“互联网 +”背景下数据挖掘在中医药领域的应用研究*

刘雅芳 杨玉赫

郭竹英

孙素芹

(黑龙江中医药大学
哈尔滨 150040)(上海交通大学医学院附属第九人民
医院检验科 上海 201999)(黑龙江中医药大学附属第二医院
哈尔滨 150001)

〔摘要〕 阐述“互联网+”背景下数据挖掘在中医药领域的应用,包括中药领域、中医药文献和中医诊断的数据挖掘研究,分析存在的信息缺失、隐私保护、数据繁杂、不规范等问题,探讨未来发展趋势。

〔关键词〕 互联网;数据挖掘;中医药;网络信息

〔中图分类号〕 R-056 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2018.08.012

Study on Application of Data Mining in the Traditional Chinese Medicine Field against the "Internet + " Background LIU Ya-fang, YANG Yu-he, Heilongjiang University of Traditional Chinese Medicine, Harbin 150040, China; GUO Zhu-ying, Department of Laboratory, School of Medical, Shanghai Jiaotong University Affiliated The Ninth People's Hospital, Shanghai 201999, China; SUN Su-qin, Second Affiliated Hospital of Heilongjiang University of Traditional Chinese Medicine, Harbin 150001, China

〔Abstract〕 The paper expatiates on the application of data mining in the Traditional Chinese Medicine (TCM) field against the "Internet + " background, including the TCM field, TCM literature and study on data mining of TCM diagnosis, analyzes the existing problems like absence of information, privacy protection, complicated, redundant and nonstandard data, and discusses the future development tendency.

〔Keywords〕 Internet; Data mining; Traditional Chinese Medicine (TCM); Network information

〔修回日期〕 2018-04-18

〔作者简介〕 刘雅芳,博士,副教授,发表论文 20 余篇;
通讯作者:孙素芹,主治医师。

〔基金项目〕 黑龙江省青年科学基金项目“基于 KDD 技术对中国古代抑郁症医案的系统研究”(项目编号:QC2014C118);黑龙江中医药中青年科技攻关项目“基于 KDD 技术对中国古代抑郁症医案的系统研究”(项目编号:ZQG-026)。

1 引言

随着社会的发展,互联网与大众生活的关系越来越密切。国家与社会对“互联网+”的发展日益重视,大数据、数据挖掘更是吸引诸多学者的关注,掀起新的发展浪潮。数据挖掘指从大量的、有噪声的、不完整的、模糊、随机的数据中通过算法搜索隐藏于其中信息的过程^[1]。以此来预测行业发展趋势的隐含模式或规律。2015 年 7 月国务院印发《关于积极推进“互联网+”行动的指导意见》,该文件的制定有助于提升“互联网+”相关技术在各

产业应用水平,增强各行业的创新能力。中医药是中华民族瑰宝,中医药的数据挖掘是其现代化研究的重要组成部分。

2 “互联网+”背景下数据挖掘在中医药领域的应用

2.1 概述

“互联网+”主要是指“互联网+各传统行业”,是指以互联网平台为基础,通过运用信息技术使互联网与其他传统行业有机结合,创造出更新颖的发展方向。数据挖掘是指通过多种方法在大量堆积以及无序数据中将人们感兴趣的有效信息剥离出来,可以将过去和现在的信息进行对比与研究,帮助人们找到潜在的规律与信息,从而更好地预测未知的事情^[1]。在“互联网+”背景下,数据挖掘能够以特殊的视角使大数据技术、信息检索技术与统计学、数据可视化、机器学习与人工智能多个不同的领域完美地组合在一起。

2.2 中药领域的数据挖掘

2.2.1 中药识别 中药对鉴定以及质量监控的现代化、统一化等方面的研究逐渐成为国内外学者们关注的热点^[2]。中药材种类多样,来源丰富。用量方法因药材品种不同而有所不同。同时中药组分复杂、交互作用多,且具有多成分、多靶点、有效成分不易确定等特点,因此对中药领域的数据挖掘至关重要。就目前技术而言,基于数据挖掘的中药识别方法主要有图像识别和指纹图谱。张小波^[3]等探究图像识别技术在全国中药资源普查中的应用情况,发现该技术在文字、卫星遥感、植物个体、药材样品和腊叶标本等 5 类照片数据的基础上对中药种类、分布和数量识别起到至关重要的作用。郝燕^[4]等对评价中药色谱指纹图谱多维多息 37 个特征参数进行数据挖掘,包括 F 和 I。研究表明 4 个主成分可以证实中药色谱指纹图谱信息规律。

