

· 临床论著 ·

基于 CiteSpace 软件中医数据挖掘文献的 可视化分析研究

林 骞 徐 浩

摘要 **目的** 通过对有关中医数据挖掘方面的文献进行计量、可视化分析,从而探究其研究历史、现状、热点以及发展趋势。**方法** 在 CNKI 检索自建库以来有关中医学数据挖掘方面的文献,应用 Excel 对纳入的文献进行初步统计,应用 CiteSpace 对文献中的作者、机构、关键词进行共现分析,绘制相关知识图谱,使用模块值和平均轮廓值评价视图的结构和清晰度,通过突现度和中介中心性发现不同聚类中的重要机构、作者以及关键词,对各主要关键词进行聚类分析并采用对数似然比方法对聚类进行标记,使用时间线视图展示相关聚类、聚类之间的相互影响以及聚类中重要关键词的历史跨度。**结果** 共获得文献 3 019 篇,发文量最大的三家机构是北京中医药大学、中国中医科学院、广州中医药大学,起“桥梁”作用的机构有 14 家;核心作者有 59 位,共发文 955 篇,占总发文量 33%,其中吴嘉瑞、任玉兰、张冰、周雪忠等学者在国内有较大影响力;在中医数据挖掘主题下共有关键词 264 个,其中每 5 年出现频率 ≥ 21 次有 45 个,起“桥梁”作用的关键词有 25 个,形成有意义的聚类 12 个。**结论** 中医证候分布、用药规律、选穴规律、传承名医经验思想依然是中医数据挖掘方面的热点;中医数据挖掘技术的热点主要集中在聚类分析、关联规则、logistic 回归,其他技术例如因子分析、复杂网络、决策树、主成分分析、频数分析、贝叶斯网络、神经网络等应用方面仍较薄弱。

关键词 CiteSpace; 数据挖掘; 中医学; 可视化分析

Visualization Analysis of Literature of TCM Data Mining Based on CiteSpace Software LIN Qian and XU Hao Cardiovascular Disease Center, Xiyuan Hospital, China Academy of Chinese Medical Sciences, Beijing(100091)

ABSTRACT **Objective** To explore the history, present situation, development trend and the research hotspots of the field of data mining of traditional Chinese medicine (TCM), relevant literature was visualized and analyzed. **Methods** In CNKI, literature on data mining of traditional Chinese medicine had been searched since the establishment of the database. Excel was used to make preliminary statistics of the included documents. CiteSpace was used to analyze the authors, organizations and keywords in the literature and draw the relevant knowledge map. Module values and average contour values were used to evaluate the structure and clarity of views. The important organizations, authors and key words in different clusters were found by using the degree of emergence and intermediary centrality. The main keywords were analyzed and labeled by log likelihood ratio. The time line view was used to show the correlation clustering, the interaction between clusters and the historical span of key words in clustering. **Results** A total of 3 019 papers were obtained. The three institutions with the largest number of articles published were Beijing University of Chinese Medicine, China Academy of Chinese Medical Sciences and Guangzhou University of Chinese Medicine. There were 14 institutions that play a role of “bridge”. There were 59 core authors, with 955 papers distributed, accounting for 33% of the total volume, among which Wu Jia-rui, Ren Yu-lan, Zhang Bing, Zhou Xue-zhong and other scholars had great influence in China. There were 264 key words under the theme of TCM data mining, and 45 of them are frequency greater

基金项目: 国家中医药管理局国家中医临床研究基地业务建设科研专项课题 (No. JDZX2015263); 首都卫生发展科研专项项目 (No. 2018-1-4171)

作者单位: 中国中医科学院西苑医院心血管科 (北京 100091)

通讯作者: 徐 浩, Tel: 010-62835341, E-mail: xuhaotcm@hotmail.com

DOI: 10.7661/j. cjim. 20191209. 240

than or equal to 21 times every five years. There were 25 key words to play the role of "bridge". Finally, 12 meaningful clusters were formed. Conclusions The distribution of TCM syndromes, the regularity of drug use, the law of selecting points, and the idea of inheriting famous doctors are still the focus of data mining in TCM. The focus of TCM data mining technology focuses on cluster analysis, association rules, logistic regression, but the application aspects of other technologies such as factor analysis, complex network, decision tree, principal component analysis, frequency analysis, Bayesian network, neural network and other applications are still weak. Using the new platform and new method to study TCM may be the trend of data mining of TCM in future.

