Advances in Applied Mathematics 应用数学进展, 2025, 14(2), 362-375
Published Online February 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.142077

Hans汉斯

基于文本挖掘的中医药学主题提取及演化

张　恒，邹晨晨*

青岛大学数学与统计学院，山东　青岛

收稿日期：2025年1月26日；录用日期：2025年2月19日；发布日期：2025年2月26日

摘　要

文章挖掘中医药学领域过去30年的研究主题，总结中医药研究主题的主流、变迁及演化，爬取中医药学领域硕博论文及权威期刊，划分时间段分析研究方向与方法，运用词云图、词频统计、LDA主题模型分析研究主题热点。查找中医药学领域的硕博论文及期刊，最终整合得到14个主要研究主题。硕博论文主要研究信号通路，中药和疾病都有涉及；《中国中药杂志》以中药研究和统计分析为主；《中医杂志》更关注具体疾病的诊治。LDA主题模型能有效挖掘中医药学文献的研究主题，80%都能被相应领域的综述类文献所验证。

关键词

文本挖掘，爬虫，LDA主题模型，中医药学

Extracting and Evolving Traditional Chinese Medicine Themes Based on Text Mining

Heng Zhang, Chenchen Zou*

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Jan. 26th, 2025; accepted: Feb. 19th, 2025; published: Feb. 26th, 2025

Abstract

This article explores the research topics in the field of traditional Chinese medicine over the past 30 years, summarizes the mainstream, changes, and evolution of traditional Chinese medicine research topics, crawls master's and doctoral theses and authoritative journals in the field of traditional Chinese medicine, divides time periods to analyze research directions and methods, and uses word cloud maps, word frequency statistics, and LDA topic models to analyze research topic hotspots. Analyzing master's and doctoral theses and journals in the field of traditional Chinese medicine, 14

*通讯作者。

文章引用: 张恒, 邹晨晨. 基于文本挖掘的中医药学主题提取及演化[J]. 应用数学进展, 2025, 14(2): 362-375.
DOI: 10.12677/aam.2025.142077

main research topics were ultimately integrated. The master's and doctoral theses mainly focus on signal pathways, including traditional Chinese medicine and diseases; *China Journal of Chinese Materia Medica* focuses on research and statistical analysis of traditional Chinese medicine; and *Journal of Traditional Chinese Medicine* focus more on the diagnosis and treatment of specific diseases. The LDA topic model can effectively explore research topics in traditional Chinese medicine literature, and 80% of them can be validated by relevant literature reviews in the field.

Keywords

Text Mining, Crawler Technology, Latent Dirichlet Allocation Topic Model, Traditional Chinese Medicine

1. 引言

文本挖掘是对意义丰富的文本进行分析，理解其内容和意义的过程。深入研究可以提高人们从大量文本数据中提取信息的能力。随着计算机技术的快速发展，文本挖掘技术取得了巨大的发展，逐渐成为了一种主流方法[1]-[3]。文本聚类是根据文本的相似性度量，然后采用聚类法将文本聚类。采用的方法通常使用 K-Means 聚类、LDA 等方法[4] [5]。文本结构分析可以有效地改进文本摘要、文本检索以及文本过滤的精度，将文本依据主题划分为若干层次[6] [7]。

在国外文本挖掘技术研究中，Masanori Hirano 等[8]提出了基于使用 Word2vec 并根据其与主题的相似性提取单词的方法。Wermter 等[9]提出了在生物医学领域，许多用于依赖基本的形态和句法分析方法的文本挖掘和信息提取系统。Nemrava 和 Svátek [10]构建一种文本挖掘工具，用于收集特定单词，使用网络目录从 UNSPSC 分类法中获取有关产品适当信息的方式，并提出了如何进一步处理提取信息的方法。隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)由 Blei 等在 2003 年提出，也被称为三层贝叶斯概率模型，对应文本层、主题层、词项层。

LDA 模型被用于经济领域[11]、旅游产业[12]、计算机科学领域[13]、化学领域[14]等。在国内研究 LDA 主题模型与应用领域中，王伟等[15]分析在线评论，运用 LDA 主题模型建模，分析在线评论主题动态演化。胡泽文等[16]以机器学习论文为研究对象，融合 LDA 和 Word2vec 方法进行主题建模和主题演化分析，分阶段展示演化规律。王一博等[17]以国内用户画像论文作为研究领域，运用 LDA 主题模型分析题目、摘要和关键词，并探究热点主题和演化趋势，得出 8 个研究热点。庞庆华等[18]运用 LDA 挖掘用户的历史微博主题，分析用户的兴趣主题，为用户推荐感兴趣的内容。

中医药是最具代表性的中国元素，中医药的发展和传播也是中国传统文化复兴的重要体现。中医药学在恶性肿瘤治疗方面也有着悠久的历史，是国内肿瘤治疗的关键组成部分。将文本挖掘技术应用到中医药学领域中对分析中医药学的研究主题及趋势有着重要意义。由于在中医药学领域的主题探索较少，本文借助 LDA 主题模型对中医药学领域硕博论文及高水平期刊进行主题提取，分析主题内容，更深入地了解中医药学领域的研究热点。

2. 数据获取与预处理

获取在中国知网(CNKI)上中医药学有关的硕博论文和高水平期刊的文献。通过网络爬虫获取数据，

文献数据主要由两部分组成。

因 2000 年之前论文数据存在空白，选取 2000~2022 年 CNKI 上的中医药学领域的学位论文。数据包含中文题目、作者、学位授予单位、学位、学位授予年度、摘要和关键词，共 25,017 篇文献。选取 1993~2022 年中医学中影响因子最大的《中医杂志》，中药学中影响因子最大的《中国中药杂志》并且被 SCI 和北大核心收录的文献。期刊文献数据包含题目、作者、发表时间、摘要和关键词等，共 36,630 篇文献。本文对数据进行预处理，包括期刊及论文的筛选、术语统一化处理、自定义分词词典。

