

现代名老中医医案数据库的构建与数据处理

赵泽鹏¹, 戴国华²✉, 高武霖²

1. 山东中医药大学, 山东省济南市历下区经十路 16369 号, 250014; 2. 山东中医药大学附属医院

摘要 运用数据挖掘技术构建医案数据库是归纳和总结名医经验的常用方法。为了提升数据挖掘结果的使用价值, 探讨现代名老中医医案数据库的构建方式, 在优化数据库结构与常用字段的基础上, 借助 HanLP 分词系统, 应用结合最大概率算法的隐马尔可夫模型 (HMM) 分词算法进行医案文本的切分与语义标注, 并通过 python 根据语义标注结果进行医案信息提取。该方法有助于优化医案数据结构, 提升了数据库的构建效率和数据挖掘质量。

关键词 中医医案; 数据库; 名老中医; 数据处理

中医医案具有较高的学术价值, 是后世中医学发展与创新的源头。随着信息技术的发展, 传统纸质医案已不能满足人们快速获取与利用的需求, 借助数据库与数据挖掘技术对大量中医医案进行储存、查询、统计分析成为中医信息化发展的新方向。

目前, 国内已建成的综合性医案数据库多由各大中医院校开发^[1], 如中国中医科学院中医药信息研究所的中国中医药数据库^[2], 万方数据开发的万方中医药数据库^[3], 上海中医药大学伤寒论教研室开发的历代医案分析统计系统^[4], 山东中医药大学研发的中医历代医案数据库^[5]等。这些综合性医案数据库大致可分为两类: 一类是基于本体整合医案信息的中医知识库, 另一类是对医案信息进行统计分析的关系型数据库。其中, 中医知识库侧重医案信息的存储与查询, 绝大部分按照自然语言录入, 但由于医案语言存在语义模糊、症状复合等现象, 这种未经处理的信息广度有余而深度不足, 统计分析时不利于数据特征的提取, 亦难免形成信息“孤岛”, 造成信息资源的浪费。关系型数据库虽侧重信息的统计分析, 但往往对计算机技术有着较高的要求, 症状信息的储存形式多为二进制, 仅从“是”或“否”去表述一个症状, 往往会忽略症状的派生性质, 不能很好地满足中医症状多样化特点。此外, 关系数据库还存在着数据结构

不一、术语标准化不足、同义词重复等问题, 降低了挖掘结果的实用性。除上述综合性医案数据库, 其他大多是围绕某种疾病或某位医家建立的专题数据库^[6-7]。这些数据库专指性较强, 数据来源较集中, 为体现专题特色, 不同数据库之间的结构设计及数据处理方法存在较大差异, 彼此缺乏信息共通, 难以进行系统化的应用。总体而言, 面对庞杂的医案信息, 大型综合性医案数据库储备充足但不便利用, 小型专题数据库处理灵活却难以整合, 与此同时, 名医临床决策系统等一系列人工智能技术又对数据质量产生了更高的要求。鉴于此, 探索一种能够系统化、广泛化地运用于构建医案数据库的数据处理方法尤为必要。

本研究基于现代名老中医医案, 对医案数据库的构建与数据处理从方法学的角度进行了讨论。旨在通过科学的数据处理方法, 提升数据挖掘质量, 归纳名老中医辨证论治经验, 为建立中医辨证论治大数据辅助诊疗云平台做好数据准备工作, 同时也为医案数据库的构建与数据处理提供思路。

1 医案数据库的建立

1.1 数据采集

1.1.1 数据来源 现代名老中医医案主要来源于国家中医药管理局评选的国医大师及全国第一、二、三、四批 500 位名老中医专家 1949 ~ 2020 年出版的临证验案集。

1.1.2 医案纳入标准 1) 语言叙述符合现代化用语习惯且内容完整; 2) 首次就诊; 3) 中药治

基金项目: 国家重点研发计划(2019YFC1710401); 国家自然科学基金(81774047)