KEYWORDS CiteSpace; data mining; traditional Chinese medicine; visualization analysis

临床是最重要、丰厚的研究资源,临床研究是转化医学研究的关键环节。中医学历经数千年的发展,积累了丰富的临床实践经验,以辨证论治、个体化治疗为其一大特色^[1]。目前公认的评价手段随机对照试验要求简单明确的干预措施、成功的对照措施(如安慰剂)、高度同质的研究人群等,不能充分体现中医药整体观念和辨证论治的特点。针对临床实际问题,探求新型研究模式,对于揭示临床疾病规律,改善临床实践,提升临床诊疗水平具有重要意义。随着互联网、移动互联网和物联网的发展,我们迎来大数据时代。大数据平台与数据挖掘技术为探索传承中医药的方法提供了新的支撑^[2,3]。CiteSpace 是美国德雷塞尔大学(Drexel University)陈超美博士及其团队开发的科学文献可视化分析软件。应用 CiteSpace 软件对某个领域已有的文献进行共被引分析或共词分析,可以直观地了解该领域的发展趋势与现状,在信息分析领域已成为影响力较大的软件^[4]。本文以中国知网(CNKI)数据库为数据来源,应用 CiteSpace V 软件的共现分析功能,对数据挖掘在中医方面应用的有关文献的作者、研究机构以及关键词进行可视化,分析了该研究领域的历史、现状、热点及发展趋势。

资料与方法

1 数据来源以及检索策略 本文数据全部来源于中国知网,检索策略如下:点击高级检索—期刊—专业检索,于检索框中输入:SU='数据挖掘' OR SU='关联规则' OR SU='apriori 算法' OR SU='逐步回归' OR SU='Logistic 回归' OR SU='K-邻近' OR SU='贝叶斯' OR SU='神经网络' OR SU='判别分析' OR SU='支持向量机' OR SU='随机森林' OR SU='决策树' OR SU='聚类' OR SU='K-means' OR SU='复杂网络',时间跨度起始不限至 2018 年 3 月,来源期刊为“全部期刊”,在“文献分类目录”中

检索“中医学”以及“中西医结合”相关文献。文献检索结果如下,共检索相关文献 3 019 篇,发现自 2005 年开始相关主题发文量陡然上升,发表年度趋势如下图 1 示。

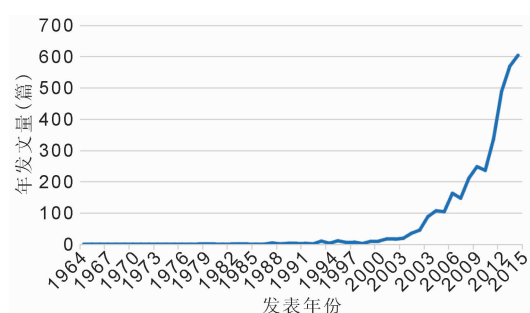


图 1 中医数据挖掘文献的发表年度趋势

2 软件参数设置 将上述文献以 Refworks 格式导出,导入 Citespace V 软件中,将上述文献进行机构、作者、关键词共现分析并对关键词进行聚类分析,软件参数设置如下:Time Slicing 模块下,设置时间跨度为 2005—2018 年,时间分区为 5 年;Text Processing 模块下,设置“Title”“Abstract”“Author Keywords”“Keywords Plus”作为词源(Term Source);节点类型(node type)选择分别为“Institution”“Author”“Keywords”;对象间的连接强度(link strength)选择“Cosine”,范围(scope)选择“With Slices”;数据删选标准(selection criteria)模块,设置“Top N”分别为 100,即筛选出每 5 年中出现频率最高的机构、作者以及关键词的前 100 名;Pruning 模块下,选择“MST”修建算法,“Pruning the merged networks”修剪策略^[5]。

3 数据整理 对于机构名称,将其统一到机构的一级,例如“北京中医药大学研究生院”改为“北京中医药大学”,“中国中医科学院广安门医院心内科”改为“中国中医科学院广安门医院”;统一同义词,例如“中国中医研究院”统一为“中国中医科学院”,“证候”统一为