1) 删除不相关的会议和信息缺失的文献。期刊文献中包含会议通知、新闻等与中医药学研究无关的文章，需要爬取数据后删除；由于摘要和关键词对文献分析至关重要，因此对缺失摘要和关键词的文献给予删除。

2) 对标准术语进行统一化。

3) 文本分词。在词汇表中添加新词，创建一个自定义的分词词典以获得更好的分词结果。

4) 去除停用词，参考停用词表和自添加停用词构建停用词词典。

3. 实验结果与分析

3.1. 描述性分析

图 1 展示了 2000~2022 年论文及 1993~2022 年期刊文献的数量变化趋势。2000~2016 年论文整体呈现上升趋势，随后出现下降趋势。期刊文献 1993~2002 比较平稳，2003~2012 呈现上升趋势，2013~2022 出现下降趋势。将文献拆分为三份，考虑到查询不到 20 世纪之前的论文以及 2005 年之前论文数量较少，将硕博论文分为 2000~2012 年、2013~2017 年和 2018~2022 年三部分。期刊文献分为 1993~2002 年、2003~2012 年和 2013~2022 年三部分。图 2~4 为论文和期刊的关键词词云图。

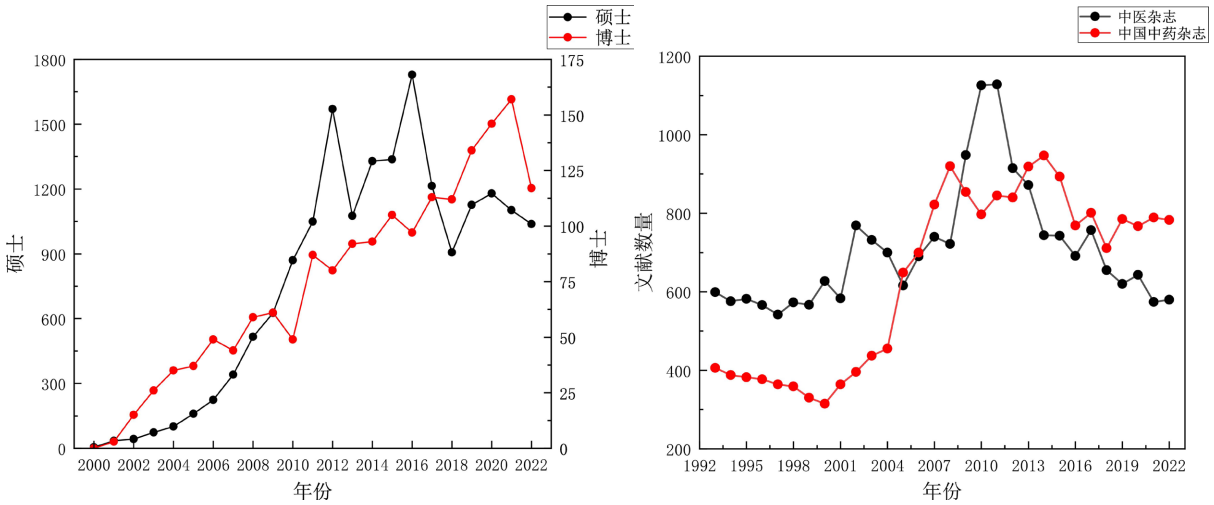

Figure 1. Trend chart of changes in papers and journal literature
图 1. 论文和期刊文献变化趋势图

由图 2 可以看出硕博论文研究方向主要有代谢组学、药代动力学、网络药理学等。研究方法主要有指纹图谱、MS、HPLC、UPLC、数据挖掘、针刺等。研究内容主要涉及到中药化学成分、含量测定、制备工艺、抗肿瘤、抗氧化、信号通路等。

Figure 2. Keyword cloud diagram for master's and doctoral theses
图 2. 硕博论文关键词词云图

Figure 3. Keyword cloud diagram of *China Journal of Chinese Materia Medica*
图 3. 《中国中药杂志》关键词词云图

由图 3 可以看出《中国中药杂志》研究方向主要为中药的药理作用、网络药理学、代谢组学等。研究方法主要为 UPLC、meta 分析、HPLC、MS、指纹图谱方法等。研究内容主要涉及到药物作用机制、中医药及中成药含量测定，其中出现较多的中药材有黄酮、黄芪、甘草。

Figure 4. Keyword cloud diagram of *Journal of Traditional Chinese Medicine*
图 4. 《中医杂志》关键词词云图

由图 4 可以看出《中医杂志》的研究方向主要为中医药疗法、中医证候、中西医结合疗法等。研究方法主要为随机对照试验、针刺法、辨证论治等。涉及到的疾病主要有冠心病、糖尿病、肿瘤、新型冠状病毒肺炎等。

统计词语词频可以反映出在一篇文章中出现的频率。对论文和期刊关键词进行词频统计。

Table 1. Keyword frequency in master's and doctoral theses
表 1. 硕博论文关键词词频

2000~2012 年		2013~2017 年		2018~2022 年	
关键词	词频	关键词	词频	关键词	词频
化学成分	559	质量标准	314	化学成分	269
质量标准	552	化学成分	310	质量标准	260
指纹图谱	316	指纹图谱	226	MS	211
临床研究	313	临床研究	218	代谢组学	206
制备工艺	255	含量测定	187	网络药理学	198
含量测定	183	MS	179	指纹图谱	179
临床观察	172	制备工艺	170	含量测定	150
高效液相色谱法	263	HPLC	130	肠道菌群	145
HPLC	147	针刺	120	制备工艺	119
针刺	147	药代动力学	116	作用机制	119
结构鉴定	123	抗氧化	114	质量评价	107
药效学	114	抗肿瘤	107	氧化应激	104
提取工艺	99	电针	102	药代动力学	88
镇痛	96	中医证型	87	质量控制	83
黄酮	92	质量控制	86	UPLC	79
提取	82	代谢组学	76	NF	79
人参	80	临床疗效	76	抗肿瘤	77
中医证型	80	提取工艺	73	PI3K	73
质量控制	79	细胞凋亡	70	抗氧化	70
合成	73	数据挖掘	69	抗炎	68

