

正则表达式对疫病古方药物标准化处理方法的探讨

朱靓贤¹ 苏前敏² 刘晨蕾² 文小平^{1*}

(1. 上海中医药大学中药学院方剂教研室, 上海市浦东新区蔡伦路 1200 号, 201203;

2. 上海工程技术大学)

基金项目: 国家中医药管理局中医药科学技术专项(06—07JP18)

* 通讯作者: wxpglj@163.com, 13636636328

[摘要] 目的 将历代治疗疫病古方中药物名称与现代国颁标准名称统一, 以实现古方药物名称标准化, 为挖掘药物配伍关系、方证(症)关系及新药开发等提供依据。方法 根据国颁标准制作药物标准文档, 运用 SQL Server 2005 和 Visual Studio 2008 软件, 采用正则表达式批量化处理古方药物数据, 实现计算机批量处理非标准药物。结果 实现了“疫病”数据库药物组成的标准化。结论 计算机程序可辅助实现古方药物名称标准化, 减少人工处理海量非标准化中药数据, 对中医现代化研究与发展有推进作用。

[关键词] 中药名称标准化; 正则表达式; 中医文献整理

在疫病方剂数据库的建立和整理过程中, 首先要对药物名称进行标准化处理, 在此过程中遇到了诸多问题。我们将传统处理方法与现代计算机相结合, 运用正则表达式提取所需药名信息, 批量化去除药物剂量、炮制方法等冗余数据, 并编程将药名全部替换为国颁标准药名, 证明了批量处理数据的可行性, 兹介绍如下。

1 中药名称标准化的研究现状与问题

方剂数据挖掘类文献报道中, 关于中药名称标准化问题的论述目前没有统一标准和规范, 有以《中华人民共和国药典》(简称《药典》)为准, 有以《中华本草》为准, 有以《中药学》教材为准, 有的没有明确表示所采用的标准。如王家平^[1]研究唐宋金元时期虫类药治疗痹证方, 以《药典》(2005 年版)为主, 次用《中华本草》对药名进行统一; 吴童^[2]研究消渴病文献中的药物则以六版《中药学》教材为主, 参考《中药辞海》规范药名。戴水平^[3]的活血化瘀方剂数据库参考《中华本草》、《中药学》教材, 《中医药学名词》、《中药名大辞典》、《中药别名辞典》等标准药名。而毛文姣^[4]的疮疡古代文献整理及研究, 张林^[5]的古代肝病方剂组方规律的研究等文章, 未明确表示以何书为参考来进行药名标准化。

方剂药物名称在历史长河的变迁中的确称谓很多, 使其按国颁标准统一问题太多, 诚如阮时宝等指出^[6], 古代方剂文献中药物标准化的难点在于药物数据庞杂。历史跨度大导致各朝代药物称谓差异、度量衡不统一、传抄时的错误以及数据处理过

程中的人为错误等, 均是影响药物标准化的原因。因此药物标准化的过程需要耗费大量的时间和人力资源。

2 选择正则表达式处理数据的原因

正则表达式由一些普通字符和一些元字符组成, 是记录文本规则的代码, 可用于查找符合某些复杂规则的字符串, 经常运用于编写处理字符串的程序或网页等情况^[7]。由于古方药物文献包含大量文本信息, 成分复杂, 若仅靠人工将古代记载药名修改为现国颁标准药名, 去除非药名数据, 工作量相当繁重而费时费力。若结合正则表达式保留有用信息, 批量化去除冗余数据, 可取代大部分人工修改数据的工作, 且能减少人为错误。如古方中书写用药信息多为“当归三两、白芍、川芎各二两……”等, 正则表达式可以设计出提取有效文本信息程序, 达到批量化去除药物名称以外的数量词等冗余数据, 保留药物名称的目的, 为药名标准化节省了大量人工处理时间。

3 药名标准化处理步骤介绍

3.1 数据库简介

本课题组收集与“疫病”相关的方剂内容, 查阅了东汉至明清的古医籍约 400 余册, 在众多疾病中选择确立“疫病”21 种; 涉及方剂 8779 首, 建立了历代“疫病”方剂数据库, 涉及朝代、作者、书名、方名、组成、剂量、功效、主治、用法等内容。

3.2 数据处理步骤与注意事项

3.2.1 批量化处理剂量 首先运用正则表达式去

除药物附带的数量词,我们运用 SQL Server 2005 作为数据存储平台,通过 Visual Studio 2008 用 C 语言对数据进行调用,采用正则表达式实现数量单位的批量去除。在初次删除数量词后,我们发现,尚存诸多具有中国历史特色的量词需要处理,如“一方寸”、“三筒子”、“九十箔”、“三挺”、“一斛”,还有估计大小的词如“二皂子大”、“半栗子大”、“如拳大者”、“豆大”等。为此我们将这些特殊量词再次补充到正则表达式中完善程序,并统计出经计算机批量处理后的这些非通用古代度量衡单位约 60 个,可以为今后研究古代度量衡演变提供参考依据。在上述批量处理过程中对含有数量单位和数字的中药、方剂,如半夏、三七、五加皮、百合、三颗针、半枝莲、两头尖、六一散等,要进行特殊处理,否则计算机将按数字处理而将其删除,导致数据丢失,使资料统计不完整。