✉ 通讯作者: daigh2004@163.com, (0531) 68616065

疗有效。

1.1.3 医案排除标准 1) 语言晦涩艰深、难以理解; 2) 四诊资料记录不完整、缺失具体方药、诊断不明确; 3) 内容重复。

1.2 研究方法

运用正则表达式对数字化医案进行文本清洗; 再通过 HanLP 分词系统应用结合最大概率分词算法的隐马尔可夫模型 (Hidden Markov Model, HMM) 对清洗后的医案信息进行中文分词与词性标注, 利用基于 HMM 的命名实体识别功能结合人工进行语义标注; 最后, 建立医案信息提取模型, 依据语义标注结果进行特定词性词语提取与聚类, 形成结构化数据并储存, 完成医案数据库的内容填充。

1.3 数据库设计

1.3.1 数据库结构设计 该数据库系统逻辑框架结构分为三层, 依次为用户界面层、计算处理层和数据存储层。用户界面层是用户与管理员访问数据库的窗口, 包含医案采集、医案目录、医案检索、数据挖掘、用户登录等链接, 用户可通过系统程序下达操作指令, 实现数据库的各种功能, 而管理员具有特殊权限, 负责医案数据与系统程序的维护。计算处理层由搜索引擎模块、数据处理模块与用户管理模块三个高内聚低耦合的模块组成, 以确保计算处理的高效性与稳定性, 可根据用户下达的操作指令, 对数据存储层中的数据依照既定原则进行处理, 并将结果返还至用户界面层。数据存储层选用 SQL Server 关系型数据库存储医案信息, 具有容量大、可用性强、安全快捷、易于管理与维护等特点^[8]。为兼顾数据库系统的完备性与执行效率, 根据存储信息类型的不同, 分为医案原文数据库、医案信息数据库、中医方剂数据库、中药信息数据库、规范用语数据库, 并从文本信息中抽取关键词构成关键词库, 建立索引机制以加强数据间的联系。将中药信息数据库中药物的相关信息, 如性味、归经、功效等, 与医案信息数据库中药物的正名相关联, 形成数据库间信息的映射^[9], 即可实现药物相关信息的自动分析。如图 1 所示。