由表 1 可以看出 2000~2022 年硕博论文的共同研究方向主要有药物质量标准、化学成分测定、临床研究、制备提取工艺等。研究方法主要有电针、针刺、HPLC 检测法、指纹图谱法等。2013~2017 年新增了 MS 信号通路和抗肿瘤的研究，2018~2022 年增加了 UPLC、肠道菌群、NF 信号通路和 PI3K 的研究。

Table 2. Keyword frequency in *China Journal of Chinese Materia Medica*
表 2. 《中国中药杂志》关键词词频

1993~2002 年		2003~2012 年		2013~2022 年	
关键词	词频	关键词	词频	关键词	词频
高效液相色谱法	75	化学成分	526	化学成分	467
化学成分	72	HPLC	321	MS	403

续表

含量测定	68	中药	165	网络药理学	218
HPLC	60	高效液相色谱	156	UPLC	188
炮制	49	含量测定	142	含量测定	171
挥发油	45	黄酮	119	HPLC	161
薄层扫描法	28	指纹图谱	117	指纹图谱	157
高效液相色谱	27	MS	101	Meta 分析	142
薄层扫描	21	色谱	89	作用机制	135
中药	20	大鼠	80	分子对接	117
气相色谱	20	挥发油	80	质量控制	115
黄芪	19	细胞凋亡	69	系统评价	114
黄连	19	丹参	64	黄酮	106
药用植物	19	凋亡	64	代谢组学	98
齐墩果酸	18	半夏	49	药理作用	91
黄酮	18	药用植物	49	丹参	86
多糖	18	甘草	48	研究进展	81
本草考证	16	质量控制	48	主成分分析	79
镇痛	16	生物碱	48	药代动力学	78
生物碱	16	三萜	46	随机对照试验	78

由表 2 可以看出 1993~2022 年《中国中药杂志》的共同研究方向主要有中药化学成分检测、含量测定、HPLC、中药作用机制等。共同研究方法主要有高效液相色谱、指纹图谱等。2003~2012 年研究方向新增质量控制、药用植物等，研究方法有指纹图谱法、MS 信号通路，涉及到的中药包括黄酮、丹参、甘草等。2013~2022 年增加了 Meta 分析法、UPLC、代谢组学、主成分分析法、药代动力学和随机对照试验的研究。

Table 3. Keyword frequency in *Journal of Traditional Chinese Medicine*
表 3. 《中医杂志》关键词词频

1993~2002 年		2003~2012 年		2013~2022 年	
关键词	词频	关键词	词频	关键词	词频
中医药疗法	1416	中医药疗法	903	名医经验	486
中药疗法	189	名医经验	221	冠心病	103
治疗应用	149	中西医结合疗法	103	新型冠状病毒肺炎	101
针灸疗法	94	诊断	77	中医药疗法	89
名医经验	93	参芪五味子片	70	针刺	73
药物作用	91	治疗应用	69	随机对照试验	70
诊断	82	证候	66	肿瘤	69
并发症	77	针灸疗法	64	中医证候	69

续表

中西医结合疗法	77	药物作用	60	黄帝内经	66
肝炎	62	辨证分型	57	慢性阻塞性肺疾病	64
代谢	52	糖尿病	57	辨证论治	62
病理学	50	辨证论治	55	恶性肿瘤	58
血液	47	针刺	53	糖尿病	56
中医病机	47	中医证候	49	生活质量	54
糖尿病	46	伤寒论	44	伤寒论	54
药理学	44	并发症	44	2 型糖尿病	54
病因学	42	经验	43	抑郁症	51
辨证论治	42	冠心病	42	文献研究	49
慢性	39	中医病机	40	炎症因子	49
胃炎	39	高血压	40	真实世界研究	48

由表 3 可以看出 1993~2022 年中医杂志的共同研究方向有分析老中医教授们的名医经验、中医药疗法、中西医结合疗法等。涉及到的疾病包括肝炎、胃炎、冠心病、糖尿病、高血压、肿瘤、慢性阻塞性肺疾病、新型冠状病毒肺炎等。共同研究方法有针刺、辨证论治等。1993~2002 年的研究还包括病理学、病因学；2003~2012 年新增了高血压、冠心病、伤寒论方向的研究；2013~2022 年新增加了新型冠状病毒肺炎、慢性阻塞性肺疾病、抑郁症、恶性肿瘤等研究方向，研究方法新增随机对照试验分析等。

3.2. LDA 主题模型

本文利用 Python 中的 Gensim 库训练 LDA 主题模型，主要涉及到三个参数 α、β、K，参数 α、β 分别是文档－主题分布的先验参数和主题－词分布的先验参数，在训练 LDA 模型时将这两个参数设置为 auto，Gensim 通过迭代自动选择最优参数值。最优主题数 K 通过计算困惑度确定，不同主题数对应的困惑度如图 5~7 所示。

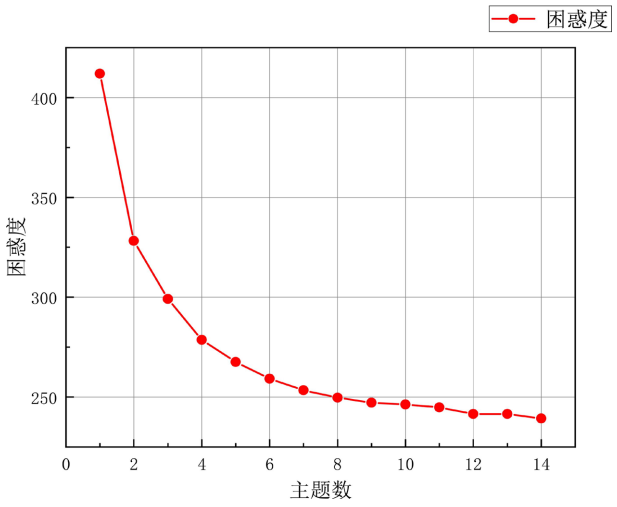

Figure 5. Curve chart of perplexity for master's and doctoral theses
图 5. 硕博论文困惑度曲线图

