高、价值大的电子病历作为制定诊疗计划的参考,因此系统设计主要包括 3 大方面:电子病历数据库、价值网络和策略网络。针对电子病历数据库的设计,尽量减少存储空间,同时优化数据库结构以提高系统的检索速度;针对价值网络的设计,制定并评估电子病历的价值指标以帮助筛选出价值高的电子病历;针对策略网络的设计,以病历文本、化验结果和医学图像 3 类数据为指标并结合深度学习的方法计算电子病历之间的相似度。系统的总体框架,见图 1。

2.2 电子病历数据库

电子病历数据库根据电子病历中文本、化验结果和医学图像存储相应的特征。考虑到如果将电子病历中所有内容作为数据库中的项目会导致数据库异常庞大,增大电子病历检索时的困难,因此本研究将价值网络和策略网络计算得到的在价值、文本、化验结果和医学图像上的量化结果作为存储项

目,电子病历数据库构成,见图 2。

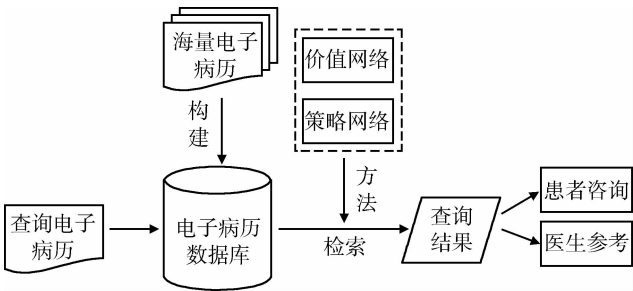


图 1 系统总体框架

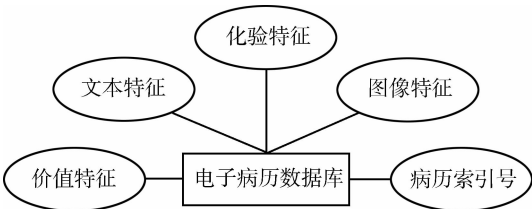


图 2 数据库构成

2.3 价值网络

参考李怡雪等^[6]对电子病历质量的评价方法,本研究确定电子病历的价值特征为质量指标、主治医师、相应科室和治疗效果的量化值,见图 3。

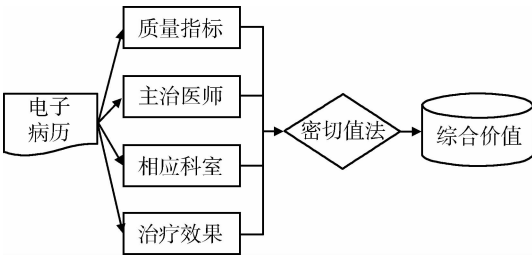


图 3 价值网络

电子病历的质量指标包括规范性、逻辑性和完整性 3 方面。规范性是指电子病历中各项的值是否符合书写的规定;逻辑性是指电子病历中各项目之间是否存在逻辑相悖之处;完整性是指病历中各项记录的完整程度。本研究采用 0~1 的数值来评价电子病历规范性、逻辑性和完整性,将得到的乘积作为电子病历的质量指标。主治医师和相应科室这两个指标是一定区域内的统计量。主治医师等级的计算方法为统计区域内所有医院相应等级的医生对

疾病的治愈率，相应科室等级的计算方法是统计医师所在医院科室对疾病的治愈率。此外，很明显治愈的病历比未治愈的病历更具有参考价值，因此在量化治疗效果时分别以 0 ~ 1 的数值代表无效、稍有好转、明显改善和治愈等情况。电子病历数据库中的价值特征包括以上 4 个指标。当需要对新电子病历进行检索时，根据这 4 种特征指标采用密切值法便可计算出数据库中每份与待检索电子病历相关的病历的综合价值，其检索过程在价值网络中进行。

2.4 策略网络

2.4.1 网络架构 策略网络是量化电子病历文本特征、化验结果和图像特征以及计算待检索病历与数据库病历相似度的重要结构。量化的 3 类特征存储于数据库中用于电子病历相似性的计算，对于待检索电子病历的相似度检索过程都在策略网络中进行。策略网络构架，见图 4。