2.2.2 新药开发 经过长期医疗方面的探索与实践,中药已经有明确的功能应用和性味归经。在已有的记录中寻找新的医用成分或不同的先导化合

物,都是研发新型药物的有效途径。通过数据挖掘能够缩短开发和研究新型药物的时间。向阳峰^[5]利用计算机数据挖掘的优势,提出新的新药研究方案,通过对经典数据挖掘算法的改进,使其在新药研发的系统平台得到有效运用。黄牛^[6]在研究中发现利用计算机数据挖掘辅助药物分子设计已经成为新药研制的重要途径,利用图形学显示系统,结合分子力学、分子动力学和量子化学等计算方法来指导药物设计已成为当今药物化学发展和新药开发的前沿领域。

2.2.3 中药方剂及配伍规律 利用数据挖掘技术可对单一复方配伍规律以及许多复方开展分析,总结相关规律,寻找到一些隐匿的规律^[6]。蒋志滨^[7]将方剂配伍规律数据挖掘文献做出口,对常用的数据挖掘技术及方剂配伍规律研究中的应用进行系统梳理,指出相关研究中存在的不足,通过晚期胃癌方剂数据挖掘分析实例,结合自主设计的数据预处理系统与开源数据挖掘平台说明该策略与方法在方剂配伍规律研究中的具体应用。李振岳^[8]采用频数分析、频繁集、关联规则方法进行数据统计:对证型、方剂、药物等进行频数分析,再运用频繁集和关联规则对药物组合进行分析,探讨药物的配伍规律。温俊茂^[9]等采用数据挖掘方法分析慢性阻塞性肺疾病稳定期的方剂配伍规律,将 133 个方剂按照药物矩阵和簇集分为不同类型,得出慢性阻塞性肺疾病稳定期的治疗方法重在益脾肺气、理气化痰、活血化瘀、润肺敛阴、滋肾温阳的结论。

2.3 中医药文献的数据挖掘

2.3.1 概述 中医药文献数据主要来自古典文献、验尸验案等。数据挖掘技术可以对名家的中药复方、学术思想、辨证方法等研究成果重新发现和总结,将中医药继续传承和发扬。中医药数据具有非线性和模糊性特点,针对中医药数据具体性与挖掘目标的不同性,将不同方法结合进行处理,通过这种方式来解决面临的诸多问题。

2.3.2 频数分析 目前已有相关研究人员证实在中医诊断中围绝经期综合征中主要存在近 20 个证候,包括肝肾阴虚、肾阳虚、肾阴虚及肾阴阳两虚

等^[10]。心肾不交、肝郁气滞及脾肾阳虚最为常见。其中常见症状共 65 种, 涵盖如月经紊乱、易怒、失眠与烘热汗出、头晕耳鸣、纳呆以及腰膝酸软、心悸等。在病位方面, 主要以心、肝、脾、肾为主。在病性方面, 主要是虚主。在脉象方面, 主要是细数、沉细、弦细为主。在舌象方面, 主要是舌淡或舌红、苔薄或少苔多见^[10]。通过诸如此类的研究证实围绝经期综合征的中医证候分布相对集中^[10]。吕倩倩^[11]等通过收集中药治疗癌性发热文献中的处方用药, 以频数分析探讨现代临床中药治疗癌性发热的组方原则及规律, 认为中药治疗癌性发热应清补结合、标本兼顾、以治本为主。彭逢春^[12]等通过对临床数据的分析探索眩晕口服方药组方的常用药物, 得出有效中药 132 味, 发现临床用药以味温平、味甘辛, 入肝脾经和补气类的药物为主。

2.3.3 聚类分析 有研究者根据文献中的记载对代谢综合征 (Medtabdic Syndrome, MS) 中医辨证分型进行聚类分析与频数统计。研究证实痰淤互结、脾虚痰湿、肝胃郁热、气阴两虚、肾阳亏虚、血癖脉络、阴虚热盛是 MS 的常见证型^[13]。钟小雪^[14]通过收集 1 019 例血脂异常患者, 采用聚类分析和对应分析对其四诊信息进行中医证候及证候要素的非线性降维和相关性研究, 血脂异常的中医证候可分为 6 类: 气滞证、滚痰互阻证、痰浊阻滞证、脾肾阳虚证、脾气虚证和肾阳虚衰证。任毅^[15]等基于数据挖掘技术采用聚类分析法对当代名老中医治疗冠心病病案方药进行总结, 共分析方药 386 个, 结果显示老中医多以健脾化痰、行气活血等药物结合养心安神、调补阴阳等药物组合治疗冠心病。