3.2.2 批量化处理炮制 除了上述提及的药物数量词外,古籍文献中记述用药方法时,还会夹带某些药物的炮制方法,如“茯苓(去皮)”、“罌粟壳(蜜炙)”、“人参(去芦)”等,这类情况也可运用正则表达式批量化去除带括号部分的数据。以上两步均运用正则表达式实现批量化处理,现举例说明处理效果,见表 1。

表 1 药物名称正则表达式处理效果举例

原始数据	正则表达式去除数量词	正则表达式去除炮制
金头蜈蚣一枚(去头足,炙)、草乌头半两(去芦头)、天麻半两、全蝎十个、香白芷	金头蜈蚣(去头足,炙)、草乌头(去芦头)、天麻、全蝎、香白芷	金头蜈蚣、草乌头、天麻、全蝎、香白芷

由表 1 可见,结合使用正则表达式能批量化去除药物组成中包括数量词、药物炮制方法等非药物名称的数据,较传统手工去除冗余数据而言减少了大量工作。

3.2.3 药物名称唯一化 去除数量词等冗余数据的中药名会有很多重复,计算机应用程序将发挥优势实现去除重复中药名,达到每一味药物只有唯一药名的效果。

3.2.4 药物名称标准化 以《药典·一部》(2010 版)^[8]、《中华本草》^[9]、《中药大辞典》^[10]、《中药学》^[11]、《中药别名速查大辞典》^[12]等为药名标准化的参考书籍。在标准化过程中遇到的问题与解决方法如下:①异名同药的处理:根据国颁标准以及原始药物信息,整理出药物正名、异名信息表,用于药物标准名替换(举例见表 2),该表中的内容补充了

表 2 黄连、黄芩、黄柏正名异名对照表(部分)

标准药名	原始药物数据
黄连 (31)	炒川连、炒黄连、川古勇黄连、川古勇连、川黄连、川连、川雅连、大宣连、顶上真川连、好黄连、黄连、黄连末、姜炒黄连、净黄连、酒炒川连、酒炒黄连、酒连、酒蒸黄连、上川连、生川连、生黄连、小川连、小云连、宣黄连、宣连、雅连、元连、云黄连、云连、真川连、正雅连
黄芩 (20)	北黄芩、炒黄芩、炒芩、淡黄芩、淡芩、黄芩、酒炒黄芩、酒黄芩、酒芩、枯黄芩、枯芩、片黄芩、青子芩、生黄芩、实黄芩、条芩、条子芩、细黄芩、以黄芩、子芩
黄柏 (20)	炒黄柏、川柏、川柏末、川柏炭、川黄柏、厚黄柏、黄柏、黄柏末、黄柏皮、黄蘗、酒柏、酒炒川柏、酒炒黄柏、酒黄柏、片黄柏、生川柏、生黄柏、生少柏、真关黄、炙川柏

《中华本草》等书籍中未列出的药物异名信息。②药名连写处理:历代医家在案例书写中常有药物名称连写现象,如“乳没”、“紫苏茎叶”、“炒知柏”等,为使其标准化,我们对对应制作出替换文档程序,将“乳没”对应“乳香、没药”;“紫苏茎叶”对应“紫苏茎、紫苏叶”;“炒知柏”对应“知母、黄柏”等,对数据库中的连写数据全部替换成为国颁标准表达形式,以便今后药物的统计分析。③无国颁药名的处理:有部分古籍记载药名在现有《中华本草》、《药典》、《中药学》等书籍中均没有记载,如:桑甲、肉灰、栾荆等。对此,我们将其统一标记,咨询相关中药专家来确定其身份。如“真关黄”是“黄柏”,“炮甲”为“炮穿山甲”,“好信”是“砒石”,“大云”指“肉苁蓉”等,为更完善有效的数据利用提供了标准。同时我们也对用此类方法处理的数据标记存疑,以待今后有国标提供时再最终确认。④同名异药问题:有些同名药无法鉴别其究竟为何药,如“香藤”,《药典》、《中华本草》等“正名”中无此药,而在“异名”或“又名”中有香藤的药物有“天仙藤”、“山药”、“瓜子藤”、“筋藤”、“金狮藤”、“鸡屎藤”等多味,一时无法确认它该是何种药物,此类记载处理时需特殊标记数据,保存原貌不做任何改动,待今后有足够资料考证后再另行处理。⑤方名代药问题:方剂组成中常夹有方名,如《时方歌括·卷上》中治疟病初起用“小柴胡去参加青皮汤”,该药物组成中的“小柴胡”在《中医方剂大辞典》^[12]中记载了历代同名方 19 首,其中 12 首药物包含人参,该条描述中的“小柴胡”是指哪首方,该用哪些药物来替换这一方剂组成,即以何方为标准拆方缺乏依据时,则需要全面分析此方的出处、时代背景、作者等资料的情况下,才能做初步判断则拆之,而不可主观臆断为某方而拆之。