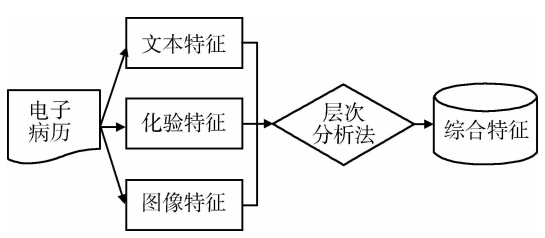


图 4 策略网络

2.4.2 文本特征 病历的文本内容具有结构化、术语话等特点，与自然语言结构有着较大区别。在病历中，有用的信息除了患者的基本信息（如性别、年龄、嗜烟嗜酒等）外，往往是症状和治疗方法，因此基于关键词的文本相似度计算方法适用于对电子病历文本内容的分析。病历的文本中存在着多词同义现象，本研究将这些词归纳为同义词库，在处理文本内容时对出现在同义词库中的关键词进行合并；此外对于某些病症，总是对应某个或某几个特异性症状，特异性症状与相应疾病之间虽然不是同义词，但其在病历中能确定的意义是相同的，本文将这些词归纳为特异词库，特异词库中的关键词在处理时也需要进行合并；最后，文本中一些没有意义的停用词，如“的”、“我院”、“于”等在获取关键词时被去掉。

系统处理文本特征的结构，见图 5。

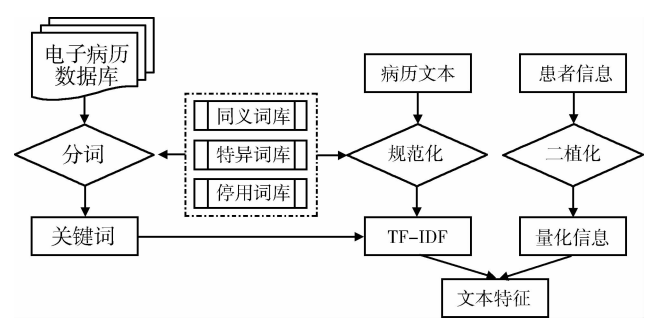


图 5 文本特征构架

存储于数据库中的文本特征来源于两种数据：一是描述患者病情的诊断报告；二是患者的基本信息。策略网络在计算文本特征时，首先根据电子病历数据库中所有的诊断报告，采用中文分词的方法得到关键词；对于待检索的电子病历，利用相同的同义词库、特异词库和停用词库规范化其中的诊断报告，根据关键词计算其 $TF-IDF^{[8]}$ 的值得到量化的诊断报告；对待检索电子病历中的患者信息，采用 0 和 1 二值化每一项（如“男”置 1，“女”置 0）得到量化的患者信息；最后将量化的关键词 $TF-IDF$ 值和量化的患者信息加权合并为文本特征向量，这个文本特征向量即为存储于数据库中的文本特征项。策略网络在计算电子病历间的文本相似度时采取的方法是计算电子病历对应的文本特征向量之间的余弦相似度^[9]。

2.4.3 化验特征 电子病历中的各种检查项目是医生诊断病情、确定治疗方案的重要依据，因此这些信息需要存储于数据库中作为电子病历的特征，本研究将其称为化验特征。由于不同的化验项目参照标准、量纲和数值意义等都不相同，将所有化验项目的原始值存入数据库中既增加存储器的负担，也不利于提高检索速度，所以很直观的解决办法便是将各项化验项目进行离散数值化。虽然临床上对于不同疾病的化验项目不尽相同，但是对于同类疾病化验项目都大同小异，而这些项目的正异常往往能确定某种或某些病变，鉴于此，本研究采用“偏低置为 -1、正常或阴性置为 0、偏高或阳性置为 1”的方法量化各个检查项目，将得出的化验项目向量作为存储于电子病历数据库中的化验特征。系

统检索时，采用结合名医建议的项目重要性的层次分析法来加权化验项目特征向量，用处理后各特征向量的余弦相似度代表化验特征之间的相似性。系统处理化验特征的结构，见图 6。

