Educational and teaching research, 教育教学研究 (16)2022,4
ISSN:2705-1277(Online); 2737-4130(Print)

基于LDA模型的国内新冠疫情文献主题聚类分析

王雅娇　曾骏程　崔基哲 *

（延边大学经济管理学院　吉林　延边）

摘　要：新冠疫情自2019年爆发以来，各领域学者深入研究分析自新冠病毒到社会防控乃至国际经济等重要议题。本文以中国知网为文献数据来源，对2020年至2022年期间发表的新冠疫情相关的各领域文献进行研究，以LDA模型挖掘文献主题类别，呈现文献主题聚类结果，并将该结果与Citespace的聚类结果进行对比并检验。筛选后得到2020-2022年期间关于新冠主题研究的中文文献8099篇，根据LDA聚类结果划分为国际关系、经济发展和信息传播等七个主题，并利用Citespace针对文献中的关键词进行聚类分析和可视化展示。通过本文对新冠疫情相关文献的类别梳理及分析，获得了后疫情时代的社会维稳相关的有效数据支撑，为后续相关研究及公共卫生领域提供可预测发展势态相关因素为基础的分析方法，提出了一种社会稳定趋势把控理论依据。

关键词：新冠疫情　主题聚类　LDA　Citespace

Theme clustering analysis of the domestic COVID-19 literature based on the LDA model

Yajiao Wang, Zeng Juncheng, Cui izhe *

（**School of Economics and Management, Yanbian University, Yanbian, Jilin**）

Abstract: Since the outbreak of COVID-19 in 2019, scholars in various fields have deeply studied and analyzed important issues from novel coronavirus, social prevention and control and even the international economy. In this paper, the literature related to COVID-19 published from 2020 to 2022 was studied, and the LDA model was used to mine the topic categories of the literature, and the clustering results of the literature topics were examined, and compared the results with the clustering results of Citespace. After screening, 8,099 Chinese documents on COVID-19 from 2020-2022 were obtained. According to the LDA clustering results, they were divided into seven themes, including international relations, economic development and information dissemination, and cluster analysis and visual display of the keywords in the literature using Citespace. Through the category sorting and analysis of the COVID-19 related literature, this paper obtains the effective data support related to social stability maintenance in the post-epidemic era, which provides an analysis method based on the factors related to the predictable development situation for the subsequent relevant research and public health field, and puts forward a theoretical basis for controlling the trend of social stability.

Key words: COVID-19 theme cluster LDA Citespace

1 引言

2022 年 12 月 7 日，国务院联防联控机制综合组出台了《关于进一步优化落实新冠肺炎疫情防控措施的通知》，该新十条的出现，标志着经过三年全国各族人民团结一致，听从党和国家的统一指挥，艰苦卓绝的斗争后，对于自 2019 年末爆发的新型冠状病毒肺炎疫情，在党和政府的领导下防控政策得到了优化，防疫正式步入了后疫情时代。抗击疫情不仅仅是人民群众与新冠病毒的抗争，更是科学工作者与时间的争夺。自 2019 年以来，各领域学者对新冠疫情所带来的诸多影响都开展了各自领域的研究。在大量研究文献涌现的背景下，利用定量分析探讨研究的热点和发展方向，为各领域后续研究与总结提供参考则至关重要。

现有针对新冠肺炎疫情各领域研究现状与发展趋势的研究，主要聚焦在以下两个角度：一方面，部分学者采用内容回顾等定性分析的方法，针对新冠疫情与群众之间的情感联系，综述新冠肺炎疫情对社会舆情与群众情感的影响。另一方面，少数学者运用数学模型法等定量分析方法，对疫情爆发以来新专利进行探讨，从而识别研究主题。上述研究一定程度上揭示了新冠肺炎疫情的前沿研究与发展态势，但缺乏对于新冠疫情整体的热点分析与趋势把控，因此本文将在现有的研究基础上，采用 LDA 模型和 Citespace 工具，完成对国内新冠疫情文献主题聚类分析，实现对新冠疫情相关文献的研究内容和趋势的把控。

2 研究方法

2.1 LDA 的研究理论及运用

LDA(Latent Dirichlet Allocation)，即潜在狄利克雷分布，是基于贝叶斯模型的话题模型，是一种无监督学习算法，用于识别文档集中潜在的主题词信息。该模型在训练时不需要手工标注的训练集，需要的仅仅是文档集，并指定主题的数量 K，并对每个主题聚类词语进一步描述。正因如此，LDA 在文本信息处理等领域被广泛使用。