1.3.2 常用字段设计 医案信息库常用字段及分类体系如图 2 所示。

2 医案数据库的数据处理

2.1 初步数字化

录入员通过扫描识别、格式转化或人工录入的方式, 使医案信息转化为可被医案录入程序识别处

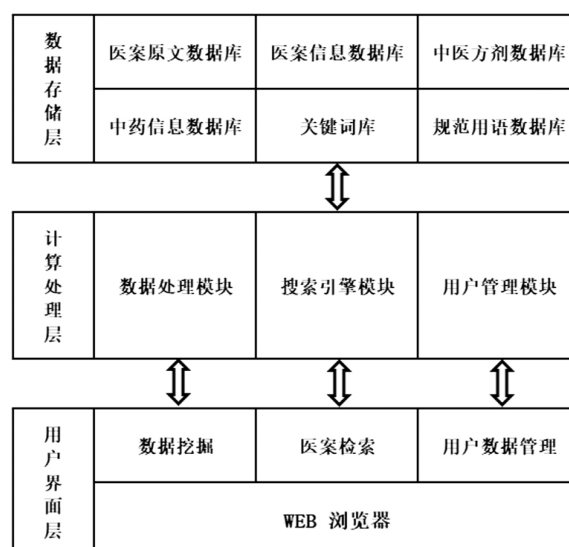


图 1 医案数据库结构设计图

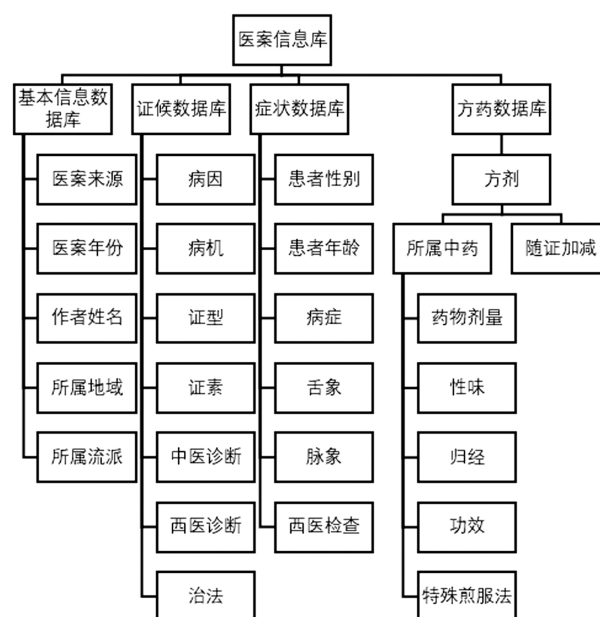


图 2 医案信息库常用字段及分类体系

理的文本格式, 并上传医案原图片, 建立医案原文数据库。审核员进行两次文字校对, 合格者纳入医案数据库, 以保证医案信息的完整性与准确性。

2.2 名词术语规范化

参考《中医临床诊疗术语》《中国中医药学主题词表》《中国图书馆分类法·医学专业分类表》《中医药学名词》《中医大辞典》《中药大辞典》《中医临床常见症状术语规范》、全国中医药规范教材等建立标准库, 为后续医案切分等数据处理提供依据。鉴于中医病名与西医病名并非一一对应的关系, 采取西医病名与中医病证结合的模式更有利于疾病的研究。对症状的处理主要是将复合症状拆分为单独症状, 张启明等^[10]总结了复合症状的拆

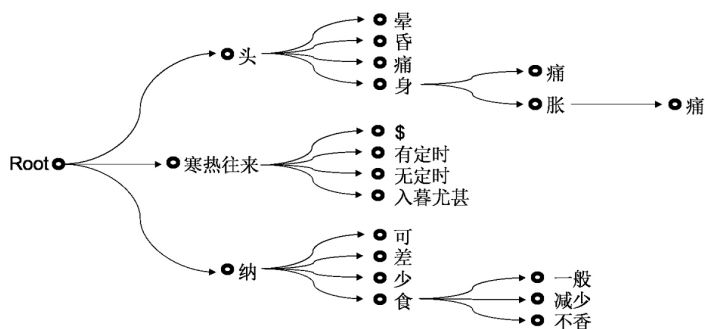
分原则,可为此提供参考。临床检验与影像学检查结果名称要规范,数值单位要统一。两方甚至多方合用的方剂需拆分为单个方剂,名老中医自拟方剂需先进行标记,其中未命名方剂可根据方药组成等自行命名,并添加标记。药物因取材部位、炮炙方法、药源产地等差异使用别称者,均统一为规范名称。初步的名词术语规范化处理能够消除冗余信息,清除部分数据噪声,使数据结构更加合理,便于分词算法的统计与识别,以提升切分质量。

2.3 医案切分

自然语言处理(NLP)是信息时代最重要的技术之一,HanLP 作为 NLP 最常用、最普遍的工具,相较于其他分词软件,具有基于多种算法的中文分词、词性标注、命名实体识别、关键词提取等功能,同时支持用户自定义词典,其性能高效、架构清晰的优点,为医案切分和语义标注提供了极大便利。

2.3.1 数据准备 运用正则表达式对数字化的医案文本进行初步清洗,去停用词处理。匹配以“无”“不”等否定词开头并含有顿号的字符串,并将该否定词添加至所有顿号之后;清除“量”“色”“质”等相关短语前的逗号,合并至前一字符串;去除对文本特征无贡献作用的字词,如“今日症见”“就诊可见”及一些语气词、助词等,得到清洗后的数据。

2.3.2 自动分词 首先,基于标准库中有关症状的语句,如图 3 所示构建 Trie 树结构的症状自定义词典。其次,人工标注一定数量的医案文本构成语料库,并借此训练症状术语识别模型;再通过少量未经标注的生语料对模型进行测试与修正,得到最终分词模型。最后,运用最大概率分词算法结合自定义词典对上一步清洗后的文本进行粗切分,再通过基于最终分词模型的 HMM 分词算法进行新词识别与词性标注。



注: Root, 根节点; \$, 终止符。

图3 医案数据库症状术语的 Trie 树结构词典示例

2.3.3 人工核查 对分词与词性标注结果进行人工核查,可利用字段匹配结合同义词替换功能对医案文本进行深度标准化,将未出现过的新词经人工审查后更新至标准库,以确保输出数据的准确性与数据存储的逻辑性。