“中医证候”,“聚类”统一为“聚类分析”等。

结 果

1 机构与作者共现分析 CiteSpace 依据网络结构和聚类的清晰度,提供了模块值(Q 值)和平均轮廓值(S 值)两个指标,它可以作为评判图谱绘制效果的一个依据。一般而言,Q 值一般在[0,1]区间内,Q 值>0.3 就意味着划分出来的社团结构是显著的,当 S 值在 0.7 以上时,聚类是高效率令人信服的,若在 0.5 以上,聚类一般认为是合理的^[6]。图 2 的 Q 值为 0.74,共囊括了 230 家机构。紫色外环的节点是图谱中中心性(centrality)较高(中心性≥0.1)的节点。中心性亦称为中介中心性(betweenness centrality),其取值在 0~1 之间,是衡量一个节点在网络中重要性的指标。中心性高的点往往位于连接两个不同聚类的路径上,在图谱中发挥了节点间相互联系的“中介”与“桥梁”作用^[7]。从图 2 与表 2 中可以看出各个大学以及其附属医院以各自为中心形成网络,而且各个网络中心在不同的时间段形成密切的合作网络。这些机构中,发文量≥10 篇的有 54 家,共发文章 2 017 篇,占总发文量 76%。其中从表 1 可以看出发文量最大的三家机构是北京中医药大学 200 篇,中国中医科学院 178 篇,广州中医药大学 124 篇,共占总发文量的 19%。

综合图 2 与表 1,可以看出北京中医药大学的吴嘉瑞、张冰以及以成都中医药大学的任玉兰、梁繁荣已经在各自机构该领域中形成优势。图 3 的 Q 值为 0.93,S 值为 0.63,从图 2 可以看出以北京交通大学的周雪忠、北京中医药大学的吴嘉瑞以及河北中医学院的贾春生为中心的三个较大的作者合作网。在最大的网络中形成了任玉兰、刘保延、张润顺较小的子网络。根据普莱斯定律,同一主题中核心作者的最低发文量 $MP = 0.749 \sqrt{Np_{max}}$ (Np_{max} 为同一主题中最高 的发文量,表 1 示其值为 40)^[8],即本主题中 $MP = 10$,故核心作者有 59 位,共发文 955 篇,占总发 文量 33%。图谱中聚类的衡量主要有两个指标,聚类 中节点数(Nodes)与轮廓值。某一聚类中的节点数 越多,说明该聚类越大,包含的信息也就越多。轮廓值 越大聚类的相似性也就越大。对作者合作网中最大的 网络进行聚类分析如图 4,能看到各个合作团体相对 固定,但各合作团队合作并不紧密。表 3 为作者合作 网聚类总结表,从表中总结了数据挖掘在中医学应用 中较大的聚类。大致可分为五类,#0、#3、#8 以冠心 病、血瘀证用药规律、证候分型为主要研究方向;#4 以 肺病用药规律、证候分型为主要研究方向;#5 以类风

湿关节炎用药规律、证型分布为主要研究方向;#6 以 舌诊、脉象、证候、疾病相关性为主要研究方向;#7 以 针灸选穴规律为主要研究方向;#12 以中医体质与疾 病相关性为主要研究方向。此外,周雪忠作为唯一中 心性≥0.1 的作者,在较大网络中起了重要桥梁作用。