LDA 模型分析包含文档、主题和词语三层结构。其主要思想是：文档是由若干主题组成的，主题是由文档中一组特定词汇组成的，文档中的每个词都是以一定的概率分布的，由此可将一篇文档的主题以出现概率最高的一组词汇表示。LDA 主题模型在文档、主题、词语三个层次上进行概率建模，计算主题与文档、主题与词语之间的语义关联度，为科学文献主题挖掘提供了方法和思路。

对于主题模型而言，选定的算法一旦确定，需要人为确定的选项（超参）通常是主题数量，LDA 算法也不例外：主题数量通常视不同场景进行调整，对于 LDA 来讲主题抽取的效果与潜在主题数 K 相关，本文中采用计算困惑度（Perplexity）的方法来衡量选取的主题数量的优劣[3]：

$$\text{Perplexity}(D_{\text{test}}) = \exp\left\{\frac{-\sum_{d=1}^{M}\log(p(w_d))}{\sum_{d=1}^{M}N_d}\right\} \quad (1)$$

其中，M 是测试语料的大小（即文献的数量），Nd 是第 d 篇文本大小（word 或 token 个数）

$$\sum_{z} p(z)p(w \mid z,r) \quad (2)$$

其中，z 是主题，w 是文档，r 是基于训练集学习的文本 - 主题分布，Perplexity 对数函数的分子部分是生成整个文档集的似然估计（表示训练集训练出的参数的生成能力）的负数，由于概率取值范围为 [0，1]，按照对数函数的定义，分子值是一个正值且与文本生成能力正相关；而分母是整个文档集的单词数目，即模型生成能力越强，Perplexity 值越小。

本文将基于 LDA 模型对新冠疫情爆发以来学界针对领域新冠疫情相关文献的摘要进行高频词提取聚类展示，研究过程包括文献收集、文本预处理、主题数量选取、主题聚类分析等。

2.2 Citespace 的理论及运用

CiteSpace[1] 是探测某一学科或领域结构、规律以及分布的一种可视化分析软件，借助此工具形成科学知识图谱，可直观地对学界的发文量，文献合作度，高频研究机构等全面了解，从而挖掘该领域的研究热点、前沿地带及发展走向。

本文采用 CiteSpace 6.1.R4 软件对新冠疫情相关文献进行计量分析，主题节点为文献关键词。其中软件生成的网络图谱中节点的大小表示关键词出现次数多少或被引用频率的高低，各节点之间的线条粗细表示节点之间联系紧密程度，线条越粗、颜色越深代表节点之间的联系越紧密，有着更深的联系关系 [2]。

本文将基于 Citespace 对新冠疫情爆发以来，学界针对领域新冠疫情相关文献的关键词进行可视化图谱生成，并对高频关键词完成聚类结果。

3 实验结果与可视化

3.1 LDA 模型主题分析

3.1.1 数据获取

本文选取的文献数据来源为中国知网。为更好地控制本文主题研究的检索要求，兼顾检全率与检准率，减少后续的数据规范和筛选，本文对文献的检索条件为：((主题 %=' 新冠 ' or 题名 %=' 新冠 ') OR (主题 %=xls('COVID-19') or 题名 %=xls('COVID-19')) OR (主题 %=' 新型冠状病毒 ' or 题名 %=' 新型冠状病毒 ') OR (主题 %=' 疫情 ' or 题名 %=' 疫情 ')) AND ((年 Between('2020', '2022')) AND (CSSCI 期刊 ='Y'))；检索范围为期刊。经过剔除部分缺失摘要的文献以及主题与新冠疫情偏离的部分文献后，共筛选出新冠疫情相关文献共 8099 篇。

3.1.2 数据预处理

根据上述步骤使用自定义批量以 Excel 格式导出题目、摘要等文献信息后形成初始实验数据集，并对不完整或重复出现的文献记录进行人工的补充或删除之后形成研究语料库。再将语料中的摘要部分存在一些无实际意义的高频虚词和词频过高的可能影响聚类结果的特征词如“新冠疫情”、“防控”、“疫情”等词语进行停词处理，采用 Python 中的 Jieba 模块调用停用词表进行分词处理。