2.4 语义标注

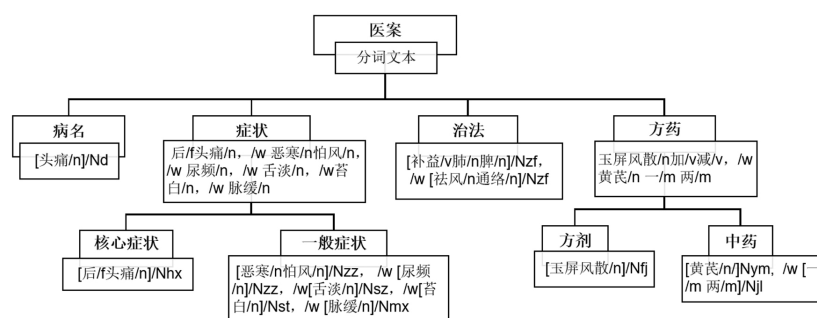
依据中医术语名词的首字母建立中医词性对照表,并基于分词结果进行如图 4 所示的语义标注。语义标注即为文本中所有常用字段添加概念标签的过程。语义标注的含义与词性标注相类似,均表示为切分所得词语添加一种类别标签,词性标注是根据语法特征进行标注,而语义标注则是根据概念。中医医案是中医诊疗活动的记录,历代医案内容虽各不相同,但记录形式却并无大异,均大致按照病名、症状、治法、方药等的顺序叙述,这种语言单元的规律性为语义标注提供了巨大的便利。利用 HanLP 分词系统基于 HMM 的命名实体识别功能,对文本中具有特定意义的实体进行识别,能够极大地加速这一进程。

进行症状信息的语义标注时,对核心症状采用多级标注的方法进行处理。核心症状是指那些对证型的判断具有较强辨识意义的主症,如患者就诊最主要的症状或体征,医案中描述较为详细的症状或体征,以及医案评述部分着重分析的症状或体征。多级标注即为将整体的症状语句按照派生部位、派生性质、限定因素和伴随症状拆分为次一级的基本短语,派生性质又可根据发作的形式、程度、特点进行再次一级的拆分。例如,发生在头部的各种疼痛,部位可位于前额、巅顶或后头;发作形式有阵发性发作、持续性发作,发作程度有重度、中度、轻度等,疼痛特点有胀痛、刺痛、空痛等;限定因素有劳累后加重、夜间加重等;伴随症状可包括眼眶痛、颈项痛等。如图 5 所示。

对证候信息进行标注时,引入中医证素对复杂的证候进行“降维”处理,使之分解为简单的证素,通过证素组合来揭示疾病的本质以及演变规律。证素是构成证候的基本要素,主要包含病位和病性两方面。将证候名称中涉及到的证素拆分为病位证素和病性证素,如两病性证素组合证候气虚血瘀证,其病性证素为气虚和血瘀;一病性加一病位证素组合证候肾气虚证,则其病位因素是肾,病性因素是气虚。三、四证素证候均按上述原则拆分。

2.5 建立关键词库

参考 HanLP 分词系统关键词提取功能所返还



注: /n, 名词; /w, 标点符号; /f, 方位词; /v, 动词; /m, 数词; /Nd, 病名; /Nzf, 治法; /Nhx, 核心症状; /Nzz, 症状; /Nsz, 舌质; /Nst, 舌苔; /Nmx, 脉象; /Nfj, 方剂; /Nym, 药名; /Njl, 剂量。