2.3.4 文本挖掘 有研究者在中国生物医学文献数据库中查阅有关中医治疗肥胖以及高脂血症的中医文献, 采用数据挖掘方法探究中医临床用药规律, 研究证实大黄、黄芩、山楂、丹参、白术、泽泻及半夏是治疗肥胖的常用药物。这些药物中最为核心药物是白术、黄芩和大黄。大黄、丹参、泽泻是治疗高脂血症常用药物。此外还有山楂、黄芩、何首乌, 将丹参和山楂作为治疗的核心药物。研究

结果说明在治疗肥胖和高脂血症方面, 中医的处方既有相同之处又有不同^[16]。赵静^[17]等对中医药治疗流行性感胃文本挖掘结果及诊疗方案进行比较分析, 得出流感证候以风热犯卫的流感轻症为多见, 治疗方药以银翘散与麻杏石甘汤为主等结论。刘国山^[18]等用文本挖掘技术探索溃疡性结肠炎在中医药领域中的研究热点、证型、治则、用药, 认为溃疡性结肠炎在中医药领域中的研究主要集中在溃疡性结肠炎的中医药辨证论治; 证型有脾胃虚弱型、脾肾阳虚型、湿热内蕴型等; 治则有清热燥湿、健脾益气、涩肠止泻; 常用的中药有灌肠方、白头翁汤、锡类散等; 给药方式主要有中药灌肠、中药保留灌肠、中药内服等。

2.4 中医诊断的数据挖掘

2.4.1 症候 (1) 关联规则。通过观察中医数据中冠心病的规律, 运用以向量法为基础改进的数据关联规则挖掘算法。以脉象、舌质和发病诱因等 18 个角度与 8 类冠心病的中医证型中研究出中医辨证规律^[19]。(2) 粗糙集方法。以粗糙集约简模型为基础, 以较高质量分类能力得到保障为基础, 高效率提取出核心证候要素^[19]。(3) 因子分析。通过因子分析法, 针对冠心病等多个症状开展数据挖掘分类研究^[19]。提取出 6 个公因子, 通过中医辨证论治后确认为心脾两虚、心肾阳虚、气滞血癖、气虚痰浊、气虚血癖、气阴两虚。冠心病心绞痛的主要病机为气虚、痰浊和血癖。(4) 主成分分析法。有研究者通过文献出现的频率来对流行病学进行调查研究, 运用主成分分析法将火热证提取出 6 个重要成分^[20-21]。将这些成分分为 4 个类型, 这与临床医学者的判断大致相同。(5) 其他方法。通过单因素分析与聚类分析、Logistic 回归模型预测效果、多因素 Logistic 回归分析、Roc 曲线评价等方法, 研究出慢性肾炎症状特点与证型的关系^[21]。

2.4.2 诊法 (1) 支持向量机。有研究者通过支持向量机开展胆囊炎患者的舌诊舌象研究。其研究结果与专家诊断的结果大致相同。这说明支持向量机对舌面图像存在分类鉴别能力^[22]。(2) 模糊聚类。有研究者以改进模糊聚类算法为方法, 设计

彩色舌图像舌质舌苔分离法。该方法可以缩短系统运算时间,还能够极大地加快聚类迭代速度^[21]。相对于 17 值分割法,此方法得到的苔质分离效果更加符合中医学的要求。(3) 其他方法。通过两个光谱仪来采集同一舌体数据。通过运用主成分分析、偏最小二乘法、神经网络、间隔偏最小二乘法等多种方法,对采集到的上千例临床样本进行数据建模与研究。结果证实脂肪肝和高博血症等疾病与舌体反射光谱之间有较大联系^[22]。