3.1.3 主题数量选取

利用 Python 对于上述公式（1）公式（2）进行实现，并对各不同个数主题的 LDA 模型的困惑度进行可视性比较，得出的结果如图 1 所示。

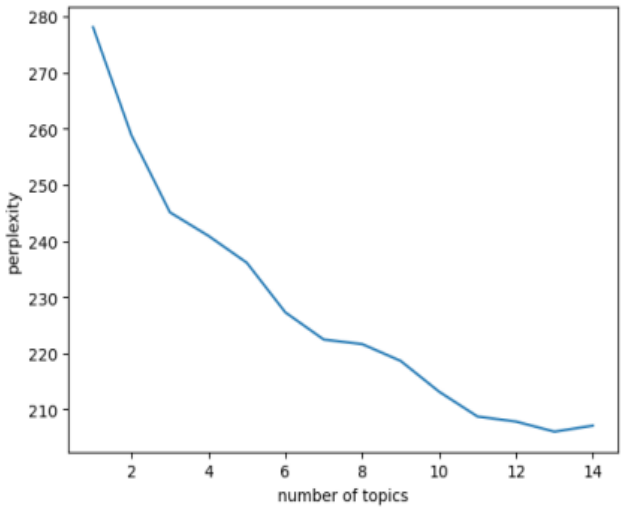

图 1. 不同个数主题下 LDA 模型的困惑度

由前文的分析，可知困惑度数值越低，代表着模型拟合结果越好，即选择的主题数量合适且具有代表性，但当主题太多时，模型可能出现过拟合情况。可以发现当主题个数为 7 时，模型的困惑度有明显的下降，所以本文考虑选择在 7 个主题个数作为最终的主题个数。

3.1.4 LDA 主题聚类

本文针对 Latent Dirichlet Allocation 类的主要输入参数设置如下表所示：

表 1. 基于 LDA 模型的主要参数设置

代码设置	代码含义
n_components=7	待识别的主题数量为 7
max_iter=50	模型在训练过程中最多迭代 50 次
learning_method=’batch’	采用变分推断 EM 算法
doc_topic_prior=0.1	“文档 – 主题”分布的先验概率的 0.1
topic_word_prior=0.01	“主题 – 词”分布的先验概率为 0.01

经过 LDA 无监督学习对导入文献摘要的文本数据进行分析之后，输出每个主题对应的高频特征词并进行聚类，得到表 2 结果：

表 2. 基于 LDA 模型的文献摘要中高频特征词聚类结果

主题编号	每个主题下的前 20 个高频特征词
主题 1	全球 国际 合作 世界 国家 人类 共同体 战略 全球化 挑战 经济 命运 关系 时代 体系 公共卫生 政治 问题 领域 外交
主题 2	经济 企业 数字 冲击 市场 产业 影响 全球 消费 贸易 政策 供应链 转型 旅游 投资 产业链 生产 金融 行业 数字化
主题 3	信息 传播 媒体 网络 社区 研究 事件 公众 数据 公共卫生 分析 新闻 舆情 平台 传染病 社会 媒介 方法 舆论 内容
主题 4	影响 风险 政策 研究 政府 分析 模型 事件 因素 机制 心理 效应 危机 韧性 程度 理论 冲击 空间 特征 地方
主题 5	防控 应急 工作 公共卫生 体系 能力 人民 管理 图书馆 国家 服务 抗疫 建设 精神 抗击 事件 制度 优势 领导 机制
主题 6	社会 危机 制度 政治 国家 价值 文化 问题 法律 理论 原则 主体 体系 逻辑 关系 政府 模式 机制 风险 生命
主题 7	教育 教学 服务 高校 研究 技术 体育 学生 时代 大学生 教师 问题 资源 融合 模式 分析 直播 方面 建设 方式

根据上述 LDA 的聚类结果，本文针对七个生成主题逐一分析：主题 1 为全球化合作交流，面对此类全球化公共卫生时间，各国之间都应当守望相助，向社会公开透明地分享本国的防控现状和本国政府的防疫政策，加强合作与交流才能更好减缓疫情的传播与影响；主题 2 为经济冲击，疫情对于国家的影响最直观的表现在经济层面，如何科学安全生产以及如何稳步复工复产，使经济水平回升也是研究热点；主题 3 为信息传播，在及时播报病例情况以及控制社会舆情方面，新媒体的发展与研究也至关重要；主题 4 为政府制度，通过对各国的防控政策的分析来进行总结；主题 5 为应急防控，新冠疫情传播迅速，面对突然出现的病毒该如何防控，应急措施是否安全有效且如何改进也是研究的热点；主题 6 为生命安全，在疫情期间疫苗的研发和特效药的制作贯穿始终，各类关于疫情本身和生命保障的研究涌现；主题 7 为网络教学，在疫情期间停工不停学，如何通过升级技术，改善制度做到高效且稳步开展教育教学也是重中之重。