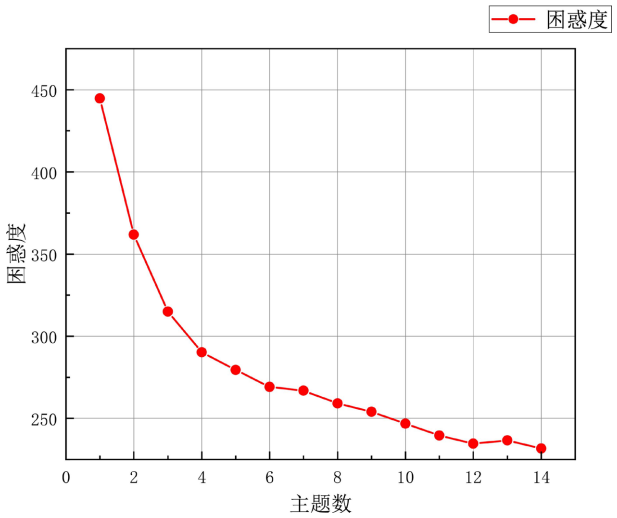

Figure 6. Curve chart of perplexity for *China Journal of Chinese Materia Medica*
图 6. 《中国中药杂志》困惑度曲线图

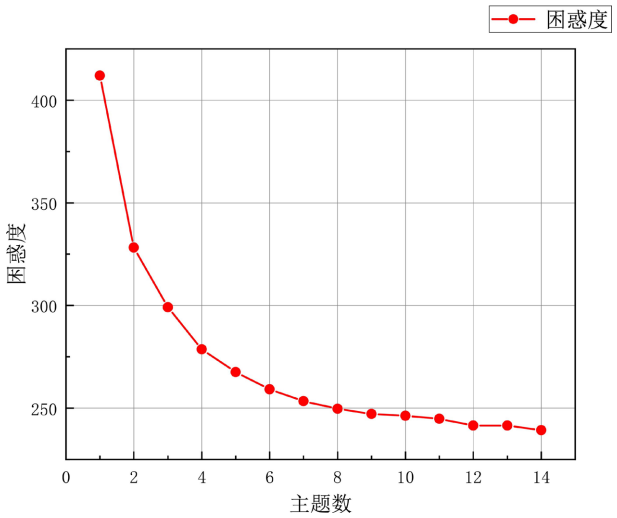

Figure 7. Curve chart of perplexity for *Journal of Traditional Chinese Medicine*
图 7. 《中医杂志》困惑度曲线图

随着主题数的增大，困惑度逐渐减小。根据手肘法和测试结果，硕博论文、《中国中药杂志》、《中医杂志》的主题数分别选取 8、10、10 时最佳。

分别对硕博论文、《中国中药杂志》和《中医杂志》建立 LDA 主题模型。(见图 8、图 9 和图 10)

表 4 展示了中医药学硕博论文的 8 个主题以及最重要的相关词汇。对中医药硕博论文的主题可以命名为：1) 细胞活性与基因分析：主要涉及到细胞实验中的检测、蛋白表达，诱导基因变化方式、分析肿瘤细胞等内容，着重分析药物对细胞活性和生长能力的影响。2) 药物制剂工艺与条件影响：主要涉及到药物的制剂条件、实验影响、中药工艺稳定性优化等。3) 化合物成分与结构研究：主要涉及到植物体内化合物提取、分析活性成分、如羟基，甲基，甲氧基等结构。4) 大小鼠模型：主要涉及到对大鼠小鼠肝脏造模，检测血清含量，设置对照实验，进行显著性检验。5) 中医药临床研究与发展：主要涉及到

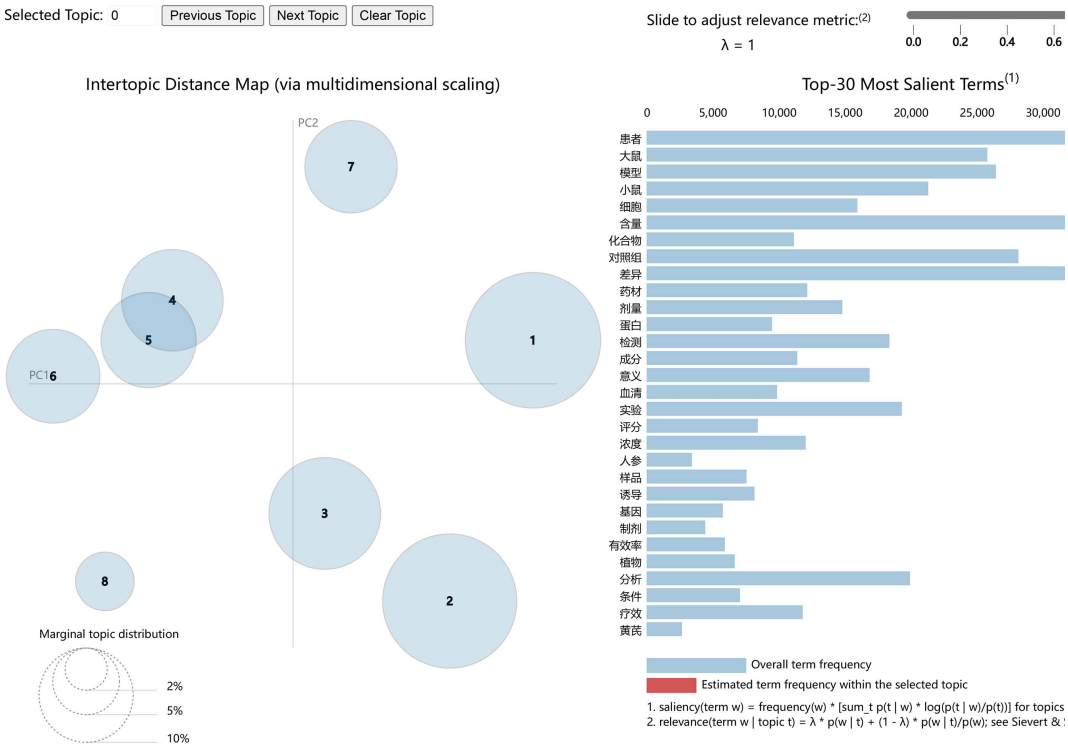

Figure 8. Visualization of LDA topic model for master's and doctoral theses
图 8. 硕博论文 LDA 主题模型可视化