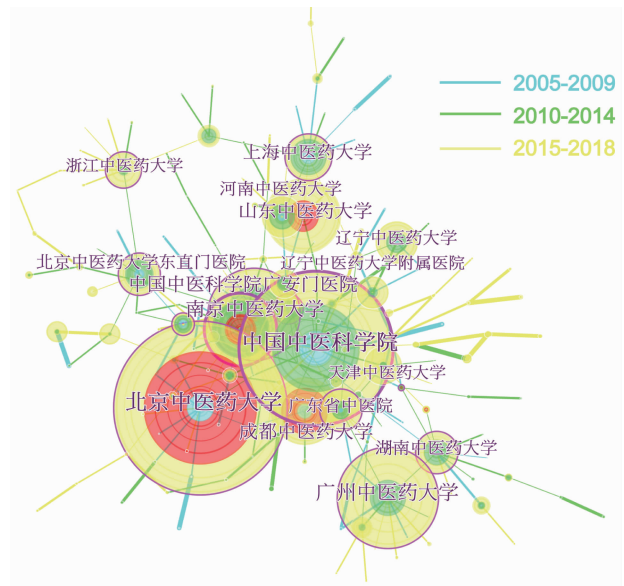


图 2 中医数据挖掘方面各家机构合作共现图

表 1 中医数据挖掘方面发文量最多的机构与作者*

机构	发文量	作者	发文量
北京中医药大学	200	吴嘉瑞	40
中国中医科学院	178	任玉兰	37
广州中医药大学	124	张冰	37
南京中医药大学	109	周雪忠	28
山东中医药大学	90	梁繁荣	26
中国中医科学院广安门医院	85	李建生	26
成都中医药大学	84	忻凌	25
上海中医药大学	61	张晓蒙	23
辽宁中医药大学	57	贾春生	23
湖南中医药大学	56	王建岭	22

注: * 表格中作者与机构非一一对应关系

表 2 起“桥梁”作用的机构

序号	中心性	机构
1	0.30	中国中医科学院
2	0.26	中国中医科学院广安门医院
3	0.19	南京中医药大学
4	0.19	上海中医药大学
5	0.19	湖南中医药大学
6	0.17	江西中医药大学
7	0.16	河南中医药大学第一附属医院
8	0.16	湖北省中医药研究院
9	0.15	北京中医药大学东直门医院
10	0.15	中国中医科学院西苑医院
11	0.14	北京中医药大学
12	0.12	广州中医药大学
13	0.12	广东省中医院
14	0.12	浙江中医药大学

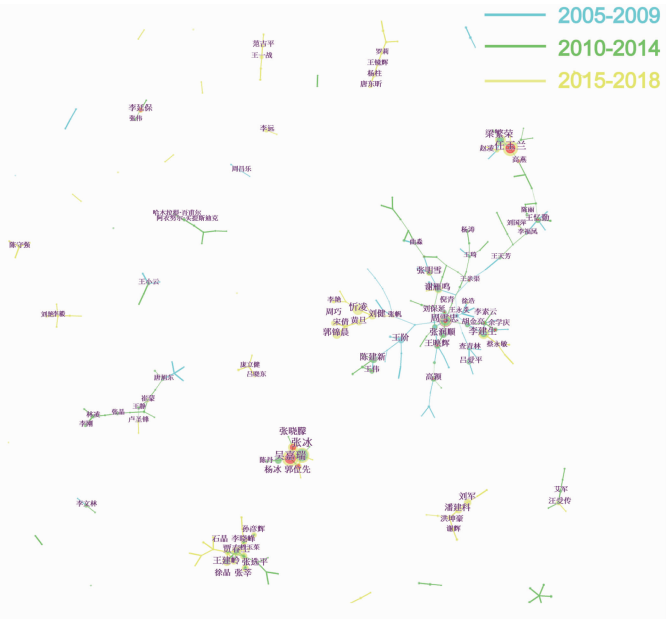


图3 中医数据挖掘方面各作者合作共现图

4.3 关键词共现分析(表4、5,图5、6) 2005年1月—2018年3月,该主题下共有关键词264个,其中出现频率≥21次有45个,如表4。除去检索时的主题词汇,还可以看到“因子分析”“主成分分析”“频数分析”“数据库”。数据挖掘也涉及到中医学各个领域,包括“用药规律”“名医经验”“经验传承”“辨证规律”以及各个疾病的证候变化规律、诊疗规律。图5为中医数据挖掘方面关键词共现图谱,Q值为0.695,



图4 中医数据挖掘方面作者聚类图

S值为0.628,共形成264个节点,368条线,起“桥梁”作用的节点有25个,具体如表5。

对各主要关键词进行聚类分析并采用对数似然比(LLR)方法对聚类进行标记,共形成有意义的聚类12个并生成时间线图,如图6。对这12个聚类主要研究方向进行归纳总结(如表6),#0、#1、#2、#3、#8、#9归纳为名家名医用药规律以及针灸选穴规律、疾病中医证候的分布规律;#4为数据挖掘在中医学方面的综述;#5、#6为肺癌、血瘀证等疾病诊断或量化诊断的研究;#7为糖尿病证候分布、中医药治疗用药规律;#10中医药治疗类风湿关节炎的研究;#11为利用支持向量机在中医辨证、舌诊、脉诊规律发现方面研究。