3.2.5 药名标准化批量处理 在完成上述数据库资料问题处理后,运用设计程序,根据“药名标准文档”由计算机批量化读取出原始文档,用程序进行多表关联并更新,把药物名称的原始数据替换为国颁标准中药名,为今后的数据精确搜索及分析提供有利支持。替换效果举例见表3。

表3 药物组成替换前后比较举例

替换前	替换后
鬼箭羽、鲮鲤甲	鬼箭羽、穿山甲
贯众、黑豆、甘草、绿豆、信	贯众、黑大豆、甘草、绿豆、砒石
金头蜈蚣、草乌头、天麻、全蝎	蜈蚣、草乌头、天麻、全蝎、白芷
香白芷	

由表3可见,运用程序将药名替换为标准药物名,既可减少人员的工作量,又可避免人工修改出错的可能。

3.2.6 校验结果 在计算机批量完成标准药名替换后,再次用数据库特有编号替换标准后药物名称,核查修改批量化处理效果,找寻未替换成编号的单词,据此既可以修缮正则表达式等数据处理程序,使“疫病”数据库药物内容准确无误,又可以作为今后古方文献的药名标准化处理的基础。

4 讨论和展望

4.1 人机结合的必要性

面对古方浩瀚的数据,在整理过程中既要符合中医专业特点,又需不断调整运用正则表达式等先进计算机数据处理方法,达到在处理大量、复杂的古方文本信息过程中,快速地去冗余数据、标准不规范的数据表达,减少人工处理数据的工作量。而无法用计算机编程实现的问题,如药名替换文档制作等,则需要传统人工检索、考证等方式完成。人机结合方能事半功倍地使数据精确和完整的,这是进一步进行数据检索、数据挖掘等研究的基础和关键。

4.2 资源整合的迫切性

中医学内容浩瀚,利用计算机对中医文献整理

是趋势也是捷径,然而古代方药内容中,尚存有部分待考证的内容,其数据能否被运用是提高数据利用率的关键。目前各大中医院校和研究机构的文献研究项目其内容和侧重点各不相同,但都涉及相同资料,因此,若能把各自研究领域中的问题和解决方法在一个统一的平台上讨论,将会对中医文献中难点和疑点的解决有一定的帮助。

参考文献

- [1]王家平.唐宋金元时期应用虫类药物治痢病的文献研究[D].新疆医科大学,2009.
- [2]吴童.消渴病中医文献与证治方药规律研究[D].黑龙江中医药大学,2005.
- [3]戴水平.中医活血化瘀方数据库及其数据挖掘[D].北京中医药大学,2007.
- [4]毛文姣.痤疮古代文献整理及研究[D].南京中医药大学,2009.
- [5]张林.计算机辅助古代肝病方剂组方规律的研究及古代肝病方剂数据库初探[D].北京中医药大学,2004.
- [6]阮时宝,刘益臻,游世晶.数据挖掘在方剂配伍规律研究中的现状与存在的问题[J].光明中医,2008,23(6):715—716.
- [7]沙金.精通正则表达式:基于.NET/ASP/PHP/JSP/Java Script[M].北京:人民邮电出版社,2008.
- [8]国家药典委员会.中华人民共和国药典·一部(2010版)[M].北京:中国医药科技出版社,2010.
- [9]国家中医药管理局《中华本草》编委会.中华本草[M].上海:上海科学技术出版社,1999.
- [10]江苏新医学院.中药大词典[M].上海:上海人民出版社,1977.
- [11]黄兆胜.中药学[M].北京:人民卫生出版社,2002.
- [12]李顺宝.中药别名速查大辞典[M].北京:学苑出版社,1996.
- [13]彭怀仁.中医方剂大辞典[M].北京:人民卫生出版社,1997.

(收稿日期:2011-07-14;修回日期:2011-10-09)

[编辑:黄健]

长期销售空心胶囊

本产品严格按照国家标准和《中国药典》规定的质量要求生产,符合药用各项指标,有0号、1号、2号等多种型号、多种颜色可供选择。本公司以合理的价格欢迎单位和个人来电来函咨询。对用量少的单位和个人将办理邮购业务。价格:每万粒135元(含邮资)。

地址:浙江新昌县一诺胶囊有限公司,邮编:312560。电话:(0575)86061888,013706852676,联系人:潘炎峰。

欢迎邮购《中医杂志》光盘合订本

1951—1986年(4张光盘)480元;1987—1994年(1张光盘)260元;1995—2004年(1张光盘)280元。光盘合订本为计算机用数据光盘,收入了同期出版的全部《中医杂志》的内容,可对所有文章进行任意字检索,方便实用。

汇款地址:北京东直门内南小街16号中医杂志北京医海林音像书店,邮政编码:100700,电话:(010)64035632。网址:<http://www.jcm.net.cn> <http://www.cnki.net>