图 4 现代名老中医医案的语义标注示例

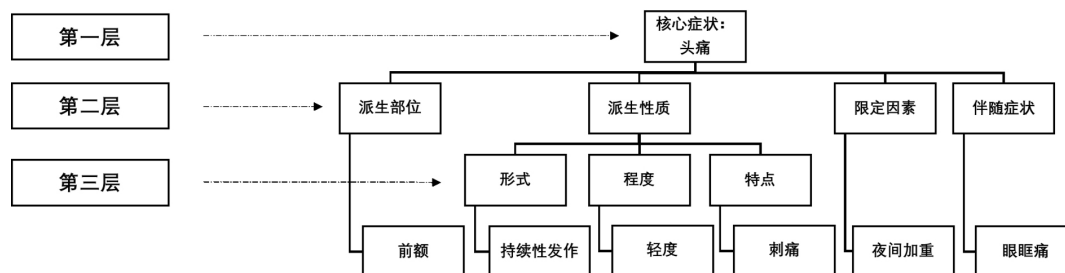


图 5 现代名老中医医案核心症状标注的概念层次

的结果, 结合人工分析为每篇医案选取 3~5 个关键词构成关键词组。选取的关键词必须要具有代表性、专指性和可检索性, 可以是病名、症状、证型、方剂、中药等, 亦可以是归纳总结的能够体现诊疗特色的医学术语或短语。

2.6 医案信息提取

语义标注后的文本具有很强的结构性, 通过 python 编程语言即可实现医案信息提取模型的建立。利用 python 结合正则表达式依据中医词性对照表依次对单个文本进行指定字符提取并聚类, 得到适配信息提取模型的语句, 形成结构化数据, 最终完成对所有医案文本的特定信息提取, 得到医案信息框架实例对象集, 形成医案信息数据库。

3 讨论

现代名老中医是中医药学术发展的杰出代表。他们的医案数量充足, 覆盖临床常见疾病, 语言结构清晰, 逻辑明确, 蕴含丰富的临证经验与诊疗特色, 具有极高的挖掘价值^[1]。尽管数据库技术已广泛应用其中, 但多集中在挖掘方法的选择与创新层面, 在实际操作过程中, 充分的数据处理是进行深度挖掘的前提, 尤其面对大量医案时, 如何兼顾数据处理的质量与效率, 是值得讨论的问题。

本研究数据处理的基本思想是借助自然语言处理技术对医案底层数据进行切分与标注, 以完成特

定字段信息的提取, 其效果优劣的关键在于分词方法的选择。目前常用的分词方法有以下三类: 基于字符串匹配的分词方法、基于理解的分词方法以及基于统计的分词方法。医案信息的数量庞杂, 运用基于字符串匹配的分词方法效率低下, 加之基于理解的分词方法目前发展并不成熟, 医案语言的复杂性亦使其难以转化为计算机直接读取的形式, 故采用基于统计的分词方法较为合适。在具体方法的选择上, 采用结合最大概率算法的 HMM 分词算法。由于医案文本所包含的大多数信息都在规范用语数据库中有迹可循, 为基于词典的最大概率算法提供了广阔的应用空间。然而, 这种分词方法并不具备新词识别与歧义处理能力, 对词典以外的医案信息分词效果较差^[2]。作为补充与完善, HMM 分词相比于其他算法更适用于医案这种信息规律较为固定的文本, 既保证了切分的质量, 其词性标注、命名实体识别等功能亦为后续的处理步骤提供了帮助, 极大的节约了人力成本与时间。

症状信息的处理是整个过程中最复杂之处。症状信息中存在着大量的同义词与同位词, 为分词识别带来了一定的困难, 延长了运算时间。通过仔细分析, 这些同义词与同位词大部分具有共前缀或后缀的特点, 如“纳可”“纳食可”, “口渴欲饮”“口渴不欲饮”等。利用这一特征, 构建症状术语的

Trie 树结构词典, 牺牲部分储存空间, 换取有序的数据结构, 能够加快这类词语的识别进程, 从而提升分词算法的运行效率。在症状信息的提取方面, 过于简化会降低挖掘结果的实用性, 过于具体则会产生庞大的噪声数据影响数据库的运行。针对医案中复杂且多样的症状描述, 刘保延等^[3]在构建中医临床科研一体化平台时提出症状具有独立性和内涵最小性, 并指出中医症状应具有 6 种特有属性和 4 种偶有属性, 对症状的表述进行了规范。有学者^[4]在对中医症状术语进行规范化研究时, 提出提取核心症状的同时应保留症状的逻辑关系, 以助于分析其病因病机。充足的信息获取是准确辨证论治的基础, 辨证论治作为中医诊病的核心环节, 其依据不仅仅是一系列的症状组合, 更在于症状特征的收集。在语义标注的基础上对核心症状进行多级标注, 整合并保留更多指导辨证的特征信息, 有利于探求症状与证型间的特征性联系, 提升数据的挖掘质量。