表3 中医数据挖掘方面作者合作网聚类总结表

聚类号	大小	S 值	平均年	聚类后标签	研究方向
#0	24	0.931	2007	急性心肌梗死	
#3	19	0.947	2009	冠心病发病	冠心病、血瘀证用药规律、证候分型
#8	13	0.976	2007	血瘀证	
#4	18	0.978	2008	动态 kohonen 网络	肺病用药规律、证候分型
#5	17	0.988	2012	关联规则	类风湿关节炎用药规律、证型分布

表4 中医数据挖掘方面主要关键词列表

频率	关键词	频率	关键词	频率	关键词	频率	关键词
869	数据挖掘	68	冠心病	26	失眠	22	药对
395	关联规则	66	中药	26	用药经验	22	肺癌
374	用药规律	62	因子分析	26	针灸疗法	22	数据库
360	聚类分析	52	文献研究	26	高血压	21	中风
332	中医证候	50	复杂网络	26	证素	21	中医辨证
141	中医药	38	综述	25	辨证	21	聚类
128	名医经验	36	证候要素	25	主成分分析	21	判别分析
98	中医传承辅助平台	35	糖尿病	25	频数分析	21	bp 神经网络
86	logistic 回归	30	辨证分型	24	选穴规律	21	证型
79	聚类算法	30	决策树	23	贝叶斯网络		
77	医案	30	方剂	22	人工神经网络		
71	针灸	27	配伍规律	22	规律		