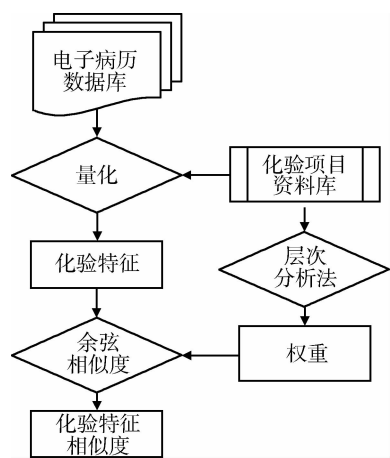


图 6 化验特征构架

2.4.4 图像特征 随着医疗技术的高速发展，医学影像已经成为临床诊断中不可或缺的一环。大多数疾病都会引起生理上的病变，从而导致生理结构改变，生理结构上的变化能通过医学成像技术客观地呈现出来，电子病历中的医学图像的相似性，特别是同种类型、同样位置的医学图像的相似性能够客观反映患者病情的相似性。但是医学图像也存在噪声干扰大、不同组织灰度相近、局部效应等缺点，会导致某些异常的图像与正常的图像在整体上可能比异常图像更具有相似性。此外，数据库中不可能存储图像本身的原始数据，需要将图像量化后的结果存储在数据库中作为对图像的数学表达。鉴于以上分析，本研究优先存储图像的正常（0）和异常（1），再在同类别中比较图像的特征。本研究所提出的问答系统在图像特征这一功能模块基于肺癌的肺部 CT 图像数据进行设计，通过前期对高斯混合模型、稀疏自编码和深度哈希编码试验，确定了深度哈希编码是提取图像特征最优的方法。深度学习中的卷积神经网络模仿人类大脑神经传输和识别的过程，理论上说只要人可以识别都可以用深度学习做到^[10]。本研究借鉴深度哈希编码思想^[11]对胸部 CT 中的胸腔壁、心血管、肺实质和肺癌区域

数据用卷积神经网络进行训练识别，用训练好的网络作为图像的哈希编码函数对电子病历中的医学图像进行量化，将量化好的图像哈希码存储于数据库中作为图像特征。图像哈希码既能表征图像的类别（正常、异常），其汉明距离^[12]又能表征同类别图像之间的相似度。系统处理图像特征的结构，见图 7。

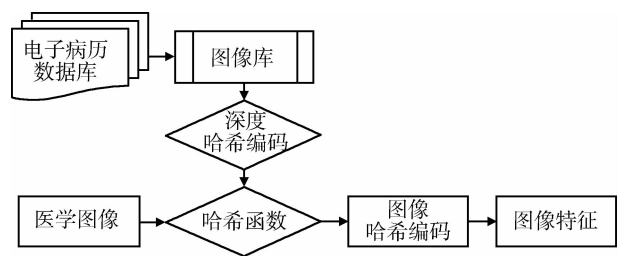


图 7 图像特征构架

2.4.5 综合相似度 策略网络最后一个结构是计算结合文本特征相似度、化验特征相似度和图像特征相似度的综合特征相似度，即系统最终确定的检索相似度。本研究对于综合相似度的计算采用加权求和的方法：根据名医对电子病历文本、化验结果、医学图像重要性的打分，采用层次分析法得到权重，3 种相似度的权重与归一化的 3 种相似度分别乘积之和作为系统检索时电子病历数据库中的病历与待检索病历的相似度。

2.5 系统工作流程

系统提供电子病历的输入和输出界面，其工作流程，见图 8。系统在文本界面、化验界面和图像界面获取电子病历的相关数据，依据策略网络计算文本特征、化验特征和图像特征，接着计算出数据库中的电子病历与待检索电子病历的相似度，然后取后前 N 份最相似的病历，利用价值网络对这 N 份电子病历的价值进行排序，最后将得到的排序结果索引项对应的原始电子病历进行输出。患者或经验不足的医生可以通过本系统检索到与待检索电子病历相似的病历，这些电子病历是名医经验和知识的载体，可为医生和患者提供参考。