Table 4. Distribution of subject words in master's and doctoral theses
表 4. 硕博论文主题词分布

Topic	与主题相关最高的词汇
1	细胞 检测 蛋白 诱导 基因 实验 药物 活性 肿瘤 通路 调控 分子 靶点
2	制剂 条件 指标 实验 工艺 乙醇 药材 质量标准 影响 处方 颗粒 复方 中药
3	化合物 植物 提取物 甘草 实验 羟基 葡萄糖 结构 活性 中药 甲基 甘草酸
4	大小鼠 模型 剂量 血清 对照组 检测 实验 空白 灌胃 显著性 造模 肝脏 阳性
5	临床 中药 文献 发展 理论 数据库 教授 基础 系统 附子 茯苓 白术 方剂
6	含量 成分 质量 饮片 植物 品种 产地 吴茱萸 炮制 挥发油 中药材 种子 栽培
7	人参 黄芪 滴丸 川芎 实验 色谱 越橘 花色素 栀子 检测 柱温 天麻 苦参
8	患者 差异 对照组 疗效 症状 有效率 临床 显著性 疗程 针刺 检验 试验

中医药在临床上的分析、中药理论发展、数据库文献分析等。6) 中药材种植及含量测定：主要涉及到中药材种植，成分差异等，研究药材植物品种的差异和培育技术对药效的影响。7) 中草药分析研究：主要涉及到中草药实验，包括人参、黄芪、川芎等，使用色谱分析法检测中草药成分等。8) 患者临床治疗差异评价：主要涉及到患者疗效、症状改善、对照组比较等，总结临床实践经验。

表 5 展示了《中国中药杂志》的 10 个主题以及最重要的相关词汇。对《中国中药杂志》的主题可以命名为：1) 小鼠实验模型：主要涉及到对小鼠造模，设置对照试验，分析肾脏，肿瘤相关疾病。2) 化合物分析和结构鉴定：主要涉及到各种使用色谱和光谱的方法对化合物的结构特征分析，主要包括羟基、

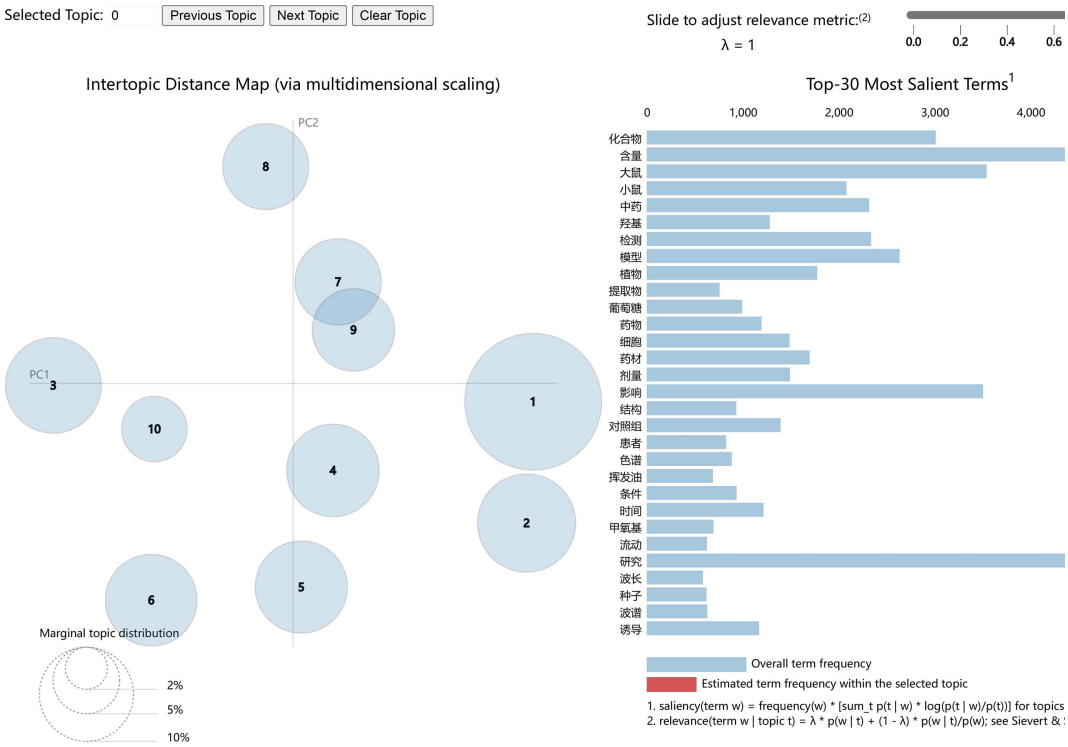

Figure 9. Visualization of LDA topic model for *China Journal of Chinese Materia Medica*
图 9. 《中国中药杂志》LDA 主题模型可视化

Table 5. Distribution of subject words in *China Journal of Chinese Materia Medica*
表 5. 《中国中药杂志》主题词分布