2.5 存在的问题

中医药在漫长的发展过程中逐渐积累大量的数据与理论,其数据具有明显的特殊性,在“互联网+”背景下数据挖掘技术的使用存在一些问题。(1) 信息缺失。对中医治疗疾病的诊断、辨证、组方等方面进行数据挖掘,文字描述占有绝大部分的比重,从而过多地增添定性内容,导致部分信息的缺失,一定程度上造成中医数据的不准确性和不完整性。(2) 隐私保护。中医诊断的某些疾病可能涉及患者隐私,以及部分中医有其个人的治疗方式、治疗药物,这些都属于不可公开的信息,过度公开就可能导出一系列的法律问题。(3) 数据繁杂。中医在记录病症的程序中可能存在许多重复和无关的内容,如临床病历数据,可能患者一段时间内的数据大致相同^[22],导致数据的重复。另外某些疾病症状相似,不能简单根据文字描述进行诊断。(4) 不规范。漫长的发展历史以及个人因素导致中医数据不够规范,如不同的地域、不同的医生对同一病症存在不同的命名或理解。每种数据挖掘方法都有各自的不足,应不断探索,运用更多的方法进行处理,从而提高数据挖掘的论证性与实践性^[24]。

3 结语

在“互联网+”背景下进行中医药数据挖掘有着广阔的发展前景以及重要的研究价值。虽然尚存不足之处,但是可以在不断发展进程中探索全新的技术和思路。数据挖掘为中医药古典文献的保留研究以及更广泛的发展应用提供良好的机遇,通过现

代技术对中医药科学内涵进行证明,加快中医药学术水平的提升。

参考文献

- 1 刘建平,张柯欣,杨钧,等. 数据挖掘技术及其在中医药领域中的应用 [J]. 辽宁中医药大学学报, 2007, 9 (6): 203 - 204.
- 2 秦雪君,施诚. 数据挖掘技术在中医药领域的应用 [J]. 医学信息, 2006, 19 (6): 945 - 947.
- 3 张小波,格小光,金艳,等. 图像识别技术在全国中药资源普查中的应用 [J]. 中国中药杂志, 2017, 42 (22): 4266 - 4270.
- 4 郝燕,董鸿晔,姜楠,等. 基于主成分分析的中药色谱指纹图谱多维多息特征数据挖掘方法研究 [J]. 中南药学, 2007, 5 (3): 267 - 272.
- 5 向杨峰. 基于数据挖掘的新药研发系统 [D]. 北京: 北京交通大学, 2010.
- 6 黄牛. 计算机模拟药物筛选在新药设计与开发中的应用 [J]. 基础医学与临床, 2001, 21 (4): 298 - 301.
- 7 蒋志滨. 基于数据挖掘的方剂配伍规律研究方法探讨 [D]. 南京: 南京中医药大学, 2015.
- 8 李振岳. 中药方剂数据挖掘研究 [D]. 广州: 广东药学院, 2010.
- 9 温俊茂,曾锐祥,廖鹏达,等. 基于数据挖掘的 COPD 稳定期中药配伍规律分析 [J]. 中国中医急症, 2016, 25 (11): 2026 - 2028, 2035.
- 10 李文林,赵国平. 数据挖掘技术及其在中医药领域的应用 [J]. 中华医学图书情报杂志, 2004, 13 (6): 4 - 6.
- 11 吕倩倩,方文岩. 基于频数分析及聚类分析法分析中药治疗癌性发热的用药规律 [J]. 贵阳中医学院学报, 2017, 39 (3): 45 - 48.
- 12 彭逢春,肖瑶,王净净. 中药复方治疗眩晕的用药频数分析 [J]. 世界中西医结合杂志, 2012, 7 (5): 369 - 371.
- 13 周华英,张启蕊. 信息技术在中医药领域应用研究概况 [J]. 中国中医药信息杂志, 2011, 3 (18): 110 - 112.
- 14 钟小雪. 基于聚类分析和对应分析的 1019 例血脂异常患者的中医证候研究 [D]. 北京: 北京中医药大学, 2017.
- 15 任毅,陈志强,张敏州,等. 当代名老中医治疗冠心病用药规律的聚类分析 [J]. 中国中西医结合杂志, 2016, 36 (4): 411 - 414.