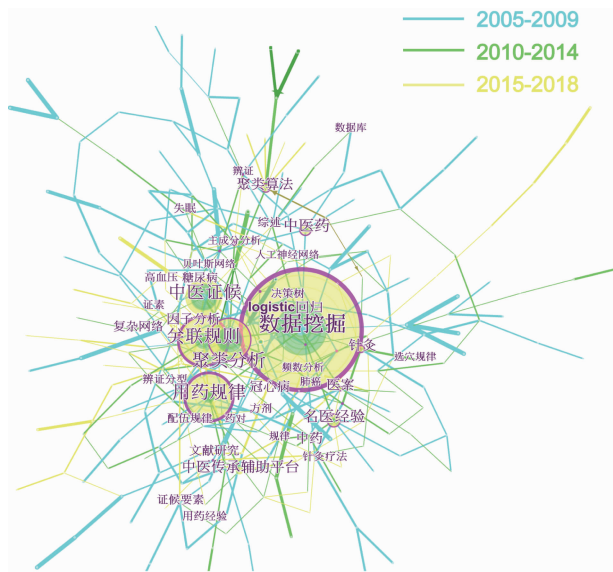


图 5 中医数据挖掘方面关键词共现图谱

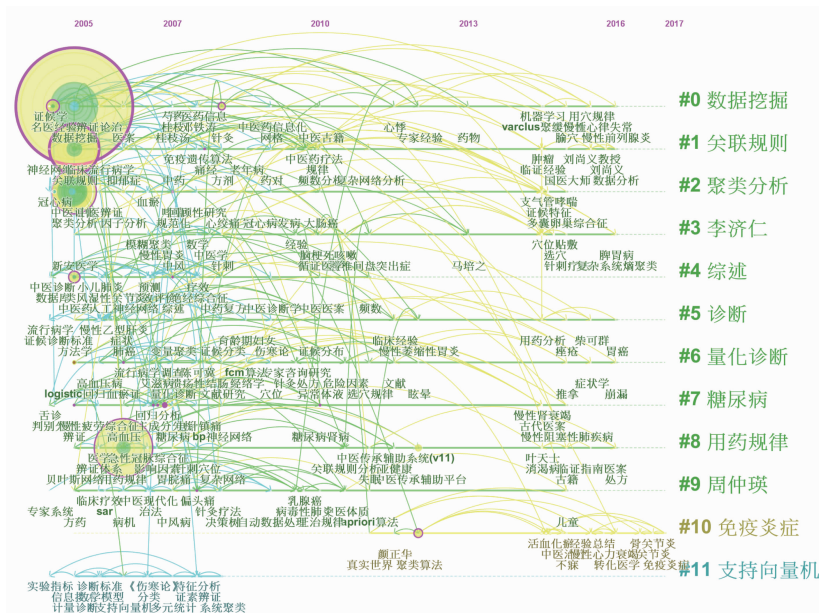


图 6 中医数据挖掘方面关键词聚类后时间线图

表 6 中医数据挖掘方面关键词聚类表

聚类号	大小	S 值	平均年	标签 (LLR)	主要研究方向
#0	27	0.891	2009	数据挖掘	
#1	27	0.795	2010	关联规则	
#2	27	0.822	2008	聚类分析	
#3	25	0.674	2011	李济仁	名家名医用药规律以及针灸选穴规律;疾病中医证候的分布规律
#8	22	0.858	2009	用药规律	
#9	22	0.735	2008	周仲瑛	
#4	24	0.773	2006	综述	中医数据挖掘方面的综述
#5	24	0.759	2008	诊断	
#6	23	0.726	2009	量化诊断	肺癌、血瘀证等疾病诊断或量化诊断的研究
#7	22	0.735	2008	糖尿病	糖尿病证候分布、中医药治疗用药规律
#10	19	0.942	2015	免疫炎症	中医药治疗类风湿关节炎的研究
#11	14	0.969	2006	支持向量机	利用支持向量机在中医辨证、舌诊、脉诊规律发现方面研究

表 5 起“桥梁”作用的关键词

中心度	关键词	中心度	关键词	中心度	关键词
0.41	数据挖掘	0.16	医案	0.12	针灸疗法
0.28	关联规则	0.16	综述	0.12	偏头痛
0.21	用药规律	0.15	针灸	0.11	聚类算法
0.20	主成分分析	0.14	方剂	0.11	文献研究
0.19	辨证分型	0.13	名医经验	0.11	哮喘
0.19	回归分析	0.13	证型	0.10	中医证候
0.18	聚类分析	0.13	血瘀证	0.10	辨证
0.17	中医药	0.13	糖尿病肾病		
0.17	抑郁症	0.13	四诊信息		

讨 论

应用 CiteSpace 对数据挖掘在中医领域应用方面进行回顾性分析,我们发现应用数据挖掘方法研究各个病种的中医证候分布、用药规律、选穴规律以及传承名医经验思想依然是中医数据挖掘方面的热点。从各个知识图谱中各个节点的连线的颜色来看,2005 年

至今,各个机构合作紧密,并且以各个大学或者大学附属机构为中心形成该领域的高产网络,尤其以北京中医药大学、中国中医科学院以及广州中医药大学最为瞩目。例如,以吴嘉瑞、张冰为核心的作者团队就颜正华、马培之等国家级名老中医治疗胃脘痛、眩晕等疾病组方用药规律做了大量数据挖掘工作,其发表的文献^[9]也对关联规则、贝叶斯网络、神经网络等数据挖掘方法在名老中医经验传承研究中的应用进行总结,并对这些常用方法的不足与优势进行分析,最终指出将数据挖掘进行集成并开发相应的软件是重要研究方向,该文献也在知网中被引 83 次,被下载 3 758 次。作者共现图谱中另一个令人瞩目的网络是以北京交通大学周雪忠为中心的巨大合作网,该合作网的作者主要来源于北京地区,表 3 也已经对其研究方向做了总结归纳,可见数据挖掘已经涉及中医学各个方面,并且方向稳定。周雪忠在 2007 年对中医临床数据库及挖掘分析平台的研究与应用做了分析论述,认为中医临床诊疗数据的结构化存储、整理、分析和应用是中医学继承、创新和发展、融合吸收现代化技术、创新中医学研究水平的重要途径^[10]。成都中医药大学的任玉兰采用关联规则挖掘技术,从古今针灸处方信息中,提取了不同维度和层次下的经穴选用规律,为针灸临床决策和科学研究提供参考依据^[11]。