在该聚类结果的基础上，进一步实现可视化验证，可以看出作为代表不同主题之间的缩放圈不存在重合，可以认为 7 个主题下的 LDA 模型拟合信度较好 [3]，结果如图 2 所示：

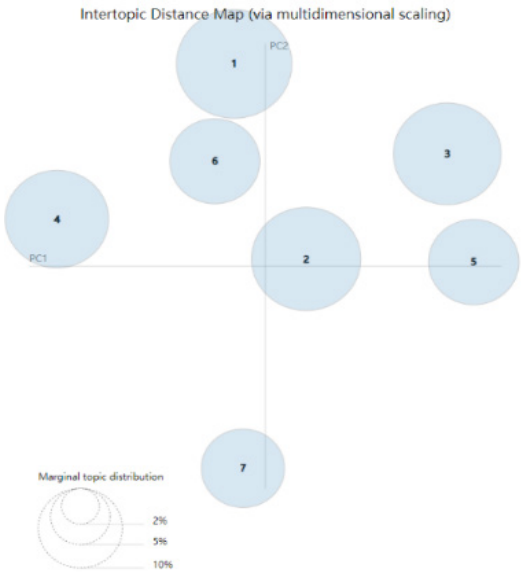

图 2.Intertopic 多维缩放距离图

3.2 Citespace 结果可视化与对比

本文以文献中初始关键词为网络节点，选择时间范围为 2020 至 2022 年，时间切片为一年。采用 Minimum spanning tree 剪枝算法，该剪枝算法运算相对便捷，可使图谱展示效果更为清晰[4]。共生成 479 个节点，434 条连线，密度为 0.0064 的关键词共现图谱，出现频次最高的关键词是疫情防控。

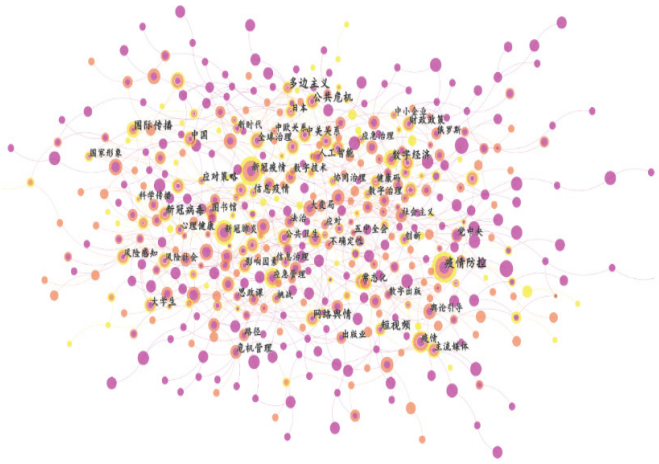

图 3.Citespace 关键词网络图谱

在图谱生成的基础上，本文对关键词进行聚类分析，聚类算法选择 LSI。研究结果显示，Q =0.47512，S = 0.7299，Q 值大于 0.3 说明划分出来的聚类结构是显著的；S 值大于 0.7，说明聚类效果明显。信度较好。共形成 7 个聚类标签，其中聚类标签序号越小，包含节点越多，所生成的关键词高频特征词的聚类结果如表 3 所示。