Topic	与主题相关最高的词汇
1	小鼠 对照组 剂量 血清 影响 胶囊 模型 肿瘤 显著性 肾脏 红细胞 灌胃 实验
2	化合物 羟基 植物 甲氧基 波谱 吡喃 黄酮 谷甾醇 甲基 硅胶 光谱 凝胶 色谱
3	中药药材 分析 药用植物 技术 中药材 文献 资源 研究所 综述 系统 传统 研究进展
4	含量 挥发油 甘草 多糖 差异 药材 质量 甘草酸 白芍 法测定 栽培 吴茱萸 指标
5	中药药物 复方 制剂 临床评价 处方 成分 中成药 组分 粒径 效应 疗效 实验 脂质体
6	基因 半夏 人参 培养基 诱导 序列 土壤 蛋白质 地黄 分化 蛋白 皂甙 遗传
7	提取物 黄芪 药材 丹参酮 检验所 口服液 水蛭 本草 党参 成分 四物汤 花粉
8	检测 色谱 药材 流动 波长 甲醇 乙腈 梯度 柱温 线性 磷酸 面积 质谱 苦参碱
9	条件 时间 温度 指标 工艺 乙醇 浓度 炮制 用量 饮片 实验 附子 提取液
10	大鼠模型 细胞 含量 检测活性 诱导 剂量 对照组 血清 蛋白 灌胃 阳性 脑组织 实验

甲氧基、甲基等。3) 中药研究与发展：主要涉及到中药药材研究，通过文献分析采用的技术与研究进展。4) 药材质量分析：主要涉及到中药药材的质量、含量测定、分析其中的指标等内容。5) 中药药物评价：主要涉及到中药药物、中成药等在临床上的使用评价，通过实验分析其疗效。6) 植物基因工程：主要涉及到诱导基因改变序列、改变蛋白质结构、分化等，研究植物的生物学特征及遗传基因的影响。7) 中草药提取与制剂：主要涉及到提取中草药有效成分，熬制提纯后对疾病有预防作用。主要的中药药材有黄

芪、丹参、本草等。8) 色谱分析：主要涉及到色谱分析药材成分实验，分析波长、柱温及面积等。9) 中医药药材炮制工艺：主要涉及到中医药炮制工艺的条件，包括时间、温度、中药材的浓度、质量等。10) 大鼠实验模型：主要涉及到使用大鼠进行实验，诱导基因蛋白变化进行分析活性，常分析的部位为脑组织。

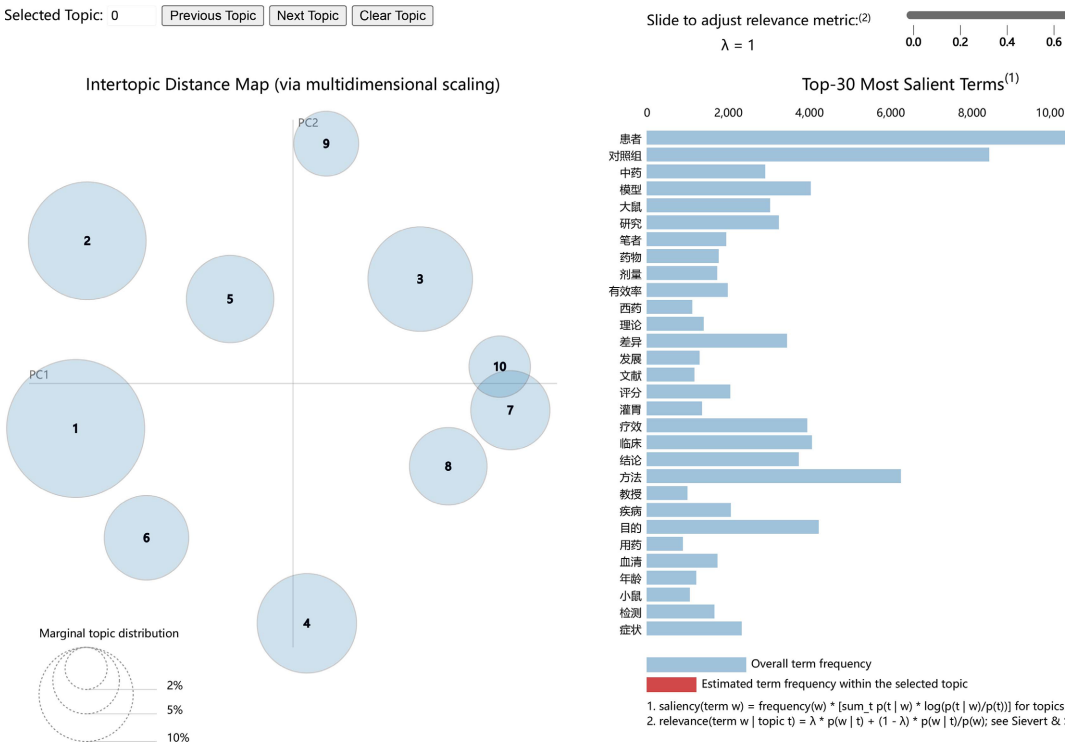

Figure 10. Visualization of LDA topic model for *Journal of Traditional Chinese Medicine*
图 10. 《中医杂志》LDA 主题模型可视化

Table 6. Distribution of subject words in *Journal of Traditional Chinese Medicine*
表 6. 《中医杂志》主题词分布

Topic	与主题相关最高的词汇
1	研究方法 文献检索 冠心病 数据分析 数据库 评价 标准 症状 指标 临床 方案 针刺
2	理论发展 临床 疾病 人体影响 分析 过程 功能 针灸 黄帝内经 内涵 系统 传统 中医
3	模型 大鼠小鼠 剂量 灌胃 检测 血清 对照组 空白 造模 蛋白 浓度 手术 细胞
4	患者 疗效 临床表现 年龄 报告 病例 性别 皮肤 症状 头痛 医院 资料 门诊
5	中药西药 基因 细胞 复方 黄芪 调节 附子 靶点 调控 神经 疗效 儿童 受体 制剂 通路
6	患者 糖尿病 化疗 症状 高血压病 年龄 血压 肺癌 高血压 舌苔 性别 比例 艾滋病
7	药物 临床 方剂 活血益气 疗效 健脾 黄芪 甘草 温病 通络 白术 伤寒论 茯苓
8	肿瘤 教授 临床经验 激素 胃癌 肝癌 黄疸 肝硬化 高脂血症 妇科 肝脏 肝病
9	疾病症状 气血 病因 病理 血瘀 瘀血 脾胃 气虚 肾虚 脏腑 脾虚 气滞 论治
10	对照组 患者 差异 有效率 疗效 疗程 指标 症状 显著性 针刺 血清 评价