值得关注的是中国中医科学院眼科医院的王一战与中国中医科学院的范吉平、广东省中医院的刘军与广州中医药大学第二临床医学院的潘建科、杨济源、洪坤豪、谢辉以及中国中医科学院的唐东昕与贵阳中医学院的杨兵、杨柱、王镜辉、黄雯琪自 2015 年起逐渐活跃,分别在运用数据挖掘知识挖掘中医脑病^[12]、中医骨伤^[13]、中医肿瘤^[14]用药规律方面逐渐崭露头角,可能在今后成为高产团队。

从图 5 可以看出,自 2005—2018 年,中医领域数据挖掘技术的热点主要集中在聚类分析、关联规则、logistic 回归,主要应用在用药规律、名医经验、中医证候、医案、针灸等方面,与上述机构作者共现分析结果一致。但在数据挖掘的其他技术例如因子分析、复杂网络、决策树、主成分分析、频数分析、贝叶斯网络、神经网络等应用方面仍较薄弱。从 2015—2018 年关键词的突现度^[7]来看中医传承辅助平台、用药规律仍较活跃。中医传承辅助平台由中国中医科学院中药研究所与中国科学院自动化研究所联合开发,于 2011 年推出,在总结名老中医经验总结、文献整理、新药研发、处方筛选方面取得了大量的进展^[15]。研究名老中医用药规律也是

继承和发展中医防治疾病的重要方法和手段^[16]。这也说明中医研究者渴望一个集临床科研于一体的平台,渴望与新的科技相结合、探索新的研究方法来传承创新中医,这也许也是未来中医药数据挖掘的趋势。

利益冲突:无。

参 考 文 献

- [1] 徐浩,史大卓,陈可冀.构建心血管病中西医结合诊疗随访大数据平台[OL].<http://www.365heart.com/show/108402.shtml>.
- [2] 徐浩,史大卓,刘保延,等.以临床实践数据为导向构建中西医结合临床指南的设想[J].中国中西医结合杂志,2009,29(6):544-547.
- [3] 李金根,姜众会,高铸焯,等.真实世界研究在中医药临床研究中的应用[J].世界科学技术-中医药现代化,2017,19(1):78-82.
- [4] Chen CM. (2006) CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature[J]. J Am Soc Inf Sci, 2006, 57(3): 359-377.
- [5] 陈悦,陈超美,胡志刚,等.引文空间分析原理与应用_CiteSpace 实用指南[M].北京:科学出版社,2014:27-69.
- [6] 陈悦,陈超美,刘则渊,等. CiteSpace 知识图谱的方法论功能[J].科学学研究,2015,33(2):242-253.
- [7] 李杰. CiteSpace 中文指南[EB/OL].<http://blog.sciencenet.cn/blog-554179-1066981.html>, 2017-07-19.
- [8] 邱均平主编.信息计量学[M].武汉:武汉大学出版社,2007:192-194.
- [9] 吴嘉瑞,唐仕欢,郭位先,等.基于数据挖掘的名老中医经验传承研究述评[J].中国中药杂志,2014,39(4):614-617.
- [10] 周雪忠,刘保延,姚乃礼,等.中医临床数据库及挖掘分析平台的研究与应用探讨[J].世界科学技术-中医药现代化,2007,9(4):74-80.
- [11] 任玉兰,赵凌,陈勤,等.数据挖掘技术在经穴选用及其特异性研究中的应用[J].中医杂志,2010,51(1):47-51.
- [12] 王一战,王玉贤,苏芮,等.基于数据挖掘的流行性乙型脑炎恢复期中医用药规律研究[J].中国中医急症,2016,25(10):1859-1862.
- [13] 潘碧琦,潘建科,刘军,等.基于关联规则和复杂系统熵聚类的痛风用药规律研究[J].中华中医药杂志,2014,29(6):2040-2043.
- [14] 杨柱,唐东昕,郭斌,等.刘尚义治疗肿瘤用药经验数据挖掘分析[J].中医杂志,2016,57(19):1641-1645.
- [15] 唐仕欢,申丹,卢朋,等.中医传承辅助平台应用述评[J].中华中医药杂志,2015,30(2):329-331.
- [16] 唐仕欢,杨洪军.中医组方用药规律研究进展述评[J].中国实验方剂学杂志,2013,19(5):359-363.

(收稿:2018-05-14 在线:2020-01-03)

责任编辑:白霞