· 医院管理 ·

基于医疗大数据平台的相似病历检索系统

杨 辉[1],薛 淞[1],顾广励[1],黄 锋[1],金赛娟[1],纪永章[2]

[摘要] **目的** 基于自然语言处理技术,实现医疗大数据平台上病历库中的相似病历检索。 **方法** 对病历中的结构化部分,采用平台的索引检索技术;对非结构化的自然语言描述部分,基于构建的医学特征库,做特征提取、相似度计算,从而检索出相似病历。 **结果** 该系统可以检索出病历库中的相似病历,用户可以基于检索结果做辅助诊断或科研分析。 **结论** 通过对检索结果的判断,证明基于自然语言处理技术的相似病历检索系统是可行的,但是在提高精度方面还需后续的改善工作。

[关键词] 索引检索;自然语言处理;相似病历

[中图分类号] R197 [文献标志码] A [文章编号] 1672-271X(2017)02-0210-03

[DOI] 10.3969/j.issn.1672-271X.2017.02.027

随着医院信息化的逐渐深入,电子病历系统 EMR(Electronic Medical Record)已被各大医院广泛使用。经过多年积累,EMR 系统已收集到海量的信息并逐渐迈入大数据时代,这些电子病历中大量的文本信息成为了各个医院的宝贵财富。然而,HIS 系统中原有相对简单的统计功能已不能满足人们日益增长的需求[1]。如何利用 EMR 系统的海量文本信息为医师及病患服务成为一个研究课题。本文利用自动分词、建立医学词汇本体库等自然语言处理技术及基于开源搜索引擎 solr 的索引检索技术,提出一种基于语义相似度计算的方法,从而实现相似病历检索功能,为电子病历文本信息的利用与电子病历的质量监控提供了参考[2]。

电子病历在各级医院中逐渐普及。除病程记录,越来越多的临床系统数据如检验、检查等数据被集成到电子病历中,因此,电子病历的数据如何存储、检索、二次利用等日渐成为研究热点[3]。

国内外均研究临床数据格式的标准,如 HL7 CDA 可作为电子病历的设计规范。国内大的电子病历厂商除遵守总体临床数据标准外,也将各模块努力做到标准化[4]。如在现病史输入环节,有些 EMR 提供症状词典,并为某些疾病设置几种模板。这些研究工作均在使电子病例的数据录入、存储尽量格式化、标准化。然而,至今还无国家或业内统一的症状词典及常用术语词典,并且大部分疾病的描述无法按照固定模板输入。

在电子病历数据检索、二次利用方面也随之存在一些困难。如医师在遇到疑难杂症难以判断或做医学研究时,希望能自定义一些输入条件进行检索历史的相似案例做参考,现有的系统很少能够满足上述医生的检索分析病历的需求。

随着响应国家建设区域医疗平台的号召,很多医院都在建设院内的医疗数据平台[5]。解放军第四五四医院已经探索搭建了一个基于医疗数据存储的大数据平台,该平台上集成了来自 HIS、LIS、EMR、PACS 等系统的各种格式的数据,并实现了基础的快速检索功能。建立大数据平台的一个重要意义,是在收集海量数据之后对数据进行分析,挖掘出在单个系统上无法发现的关联信息[6]。为了二次利用医疗数据的价值,本文设计了一个基于上述大数据平台上的相似病历检索系统,通过文本检索出相似病历以后,可以进一步查看相关的检验数据和影像数据等信息。

1 系统设计与实现

1.1 设计思想 利用大数据平台的数据收集功能,从 HIS、LIS、EMR、PACS 等系统的 DB、HTML、PDF、HL7、DICOM 等形式的标准数据或文件中抽取用户自定义的 meta data(元数据),并将该元数

作者单位:210001 南京,解放军第 454 医院,1. 信息科,2. 医务处

通信作者:纪永章,E-mail:Jyz454@sohu.com

引用格式:杨 辉,薛 淞,顾广励,等.基于医疗大数据平台的相似病历检索系统[J].东南国防医药,2017,19(2):210-212.