(下转第 72 页)

参考文献

- 1 袁泉, 赵辅昆. 蛋白质组研究新前沿: 定量蛋白质组学 [J]. 生物化学与生物物理学报: 英文版, 2001, 33 (5): 477 - 482.
- 2 Matthiesen R. Mass Spectrometry Data Analysis in Proteomics [M]. Clifton: Humana Press, 2013: 1 - 6.
- 3 Chervitz SA, Deutsch EW, Field D, et al. Data Standards for Omics Data: the basis of data sharing and reuse [M]. Clifton: Humana Press, 2011: 31 - 69.
- 4 JD Tenenbaum, SA Sansone. A Sea of Standards for Omics Data: sink or swim [J]. Am Med Inform Assoc, 2014, 21 (2): 200 - 203.
- 5 Taylor CF, Paton NW, Lilley KS, et al. The Minimum Information about a Proteomics Experiment (MIAPE) [J]. Nature Biotechnology, 2007, 25 (8): 887.
- 6 Vizcaíno JA, Côté RG, Csordas A, et al. The PRoteomics IDentifications (PRIDE) Database and Associated Tools: status in 2013 [J]. Nucleic Acids Research, 2013, 41 (Database issue): 1063 - 1069.
- 7 Mayer G, Montecchi - Palazzi L, Ovelleiro D, et al. The HUPO Proteomics Standards Initiative - mass Spectrometry Controlled Vocabulary [J]. Database the Journal of Biological Databases & Curation, 2013, 2013 (10): bat009.
- 8 Musen MA. National Cancer Institute Thesaurus [M]. New York: Springer, 2013.
- 9 The UniProt Consortium. UniProt: a hub for protein information [J]. Nucleic Acids Research, 2015, 43 (Database issue): D204.
- 10 Warnes GR, Bolker B, Bonebakker L, et al. Gplots: various R programming tools for plotting data [J]. R Package Version, 2009, 2 (4): 1.
- 11 Consortium TGO. Gene Ontology Consortium: going forward [J]. Nucleic Acids Research, 2015, 43 (Database issue): 1049 - 1056.
- 12 Yu G, Wang LG, Han Y, et al. ClusterProfiler: an R package for comparing biological themes among gene clusters [J]. Omics A Journal of Integrative Biology, 2012, 16 (5): 284.
- 13 Croft D, Mundo AF, Haw R, et al. The Reactome Pathway Knowledgebase [J]. Nucleic Acids Research, 2014, 42 (Database issue): 472 - 477.
- 14 Fabregat A, Korninger F, Viteri G, et al. Reactome Graph Database: efficient access to complex pathway data [J]. Plos Computational Biology, 2018, 14 (1): e1005968.
- 15 Yin H, Bin Li, Weimin Z, et al. Bioso! - A Search Engine & Annotation Framework for Biological Big Data [C]. Beijing: The 8th International Biocuration Conference Abstract Booklet, 2015.
- 16 Mariani TJ, Budhreja V, Mecham BH, et al. A Variable Fold - change Threshold Determines Significance for Expression Microarrays [J]. Faseb Journal, 2003, 17 (2): 321 - 323.

(上接第 61 页)

- 16 涂泳秋, 陈国华, 朴胜华. 中医药科研中用到的数据挖掘方法综述 [J]. 医学信息, 2011, 24 (7): 4045 - 4047.
- 17 赵静, 郭洪涛, 韩经丹, 等. 中医药治疗流行性感冒文本挖掘结果与诊疗方案的比较分析 [J]. 中医杂志, 2014, 55 (7): 612 - 616.
- 18 刘国山, 施丽婕. 溃疡性结肠炎在中医药领域中的文本挖掘 [J]. 中医药学报, 2017, 45 (4): 40 - 43.
- 19 周华英, 张启蕊. 信息技术在中医药领域应用研究概况 [J]. 中国中医药信息杂志, 2011, 18 (3): 110 - 112.
- 20 吴其国, 胡叶青, 查元, 等. 大数据在中医药领域中的应用现状 [J]. 广西中医药大学学报, 2016, 19 (1): 153 - 156.
- 21 陈静锋, 郭崇慧, 魏伟. “互联网 + 中医药”: 重构中医药全产业链发展模式 [J]. 中国软科学, 2016, (6): 26 - 38.
- 22 林晓航. 方证数据挖掘分析对腰腿痛辨证用药规律的探索 [D]. 广州: 广州中医药大学, 2013.
- 23 邓宏勇, 许吉, 张译, 等. 中医药数据挖掘研究现状分析 [J]. 中国中医药信息杂志, 2012, 19 (10): 21 - 23.
- 24 杜静. 数据挖掘技术在乳腺增生病中医辨证的应用研究 [D]. 济南: 山东中医药大学, 2009.