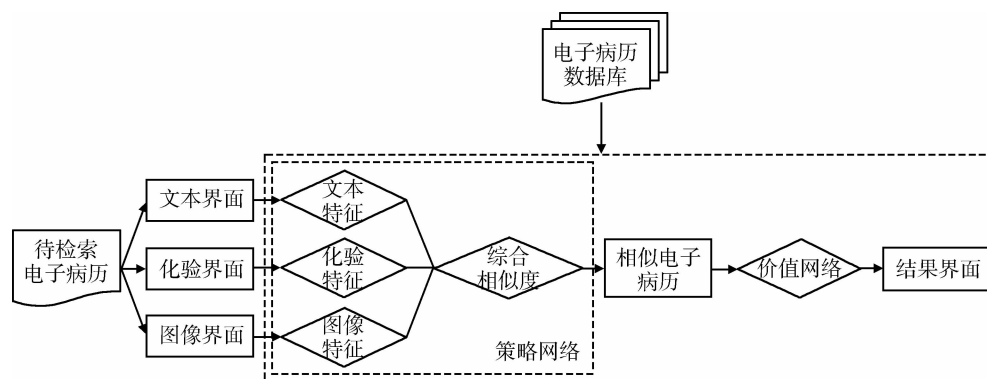


图 8 系统工作流程

3 结语

本系统价值网络和策略网络中的各个指标都可以独立计算,即使数据不全的病历(例如缺少医学图像)也可以进行检索,极大降低了系统对电子病历的限制,拓展数据的容量,提高系统的适应性。本研究以肺癌电子病历实践了问答系统的检索能力,在文本和图像检索方面都取得了与现有研究结果相当的准确率。医院信息化是社会发展的必然趋势。电子病历是医生知识和经验的载体,挖掘电子病历的价值可以共享医生的知识和经验、改进诊疗水平、缓解我国医疗资源短缺的现状,将电子病历与大数据分析相结合,是未来医疗大数据服务临床的重要研究方向。本系统建立在电子病历数据库的基础上,采用深度学习等人工智能技术挖掘电子病历数据内容上的价值和相似度,从而达到将医生宝贵经验和知识应用于辅助诊断的目的。由于系统呈现给使用者的是电子病历数据库中待检索病历相似的原始病历,这种客观的原始病历避免了人为处理造成的错误,无论是经验不足的医生还是患者都可以根据这些原始病历获取参考信息。

目前本系统尚处于试验阶段,其可靠性和稳定性需要在实际应用中进一步验证和改善。系统的文本特征分析模块,在大量的电子病历支持下也可借鉴深度学习在文本语言处理中的方法^[13-14]来进行改善。本系统所涉及的处理方法如大数据分析、深度学习等在越多的数据上越能表现其优越性,相信随着数据量越来越多,系统将会更加完善并具备更好的性能。

参考文献

- 1 沈剑峰,汪巍,杜平,等. 基于卫生信息平台的电子病历信息共享研究和实现 [J]. 中国生物医学工程学报, 2013, 32 (4): 504-507.
- 2 穆芳洁. 国内外电子病历的发展概况及思考 [J]. 中国病案, 2014, 15 (9): 40-42.
- 3 Rothman MJ, Rothman SI, Th BJ. Development and Validation of a Continuous Measure of Patient Condition Using the Electronic Medical Record [J]. Journal of Biomedical Informatics, 2013, 46 (5): 837-848.
- 4 Nucita A, Bernava GM, Bartolo M, et al. A Global Approach to the Management of EMR (Electronic Medical Records) of Patients with HIV/AIDS in Sub-Saharan Africa: the experience of DREAM software [J]. BMC Medical Informatics and Decision Making, 2009, 9 (1): 2723-2723.
- 5 李国垒,陈先来,夏冬,等. 面向临床决策的电子病历文本潜在语义分析 [J]. 现代图书情报技术, 2016, 32 (3): 50-57.
- 6 李怡雪,韩薇,王庆阳,等. 医院信息化: 临床决策支持系统构建 [J]. 医学信息学杂志, 2015, 36 (11): 6-10.
- 7 王庆阳,张书旭,白彦灵,等. 电子病历辅助临床诊断系统研究 [J]. 医学信息学杂志, 2016, 37 (7): 1-6.
- 8 Shin Y, Ryo C, Park J. Automatic Extraction of Persistent Topics from Social Text Streams [J]. World Wide Web, 2014, 17 (6): 1395-1420.
- 9 Xia P, Zhang L, Li F. Learning Similarity with Cosine Similarity Ensemble [J]. Information Sciences, 2015, 307 (C): 39-52.