据和对应的源数据文件以对象的形式保存在内容存储平台上,本平台采用了日立存储[7]。该平台还利用 Solr 对这些元数据及源文件建立了全文索引,可以快速检索并显示相关文件[8]。基于此平台上的相似病历检索系统功能设计为对病历中的结构化部分,采用平台的索引检索技术;对非结构化的自然语言描述部分,基于构建的医学特征库,做特征提取、相似度计算,检索出相似病历。系统构架图,见图 1。

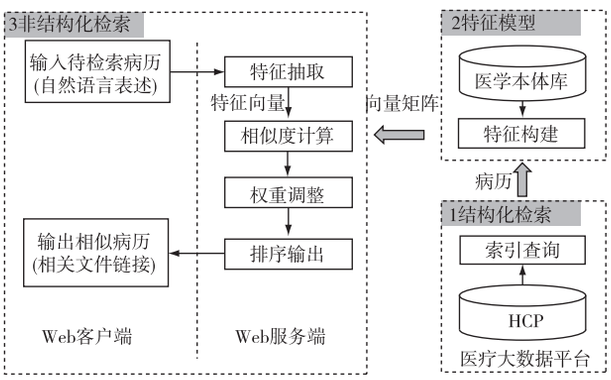

图 1　相似病历检索系统架构图

1.2　*电子病历结构分析*　该平台上收集的电子病历源文件为本院 EMR 系统导出的 HTML 文件。分析 xml 文件中包含的关键信息,见图 2,其中结构化的部分,如患者性别、年龄、体格检查结果等,直接利用 Solr 工具建立索引[9],并在检索界面上提供相应的检索输入接口;非结构化的数据,如现病史的描述部分,Solr 建立了全文索引,在检索界面上可输入其中包含的语句、关键词进行查询,但检索性能一般。

<fieldelem name="病人姓名">张三</fieldelem>
<fieldelem name= "病人性别">女</fieldelem>
<fieldelemname= "年龄">63 岁</fieldelem>
<fieldelemname= "现病史内容">患者昨晚受凉后出现畏寒、发热,体温最高达 39.0℃,伴有头疼,轻度咽痛、恶心,无呕吐,无鼻塞、流涕,大便顺解,夜尿困难,伴有尿痛,患者今晨来我院就诊,门诊拟查胸部正侧位:未见明确实质性病变,予以口服头孢克肟片和白加黑感冒片口服,患者仍有高热,今日下午门诊就诊,拟"发热待查"收入我科。病程中,患者精神欠佳,睡眠、饮食一般,大便畅解,夜尿稍多,伴有尿痛。近期体重无明显变化。</fieldelem>
<fieldelemname="体温值">42.0</fieldelem>
<fieldelemname="脉搏值">96</fieldelem>

图 2　电子病历片段

1.3　*基于相似度电子病历检索*　基于图 1 中的非结构化数据中类似现病史的描述部分,虽然 Solr 建立了全文索引,可以通过检索界面输入一些词句,但是用户需要自己组织关键语句,并且 Solr 未对各分词做特殊处理,无法区分症状词语较其他词语的重要性,因此检索结果不易控制。而对相似病历中的非结构化数据检索则作了基于语义的相似度计算。

1.3.1　*基于结构化数据的检索*　先分析待检索病历的一些有意义的特征,设置检索条件。如设置的检索条件(性别:女,年龄:60~70,体温:39~42,科室:呼吸内科等),可粗略筛选出一组病历。

1.3.2　*对非结构化数据构建特征模型*　为了构建特征模型,首先准备医疗领域的本体库,其中描述了电子病历的各种特征。症状特征可用常用症状词典表示,如畏寒、发热、头疼、咽痛、恶心、呕吐、鼻塞、流涕、尿痛等。对筛选出的每个病历中的非结构化部分,如现病史描述,通过症状词典可以构建出一个特征向量[10]。具体设计如下:出现症状词语用 1 表示;未出现用 0 表示;出现但是用"无"修饰时,用-1 表示。按规则处理"无"、"否认"、"不伴有"等属于相似词语。由于症状词典词语很多,初步构建出的向量维度较大,从运算速度和语义意义上需要做降维处理。本系统采用奇异值分解,将每个向量降到十几、几十维。至此,对所有筛选出的病历,构建了一个特征向量的矩阵模型。