结合自然语言处理技术的数据处理方法具有一定的实用性, 促进了医案信息提取过程的系统化与规范化, 为现代名老中医医案数据库的构建奠定了基础。医案数据库的构建是一个不断完善、不断发展的过程, 也是一项复杂且繁重的跨学科工作, 需要中医科研者与程序员的通力协作。深度发掘医案中不便于统计分析的隐含规律有待进一步研究。

参考文献

- [1] 周京艳, 杨萍, 王世雯. 我国中医药信息资源数字化建设的现状 [J]. 中华医学图书情报杂志, 2016, 25(7): 49-52.
- [2] 于琦, 崔蒙, 李园白, 等. 中医药文献数据库建设规范研究 [J]. 世界科学技术-中医药现代化, 2014, 16(11):

2304-2307.

- [3] 武辉, 马路. 中医医案数据库的研发现状及展望 [J]. 中华医学图书情报杂志, 2012, 21(11): 49-52.
- [4] 吴中平, 何新慧, 柯雪帆, 等. 医案查询统计分析系统介绍 [J]. 上海中医药杂志, 2003, 37(3): 54-56.
- [5] 张启明, 王永炎, 张志斌, 等. 中医历代医案数据库的建立与统计方法 [J]. 山东中医药大学学报, 2005, 29(4): 298-299.
- [6] 陈丽平, 李建生, 蔡永敏, 等. 中医肺病数据库构建及挖掘的思路与方法 [J]. 中华中医药杂志, 2017, 32(12): 5530-5534.
- [7] 郭晓燕. 基于数据挖掘技术的原发性高血压医案知识发现 [D]. 广州: 广州中医药大学, 2019.
- [8] GUIMARAES M. Data warehouse and data mining with SQL server [J]. J Comput Sci Coll, 2018, 34(1): 233-234.
- [9] 翁敏峰, 万城春. SQL Server 数据库中关联数据表的设计方法探讨 [J]. 电脑知识与技术, 2019, 15(28): 17-18.
- [10] 张启明, 张振中, 李檬, 等. 作为科技术语的中医症状的命名 [J]. 北京中医药大学学报, 2007, 49(12): 797-799.
- [11] 方旖旎, 王琦, 李竹青, 等. 名老中医学术思想传承研究的现状综述 [J]. 环球中医药, 2020, 13(10): 1818-1822.
- [12] 梁礼铿, 黎敬波. 基于最大概率法探讨中医症状信息提取与标准化 [J]. 中华中医药杂志, 2017, 32(5): 2159-2162.
- [13] 刘保延, 张启明. 构建中医临床科研信息一体化平台需要解决的症状规范问题 [J]. 中医杂志, 2011, 52(20): 1714-1716.
- [14] 张妮楠, 曹馨宇, 林睿凡, 等. 癫痫中医症状术语规范化研究 [J]. 世界科学技术-中医药现代化, 2020, 22(5): 1386-1391.

Construction and Data Processing of Medical Cases Database from Modern Famous Doctors of Traditional Chinese Medicine

ZHAO Zepeng¹, DAI Guohua², GAO Wulin²

1. Shandong University of Traditional Chinese Medicine, Jinan, 250014; 2. The Affiliated Hospital of Shandong University of Traditional Chinese Medicine

ABSTRACT Data mining technology is a commonly used method to induct and summarize the experience of famous doctors based on the constructed database of medical cases. To promote the use of the data mining results and to explore the method suitable for constructing the database of medical cases from modern famous veteran doctors of traditional Chinese medicine (TCM), this study optimized the database structure and common fields, used HanLP word segmentation system in combination with maximum probability algorithm-based hidden Markov model (HMM) framework to perform text segmentation and semantic annotation, and extracted the data according to the annotated semantics through python technique. This method helps to optimize the data structure of medical records, and improves the efficiency of database construction and the quality of data mining.

Keywords medical cases of traditional Chinese medicine; database; famous doctors of traditional Chinese medicine; data processing

(收稿日期: 2020-12-14; 修回日期: 2021-03-22)

编辑: 贾维娜