(下转第 39 页)

申请然后点击“保存”按钮，但未提交的状态。该状态需要写日志。(3) 已提交：指医生提交会诊申请给护士，护士尚未校对。该状态需要写日志。(4) 已发送：指护士已校对完会诊医嘱，会诊申请已经“发送”给会诊科室。该状态需要写日志。(5) 已预约：指会诊科室护士已经安排了会诊医生。该状态需要写日志。(6) 会诊中。(7) 会诊结束：这两个状态目前用不到，保留。(8) 完成会诊意见：指会诊医生填写会诊意见并提交给申请科室。该状态需要写日志。如果是修改会诊意见，则内容描述字段需要写入“修改会诊意见”。(9) 申请取消中：指医生执行“取消会诊”或者医嘱“取消执行”操作。该状态需要写日志。(10) 申请已取消：指护士校对了新取消执行的会诊医嘱，会诊正式取消。该状态需要写日志^[4]。

4.3 权限控制

会诊申请功能受权限点控制，总计 8 个权限点^[6]。其中“1802 - 新开会诊申请”用于控制“新建”、“保存”、“删除”3 个功能，如果有 1802 权限，则这 3 个按钮将显示。“1808 - 编辑会诊意见”用于控制会诊医生是否可以填写和提交会诊意见。会诊申请功能除受以上权限点控制外，在流程中，还需要控制病历的阅读权限，即会诊申请发出后，

只有当会诊科室护士安排了会诊医生之后，会诊医生才可以浏览患者的病历，但当会诊状态为已取消时，则失去查看病历的权限^[5]。

5 结语

山西大医院以电子病历系统为基础进行再次开发，使会诊流程电子化，利用信息化手段对院内会诊进行管理，大大地提高院内会诊的质量和临床科室的工作效率，加强科室间的有效沟通以及学科间合作、交叉互动，持续推进医疗服务质量改进^[6]。

参考文献

- 1 叶俊. 完善院内会诊流程 提高临床医疗质量 [J]. 现代医院管理, 2013, 11 (6): 56 - 57.
- 2 田柯, 耿仁文, 林凯程. 院内会诊存在的问题及对策分析 [J]. 现代医院, 2011, 11 (2): 110 - 111.
- 3 肖强, 吴伟斌, 陈联忠, 等. 电子病历系统的主要实现技术 [J]. 中华医院管理杂志, 2004, 20 (4): 207 - 209.
- 4 王秋卉, 王东, 姜素霞. 院内科间会诊质量管理的改进措施 [J]. 解放军医院杂志, 2014, 21 (1): 38 - 39.
- 5 夏克俭, 刘芸虬. 医疗系统中基于 RBAC 的权限管理的实现 [J]. 办公自动化, 2008, (12): 42 - 44.
- 6 闻大翔, 陆琳, 甘宁, 等. 基于六西格玛方法对优化院内科间会诊的研究 [J]. 中国医院, 2009, 13 (9): 8 - 11.

(上接第 11 页)

- 10 Najafabadi MM, Villanustre F, Khoshgoftaar TM, et al. Deep Learning Applications and Challenges in Big Data Analytics [J]. Journal of Big Data, 2015, 2 (1): 1 - 21.
- 11 Xia R, Pan Y, Lai H, et al. Supervised Hashing for Image Retrieval via Image Representation Learning [C]. AAAI Conference on Artificial Intelligence, 2014: 2156 - 2162.
- 12 Hamming RW. Error Detecting and Error Correcting Codes [J]. Bell Labs Technical Journal, 1950, 29 (2): 147 - 160.
- 13 Mikolov T, Karafiát M, Burget L, et al. Recurrent Neural Network Based Language Model [C]. Makuhari, Chiba, Japan: Conference of the International Speech Communication Association, 2010: 1045 - 1048.
- 14 Sukhbaatar S, Szlam A, Weston J, et al. End - To - End Memory Networks [C]. Proceedings of the 28th International Conference on Neural Information Processing Systems, 2015: 2440 - 2448.