1.3.3　*原始病历与矩阵模型做相似度计算*　将原始病历中的非结构化部分,如现病史描述执行与"1.3.2"中相同的处理流程,得到一个特征向量。通过比较该特征向量与上述特征矩阵中的每个向量的距离,得出该病历与上述病历组中的每个病历的初始相似度。如本体库中还提供各症状的权重,即反映疾病的重要程度或频率,可利用该知识对初始相似度做进一步修正,得出最终的相似度。症状权重也可利用基于词频的统计进行试验,然后经过专家确认得出。计算出相似度以后,在输出界面上按照相似度大小顺序显示。除直接显示出相似的文本信息外,还提供原始病历的链接以及相关影像等文件的链接。用户可根据自身需求,进行更深入的查看分析。

2 检索结果与讨论

本系统采用某科室的一批电子病历做初步试验。由于相似度的计算结果判定无业界标准,且无业界统一试验数据库,因此,只能人为地判断计算结果的优劣。当输入的病历也存在病历库中时,两者相似度是 100%;相似度在 80%~100%之间的病历,通常是有参考意义的;病历库越大,检索出相似度高的病历的概率越大。检索结果也反映出了很多待处理的问题:一是由于症状的描述不规范,需要收集症状的近似词典,如“乏力”、“无力”等。二是由于症状词典不够丰富,某些科室或疾病的常用语没有被作为重要特征,待常用语词典被添加后,相似度结果会更精确。三是症状修饰的部位,如“双下肢”、“左下肢”还未建立关系。在本体库中增加这样的关系后,检索结果也会更精确。病历描述语言的处理涉及复杂的自然语言处理技术,如果考虑更多的特征点,需要长期的对系统进行优化与提高。本系统基于自然语言处理和本体的相关技术,对相似病历检索做了一个初探。

3 结 语

本文阐述了一个基于医疗大数据平台的相似病历检索系统,对平台上存储文件中的非结构化数据、即自然语言描述部分,做了特征抽取和相似度计算,并将检索结果显示给用户。

检索出相似病历以后,用户可进一步查看相关的检验、影像数据信息。利用该系统,用户可以参考相似病历做辅助诊断,也可根据自己的科研需求分析某一类特殊病历并从中挖掘新的知识。

【参考文献】

[1] 宋 斌,陈海东,雷 勇,等. 数据仓库在数字化医院的应用[J].东南国防医药,2010,12(6):519-522.

[2] 赵伯诚,周 斌,吕耀欣,等. 我院监控电子病历质量的实效与经验[J].东南国防医药,2010,12(3):276-277.

[3] 张志常,娄 岩. 2013-2015 基于电子病历的 SCI 论文主题词聚类分析[J].中国数字医学,2016,11(3):26-27.

[4] 孟 岩,李姗姗,宋海庆,等. 电子病历深度应用及体会[J].中国数字医学,2016,11(7): 111-113.

[5] 安志萍,高志军,张云宏,等. 远程病案信息查询系统的构建与应用[J].医学研究生学报,2016,29(12):1325-1327.

[6] 邹北骥. 大数据分析及其在医疗领域中的应用[J].计算机教育,2014,7:24-29.

[7] 薛以锋,顾广隶,赵伯诚,等. 基于元数据文件存储的医疗大数据平台研究与实现[J].中国数字医学,2015,10(10):73-75.

[8] 周 斌,杨 辉,薛 淞,等. Solr 在医疗大数据检索中的应用[J].中国数字医学,2016,11(9):21-23.

[9] 霍 庆,刘培植. 使用 Solr 为大数据库搭建检索引擎[J].软件,2011,32(6):11-14.

[10] 王 欢. 基于领域本体和 Lucene 的语义检索系统研究[J].计算机应用,2010,30(6):1656-1660.

(收稿日期:2016-07-20; 修回日期:2016-12-29)

(本文编辑:刘玉巧)