

临床科研数据抽取研究

牛承志 骆 鑫 赵 丹

(郑州大学第一附属医院 郑州 450052)

〔摘要〕 以医院信息系统为研究基础,以临床医学文本分类为依据,从数据来源角度分析临床科研数据抽取方法和过程,使用自然语言处理、数据库检索技术对医学文本进行抽取和挖掘,取得良好效果,助力临床科研工作。

〔关键词〕 医学文本;信息抽取;自然语言处理

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2020.07.009

Study on Clinical Research Data Extracting NIU Chengzhi, LUO Xin, ZHAO Dan, The First Affiliated Hospital of Zhengzhou University, Zhengzhou 450052, China

〔Abstract〕 Taking the Hospital Information System (HIS) as study basis, the paper analyzes the extraction methods and process of clinical research data from the perspective of data sources according to the classification of clinical medical texts. Natural Language Processing (NLP) and database retrieval techniques are used to extract and mine medical texts, which achieves good results and helps clinical research.

〔Keywords〕 medical text; information extraction; Natural Language Processing (NLP)

1 引言

随着信息化迅速发展,医院业务流程已实现闭环管理以及全流程数据记录。医院信息系统中存储大量临床数据,从类型来区分,可以分为临床医学影像数据和临床医学文本数据两类。临床医学文本数据主要包括患者检验结果、诊断报告、病历文书、病理报告等。临床医学文本中嵌入大量有用信息,如何利用现有技术方法从医学文本数据中提取潜在、有价值的信息逐渐成为信息抽取领域关注和研究的重点^[1]。本文以国内某大型三甲医院信息系

统为基础,以临床医学文本分类为依据,从数据来源角度分析临床科研数据抽取方法和过程,使用自然语言处理、数据库检索手段对医学文本进行抽取和挖掘,取得较好效果。文中介绍的数据分析、抽取、集成方法均来自日常临床科研数据工作实践,希望能够为临床科研数据挖掘工作提供一些思路。

2 临床科研数据

2.1 概念

指医院信息系统运行过程中产生的各种形式数据,包括患者医嘱、费用、医疗文书和检查检验等信息,主要来自医院信息系统(Hospital Information System, HIS),电子病历系统(Electronic Medical Records System, EMRS),医学影像存储与传输系统(Pictures Archiving and Communication System,

〔修回日期〕 2020-03-12

〔作者简介〕 牛承志,硕士,工程师,发表论文10篇,参编专著1部;通讯作者:赵丹,副教授。

PACS), 检验信息系统 (Laboratory Information System, LIS) 4 大系统。医疗数据是医疗活动的数据记录, 对诊断和治疗疾病具有重要意义。临床科研数据针对课题研究使用, 研究课题不同, 所要抽取数据也不同。例如药物研究课题是针对使用某种药物的患者来展开研究, 糖尿病、肾病是以疾病为中心展开研究。临床科研数据抽取是精准、定量定性的数据抽取过程。例如狼疮肾炎课题需要抽取患者第 1 次血常规中的红细胞计数数据, 而不是所有实验室数据; 研究 2 型糖尿病课题需要从入院记录中抽取患者糖尿病病史, 而不是整个现病史记录。

2.2 分类

临床医学文本从结构化角度分析, 可分为两类, 即结构化和非结构化数据。结构化数据是按照预先设计好的数据类型和长度规则, 通过不同软件进行数据存储和管理; 是典型的关系型数据, 以行和列组成的二维表, 可以通过结构化查询语言 (Structured Query Language, SQL) 语句实现检索和使用, 主要有患者病历首页、实验室、医嘱、费用信息等^[2]。非结构化数据是未使用数据库软件进行存储管理的数据, 暂无法用二维表结构来描述, 以 HTML、XML、PDF 等形式存储在硬盘等介质, 以大段医疗文书描述为主, 优点是便于书写, 可以反映书写者连贯思维, 缺点是不易存储, 信息抽取困难。主要有患者医疗文书、影像报告描述等。

3 结构化数据抽取

3.1 内涵

从医院信息系统数据库中获取数据, 要充分考虑到信息系统安全和业务系统负载, 可以采用相对独立的数据仓库模式, 在医院信息系统中负载较轻的时间从相应业务系统中通过提取 - 转换 - 加载 (Extract, Transform and Load, ETL) 采集工具将来源于 HIS、EMRS、集成平台等系统的患者数据、门诊诊疗、门诊用药、住院治疗、住院医嘱、手术、检查检验等大量数据信息, 经过转换清洗, 存储到数据库中^[3], 见图 1。