表 3. 基于 Citespace 的文献关键词中高频特征词聚类结果

主题编号	每个主题下的前 10 个高频特征词
主题 1	新冠疫情；委托代理；分级诊疗；激励相容；机制设计	再全球化；新冠肺炎疫情；数字全球化；“慢球化”；宏观经济
主题 2	疫情防控；政府信任；媒介依赖；公众风险感知；新冠肺炎疫情	宗教领域；宗教活动场所；宗教教职人员；广东湛江；宗教界人士
主题 3	公共危机；社会治理；书香战疫；需求理论；质性研究	知识服务；传染病健康教育；协同机制；社会组织；新型举国体制
主题 4	新冠疫苗；btm 主题模型；关联规则；内容分析；主题发现	知识产权；强制许可；公共健康；trips 协定；公民服从
主题 5	新冠肺炎；数据共享；元数据标准；数据开放；媒体依赖	社会规范；媒介接触；集体行动；接种意愿；新冠疫苗
主题 6	新冠病毒；受体结合域；亚单位疫苗；中和抗体；多肽药物	科学传播；多元主体；参与模式；突发性公共卫生事件；疫情科学信息
主题 7	健康传播；情绪传播；风险感知；反霸权新闻；报道框架	报道信；反霸权新闻；报道框架；国际传播；情绪传播

从特征词的结构角度看，Citespace 的高频特征词的呈现以四字为主，主要产生聚类区别的原因在于该特征词的来源为文献的关键词。与此同时，该高频特征词的抽取过程中缺少停词的步骤，所以使“新冠疫情”等高频词出现在多个聚类结果中。从类别的结果角度看，Citespace 主题词聚类后的结果与 LDA 模型主题分类大致相同。

4 结论

本文搜集了中国知网数据库在 2020-2022 年期间所收录新冠疫情的相关文献，基于 LDA 无监督学习模型分析文献摘要生成各领域期刊论文研究热点的聚类结果，并利用 Citespace 生成关键词图谱，并对高频特征词进行聚类分析，并对不同的主题之间的结果进行了详细的解析。

本文通过对主题的聚类结果深入挖掘，发现在新冠疫情发生期间，学者们首先对于科学防控进行研究，其中包括如何在疫情中保护人民生命安全，有效地遏制病毒的扩散、蔓延和传播在疫情期间显得尤为重要；其次在当今科技传播手段多元化的背景下，这次疫情在信息传播和舆论控制的研究也相较于人类近代史上其他重大公共卫生危机更加深入，现今我国放开政策，如何在后疫情时代做好宣传，合理控制好群众情绪，正确引导社会舆论就需要从疫情期间有关数字媒体传播的文献中总结经验教训；随着全球化的发展，COVID-19 不仅对中国，而且对全世界都是一个巨大的公共卫生挑战，在类似重大公共事件上，应当展现中国政策，贡献中国力量；根据聚类结果进行分析，人类共同体和政府风险政策也是研究热点，全球疫情发展仍然变化莫测。科研人员需要与热点命题相结合，加大科研合作，更深入地对新冠领域进行研究，不断探索和拓宽公共卫生领域研究。

本文的创新点在于以文献摘要为研究数据，将 LDA 模型引入新冠疫情主题的研究中，采用困惑度确定预生成的主题数量，客观的描述了学者对其主要研究方向和视角。与此同时，本文基于文献关键词，进一步利用 Citespace 生成图谱，增强对文献文本数据的可视化，在一定程度上克服了主题聚类过程中的主观性。但本文出于对文献质量及文献数量的考虑，可能导致对部分领域文献的关注度较少，缺少对新冠疫情部分领域的主题聚类结果，如医药领域等。在未来研究中将扩大数据覆盖面，更深层次地探讨新冠疫情对于学术界以及社会面的影响。进一步为后疫情时代的社会维稳提供有效数据支撑，为从多角度识别新冠疫情相关的文献主题拓宽思路，从而为后疫情时代我国的科学防控与公共卫生领域的总结发展提供更全面的分析方法和理论支持。

参考文献

[1]Chen C,CiteSpace Ⅱ：Detecting and Visualizing Emerging Trends [J]. Journal of the American Society for Information Science & Technology,2006，57(3):359.377.

[2] 蓝蕾，瞿心远，应曜宇 . 基于 CiteSpace 对新型冠状病毒肺炎疫苗文献的可视化分析 [J]. 现代医院 ,2022,22(02):286-291.

[3] 刘华玲，王希睿，孙爱华 . 基于 LDA 主题模型舆情预测与可视化——以 COVID-19 为例 [C]//. 第十六届（2021）中国管理学年会论文集 .2021:303-316. DOI:10.26914/c.cnkihy.2021.055100.

[4] 陈悦，陈超美，刘则渊，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.

基金项目：

2021 年度吉林省大学生创新创业计划训练项目延边大学大学生创新创业训练资助项目“国际评价指标体系与制度性话语权关系的探究——以疫情期间 GHS 指数为例”（项目编号：202110184108）。