表 6 展示了中医杂志的 10 个主题以及最重要的相关词汇。对中医杂志的主题可以命名为：1) 冠心病研究与评价：主要涉及到文献检索、研究方法、病状特征等对冠心病进行深入研究和评价。2) 中医基础理论与发展：主要涉及到中医理论发展过程，以《黄帝内经》为代表的中医论文体系，探讨中医药的理论体系和治疗方法对人体的影响。3) 动物实验模型：主要涉及到对大鼠小鼠造模、构造方法、灌胃技术等；设置对照试验观察疾病的发生机制和细胞变化等。4) 临床病例分析：主要涉及到患者临床表现、疗效评估，通过病例分析探讨不同病人在治疗过程中的表现与反应。5) 中西药物疗效研究：主要涉及到传统中医药学基本概念和基本理论，如《黄帝内经》以及传统中医在临床上的应用。6) 疾病分析与治疗分析：主要涉及到通过对患者情况、症状血压等数据分析，从不同疾病角度分析治疗效果；疾病主要涉及到糖尿病、高血压病、艾滋病等。7) 中医药应用研究：主要涉及到中医药的临床应用、方剂搭配等；主要涉及到的中药有黄芪、甘草、白术、茯苓等。8) 肿瘤临床实践：主要涉及到肿瘤临床的经验总结，包括诊断治疗方案、激素使用等；疾病涉及胃癌、肝癌等。9) 中医论治与疾病症状分析：主要涉及到气血理论、病因分析等；症状有血瘀、气虚、气滞等。10) 针刺研究和对照试验设计：主要涉及到针刺法的应用、对照组设计、数据分析和评估治疗方案的有效性和显著性差异。

3.3. 研究方向总结

由词云图、关键词词频分析和 LDA 主题模型的结果可以看出，三种文献有许多共同主题，将主题合并分析得出，中医药学领域的 14 个重要主题，具体主题见表 7。

Table 7. Main research topics and contents of traditional Chinese medicine
表 7. 中医药学主要研究主题和内容

主题编号	研究主题	研究方法	研究内容
1	动物实验研究	大小鼠造模 化学因素刺激法	肾脏、肝脏、心脏细胞、神经元等
2	药物质量标准分析及制备	HPLC UPLC 指纹图谱法	药材成分、含量、质量差异； 色谱分析法分析药物成分
3	基因分析与应用	细胞模型 中药靶点 肿瘤基因诱导	基因测序，分析蛋白结构； 基因调控肿瘤细胞，多靶点抑制杀伤肿瘤细胞
4	中医药的临床研究	针刺法 对照试验	疾病诊断、药效和药理研究； 黄芪、甘草、茯苓、白术、白芍、附子等中药在临床上的应用
5	文献研究	数据库分析	文献和数据库的信息检索，分析研究方法和目的
6	活性成分分析	分子生物学 化学分析法	荧光染色细胞检测，分析肿瘤细胞 植物化合物成分活性分析
7	中药种植与环境影响	对照试验	中草药植物的栽培条件探索
8	患者特征及医疗研究	病例对照分析 随访研究	药物对患者的疗效，症状改善
9	中医辨证及理论应用	辨证分析 整体思维	教授传授经验理论； 《黄帝内经》基础理论； 主要应用于肿瘤、糖尿病、脾虚、冠心病、肺癌、肾虚、高血压等

续表

10	信号通路	PI3K 通路 MS 通路 Wnt 通路	抗炎症； 抗肿瘤； 免疫功能
11	药代动力学	分子对接 放射性同位素追踪	药物在体内的分布，代谢
12	网络药理学	网络建模 药效实验 数据库分析	药物副作用； 活性成分作用靶点预测
13	统计学方法应用	主成分分析 因子分析 Meta 分析 随机对照实验	遗传基因； 临床试验； 药效评估
14	新型冠状病毒肺炎	临床分析	连花清瘟防治新冠肺炎； 患者临床疗效

4. 结论

本文以中医药学论文、期刊文献为源数据集，通过词云图，关键词词频分段分析中医药学的研究方向与方法。运用 LDA 主题模型进行可视化及聚类分析，结合困惑度确定模型最优主题数，挖掘出 14 个潜在主题，并对主题内容细致分析，主要包括动物实验研究、药物质量标准与分析、基因分析与应用、中医药的临床研究、药物制备、活性成分分析等。通过查阅中医药学领域的综述性文献[19]-[31]，发现识别出的主题符合专家研究方向，以及其中近 10 年出现的大量新型冠状病毒肺炎研究，结合主题均符合研究规划，证明了结果是准确的，进而说明该研究方法是合理的。针对中医药学研究主题，研究者应更加关注中医理论现代化研究，包括基础理论的现代解释和中医辨证论治的科学化；中药的研究与开发；中医临床疗效的科学评估；中医药与信息技术结合；中医药的心理学、精神疾病治疗和癌症治疗，紧密结合现代科学技术，探索中医药理论、药物、疗法等在现代社会的应用。

本文的不足之处在于选取数据集时只选取了近 30 年数据，并且仅仅考虑了中医药领域的个别专业期刊，主题挖掘可能不全面，如蛋白质组学、红外热成像技术、骨质疏松等主题未被挖掘出来，无法展示中医药领域的全部研究主题和发展方向。需考虑扩大数据集进行分析，进而充分了解中医药领域的发展现状；对于主题发现，由于 LDA 主题模型不能够对主题命名，需人工总结主题，对一些主题的命名解释缺乏专业性，后续希望通过查阅更多资料和咨询专业研究者使得对主题模型的结果解释更加合理化。

参考文献

[1] 梅馨, 邢桂芬. 文本挖掘技术综述[J]. 江苏大学学报(自然科学版), 2003, 24(5): 72-76.