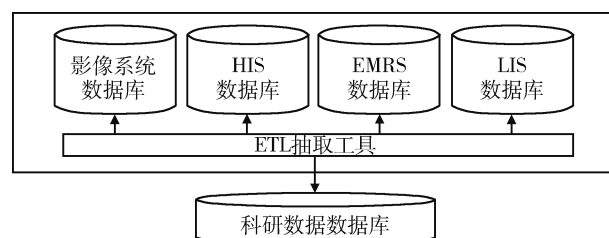


图 1 结构化数据抽取设计

3.2 步骤

结构化数据抽取需要耗费大量服务器资源, 在抽取过程中将业务系统和抽取业务相互独立, 保证医院日常业务正常运行。结构化数据抽取步骤如下: 从诊断表中按时间找到符合条件的患者住院号; 从患者变更表中筛选出在 ICU 治疗过的患者住院号, 存入本地数据库。住院号是患者唯一标识, 以此将患者所有数据串联起来; 依据住院号抽取患者首页信息, 存入本地库; 依据住院号将患者所有 LIS 数据抽取出来, 存入本地库; 依据住院号将患者医嘱抽取出来, 存入本地库; 按需求创建数据集表, 更新患者首页部分内容; 与需求者沟通确定医嘱中的打分规则, 输入每名患者分值, 更新数据集表医嘱部分内容; 按需求抽取患者 LIS 数据, 整合数据, 分别取最高值、最低值更新数据集表。

4 非结构化临床医学文本抽取

4.1 内涵

从电子病历抽取科研课题所需信息, 本文以常见多发疾病 2 型糖尿病为研究对象, 采用基于规则的自然语言处理、文本挖掘技术与算法从电子病历中入院记录自动抽取患者病史, 将其结构化处理。

4.2 数据描述

电子病历中患者疾病史会出现在入院记录中, 大多数格式为“疾病名称 + 时间陈述”, 例如“自述糖尿病 15 年”、“既往史, 患糖尿病 8 年余”、“20 年前确诊 2 型糖尿病”, 常见的糖尿病史描述模式, 见表 1。

表 1 糖尿病史描述模式示例

模式	实例
疾病 * 时间数值 * 时间单位	现病史：糖尿病史 2 年，口服二甲双胍……
时间数值 * 时间单位 * 疾病	…3 年至第二人民医院，以“①高血压 ②右眼玻璃体出血 ③糖尿病视网膜脱落④甲状腺结节入院”…
时间数值 * 时间单位 * 症状	20 年前出现口渴、多饮、多食、多尿、饮水每天 3 - 4L，尿量与饮水量相当，伴消瘦，体重下降约 5Kg，
* 诊断 * 疾病	不伴手足麻木、视物模糊，就诊于郑州市第四人民医院，查空腹血糖（具体值不详），确诊为“2 型糖尿病”

4.3 方法与步骤

分析病史描述规则后进行糖尿病史抽取与组织，见图 2。糖尿病史抽取与组织可分为 5 个步骤，首先构建糖尿病史语料库，组织在校医学硕士分为 3 组，其中两组文本标注，1 组核对结构，对内分泌科入院记录进行规范化标注，总结分析出入院记录中的描述糖尿病史模板；其次将第 1 步结果经内

分泌临床医生讨论后构建规则库^[4]；第 3 步在非结构化化的入院记录中，基于构建的规则库对糖尿病史进行自动抽取，将抽取的病史结果进行存储；第 4 步核对入院记录中有关否定描述内容，从抽取的结果中剔除有关否定的结果^[5]；第 5 步采用人工和算法核对相结合的方法对结果进行验证。糖尿病史抽取结果，见表 2。

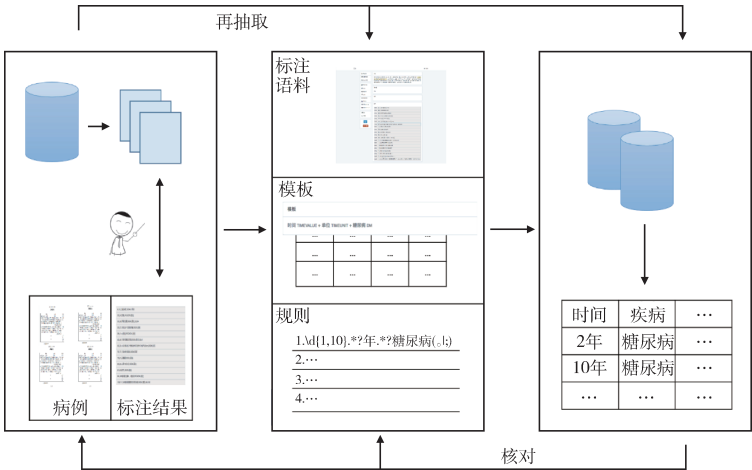


图 2 非结构化电子病历文本数据中的糖尿病史抽取与组织

表 2 糖尿病史抽取结果

患者住院号	抽取结果（章节：分类→结果）
ZY01000XXX3963	既往史：2 型糖尿病：——→2 型糖尿病 7 年
ZY0100XXX5671	个人史：2 型糖尿病：——→2 型糖尿病病史 6 年
ZY02000XXX2497	现病史：2 型糖尿病：——→2 型糖尿病 9 余年
ZY02000XXX2216	既往史：2 型糖尿病：——→2 型糖尿病 7 年
ZY04000XXX9027	既往史：2 型糖尿病：——→2 型糖尿病 15 天
ZY01000XXX5884	既往史：2 型糖尿病：——→2 型糖尿病 6 年