[2] 谌志群, 张国煊. 文本挖掘研究进展[J]. 模式识别与人工智能, 2005, 18(1): 65-74.

[3] 周茜, 赵明生, 扈旻. 中文文本分类中的特征选择研究[J]. 中文信息学报, 2004, 18(3): 18-24.

[4] 翟东海, 鱼江, 高飞, 于磊, 丁锋. 最大距离法选取初始簇中心的 K-means 文本聚类算法的研究[J]. 计算机应用研究, 2014, 31(3): 713-715, 719.

[5] 王鹏, 高铖, 陈晓美. 基于 LDA 模型的文本聚类研究[J]. 情报科学, 2015, 33(1): 63-68.

[6] 林鸿飞, 战学刚, 姚天顺. 基于概念的文本结构分析方法[J]. 计算机研究与发展, 2000, 37(3): 324-328.

[7] 徐妙君, 顾沈明. 面向 Web 的文本挖掘技术研究[J]. 控制工程, 2003, 10(z1): 44-46, 50.

[8] Hirano, M., Sakaji, H., Kimura, S., Izumi, K., Matsushima, H., Nagao, S., *et al*. (2019) Related Stocks Selection with

Data Collaboration Using Text Mining. *Information*, **10**, Article 102. https://doi.org/10.3390/info10030102

[9] Wermter, J., Fluck, J., Stroetgen, J., Geißler, S. and Hahn, U. (2005) Recognizing Noun Phrases in Biomedical Text: An Evaluation of Lab Prototypes and Commercial Chunkers. *CEUR Workshop Proceedings*, Hinxton, 10-13 April 2005, 148.

[10] Nemrava, J. and Svatek, V. (2005) Text Mining Tool for Ontology Engineering Based on Use of Product Taxonomy and Web Directory. *Annual International Workshop on Databases, Texts, Specifications and Objects*, Desná, 13-15 April 2005, 129.

[11] 宋军, 肖超. 上市公司年报风险信息披露与市场风险——基于 LDA 主题模型的文本研究[J]. 复旦学报(社会科学版), 2024, 66(2): 165-176.

[12] 陈秋英, 宋姗姗. 国外智慧旅游政策和理论的主题建模及趋势研究[J]. 科技和产业, 2024, 24(5): 56-64.

[13] 杨海霞, 高宝俊, 孙含林. 基于 LDA 挖掘计算机科学文献的研究主题[J]. 现代图书情报技术, 2016(11): 20-26.

[14] 陈壮, 贾成贺, 姜红. 显微共聚焦激光拉曼光谱结合 PCA-HCA-LDA 对便签纸的识别分类[J]. 应用激光, 2024, 44(2): 94-103.

[15] 王伟, 高宁, 徐玉婷, 王洪伟. 基于LDA 的众筹项目在线评论主题动态演化分析[J]. 数据分析与知识发现, 2021, 5(10): 103-123.

[16] 胡泽文, 韩雅蓉, 王梦雅. 基于 LDA-Word2vec 的图书情报领域机器学习研究主题演化与热点主题识别[J]. 现代情报, 2024, 44(4): 154-167.

[17] 王一博, 张鹏翼. 基于 LDA 模型的国内用户画像研究主题及演化分析[J]. 情报探索, 2024(2): 99-105.

[18] 庞庆华, 徐珣, 张丽娜. 融合微博多维特征和用户动态兴趣的主题推荐研究[J]. 数据分析与知识发现, 2025, 9(1): 110-120.

[19] 陈科旭, 王志荣, 苗明三. 基于中西医临床特点的干燥综合征动物模型分析与展望[J]. 中华中医药杂志, 2024, 39(3): 1427-1430.

[20] 赵磊, 齐芳华, 莫蕊辰, 等. 冠心病的中西医病机研究进展[J]. 中国中医药现代远程教育, 2024, 22(1): 154-156.

[21] 孙东文, 姜文君, 邰晓峰, 等. 中西医对肠道菌群与冠心病关系的研究及治疗进展[J]. 吉林中医药, 2023, 43(12): 1488-1492.

[22] 郑先丽, 严兴科, 姚小强, 等. 针刺治疗创伤后应激障碍的思路探讨[J]. 中医研究, 2023, 36(12): 89-92.

[23] 郅扶旻, 孙士博, 徐洪涛. 中西医结合治疗糖尿病性骨质疏松症进展[J]. 云南中医中药杂志, 2023, 44(10): 94-97.

[24] 帅眉江, 谢兰香. 中西医结合诊治老年高血压病研究进展[J]. 中国民间疗法, 2023, 31(4): 102-106.

[25] 李小薇, 毛浩萍. 功能性消化不良的临床研究进展[J]. 中医药学报, 2022, 50(2): 82-87.

[26] 张宜帆, 周曼丽, 罗晓欣, 等. 代谢组学在冠心病中医证候中的研究进展[J]. 中医药通报, 2022, 21(1): 57-60.

[27] 钟森杰, 李琳, 胡思远, 等. 中医病因型证候模型建立的思考[J]. 中国中医基础医学杂志, 2022, 28(2): 310-314.

[28] 赵庆大, 旋静. 肿瘤治疗中辨病论治与辨证论治相结合的应用综述[J]. 解放军医学院学报, 2021, 42(9): 993-996.

[29] 姜琦, 郭会军, 李鹏宇, 等. 艾滋病 HAART 后高脂血症的临床研究进展[J]. 中医研究, 2020, 33(4): 63-67.

[30] 戴中上, 钟严俊, 陈燕. 慢性阻塞性肺疾病合并支气管扩张症的研究进展[J]. 结核与肺部疾病杂志, 2023, 4(6): 499-505.

[31] 滕石山, 瞿小旺. 新型冠状病毒中和抗体应答研究进展[J]. 中国科学: 生命科学, 2023, 53(10): 1490-1498.