4.4 结果分析与验证

为说明抽取结果有效性，采用两种验证方法并行策略，第 1 种方法选取 1 个科室 2015 年全部数据进行基于辅助工具的人工验证，其验证结果为整体抽取结果的指标分析提供参考依据；第 2 种方法对所有科室数据进行 10% 取样，取样使用人工标注语料的 5%，其目的在于经过专业人员标注的内容可以作为糖尿病史抽取结果来验证^[6]，从而提高全部科室糖尿病史抽取准确率。抽取结果验证指标，见表 3。

表 3 抽取结果验证指标

项目	结果	指标	结果
总验证病历数	N = 1 436	PRECISION	99.5%
TP (True Positive)	765	RECALL	87.6%
FP (False Positive)	4	F - MEASURE	0.93
FN (False Negative)	108	-	-
TN (True Negative)	558	-	-

5 结语

通过对临床医学文本结构化和非结构化数据抽取，将获取到的数据组织成数据集表，供临床科研工作使用。通过基于规则的抽取方法和自然语言处理技术相结合，实现非结构化电子病历、文本数据的定制化抽取；采用基于数据库的 ETL 技术实现从结构化临床医学文本抽取信息，取得良好效果，助力临床科研工作。

参考文献

1 汪新建, 马睿, 杨孝光, 等. 数字化医院科研信息化管理平台的设计 [J]. 中国数字医学, 2017, 12 (11): 74 - 76.

2 程显毅, 朱倩, 王进. 中文信息抽取原理及应用 [M]. 北京: 科学出版社, 2010: 103 - 105.

3 冯志伟. 自然语言处理简明教程 [M]. 上海: 上海外语教育出版社, 2012: 56 - 57.

4 俞鸿魁, 张华平, 刘群, 等. 基于层叠隐马尔可夫模型的中文命名实体识别 [J]. 通信学报, 2006, 27 (2): 87 - 94.

5 赵爽. 基于大数据的医疗系统虚拟化研究与设计 [J]. 医学信息学杂志, 2018, 39 (10): 26 - 28.

6 Név  l A, Zweigenbaum P. Clinical Natural Language Processing in 2015: leveraging the variety of texts of clinical interest [J]. Yearb Med Inform, 2016 (1): 234 - 239.

2020 年《医学信息学杂志》编辑出版重点选题计划

2020 年本刊将继续以“学术性、前瞻性、实践性”为特色，及时追踪并深入报道国内外医学信息学领域前沿热点，反映学科研究动态，展示学科研究与应用成果，引领学科发展方向。现对 2020 年度编辑出版重点选题策划如下：

一、医学信息学研究

1 医学信息学基础理论研究；2 医学信息学研究新方法、新指标、新工具；3 医学人工智能的前沿技术、科研发展及临床应用；4 医学智库研究与智库服务实践；5 医学数据中心建设及数据挖掘、关联应用；6 生物信息学、生物医学信息学、基因组学理论研究及实践；7 公民健康素养培养及健康促进。

二、医药卫生体制改革与医药卫生信息化

1 “互联网+医疗健康”支撑体系、服务体系建设；2 健康中国战略背景下医药卫生信息化发展规划与方案；3 信息化助力新型医疗健康服务体系建设目标及解决方案；4 医疗卫生信息标准化与规范化建设现状和应用；5 医疗卫生信息化相关法律法规。

三、医学信息技术

1 医用机器人的系统研发、模型设计及典型案例；2 健康医疗大数据的科学管理及智能分析；3 精准医学与个性化医疗技术研究与应用；4 物联网、移动互联网在诊疗及健康管理中的综合运用；5 健康云平台功能、技术、系统架构及基础设施构建；6 5G 医疗技术应用及实践；7 临床医疗可视化及虚拟现实技术的医学应用；8 网络安全体系建设与风险评估。

四、医学信息组织与利用

1 “互联网+”环境下医学图书馆的创新举措；2 人工智能技术及新型媒体在医学图书馆中的应用；3 知识发现技术进展及医学图书馆知识服务模式演变；4 区域医疗卫生信息资源整合及共建共享；5 基于医学数据图书馆、数据馆员的新型专业实践。

五、医学信息教育

1 “互联网+”环境下医学信息专科、本科、研究生教育及继续教育面临的挑战、改革与实践创新；2 医学信息学专业课程设置及复合型人才培养；3 网络化、数字化医疗信息教育培训平台及在线课程；4 基于互联网的健康科普知识精准教育；5 中外医学信息学教育比较研究。

(《医学信息学杂志》编辑部)