

基于LDA2vec模型老年人主观幸福感主题时序演化分析

陈婉铭, 刘媛华*

上海理工大学管理学院, 上海

收稿日期: 2025年6月3日; 录用日期: 2025年7月23日; 发布日期: 2025年8月1日

摘要

关注老年人的主观幸福感在积极应对老龄化方面具有重要意义。本文采用LDA2vec主题模型对中国知网数据库中老年人主观幸福感领域文章进行主题挖掘, 运用TF-IDF算法、LDA模型结合Word2vec词向量模型, 从时间维度上深入挖掘老年人幸福感的核心主题及其演变路径, 得到“养老模式与社会演化的关系”、“社会关系与偏远地区老年人”、“跨文化视角下的老年人幸福感”和“健康与数字老龄化”四个主题。通过计算主题热度, 得到近五年的主题热度趋势结果。同时在时间维度上讨论了各主题的拐点时间和首次发文时间, 并可视化三个时间窗口上主题演化情况, 直观呈现了老年人主观幸福感文章的主题结构和演化趋势。从研究热点看, “心理健康”与“社会支持”是该领域的重要研究主题。从整体上看, 主题间的交叉融合不断发展, 研究的主题逐渐多样化。

关键词

LDA2vec模型, 主题挖掘, 时序主题演化, 老年人主观幸福感

Analysis of the Time-Series Evolution of Subjective Well-Being Themes among the Elderly Based on LDA2vec Modeling

Wanming Chen, Yuanhua Liu*

Business School, University of Shanghai for Science and Technology, Shanghai

Received: Jun. 3rd, 2025; accepted: Jul. 23rd, 2025; published: Aug. 1st, 2025

*通讯作者。

文章引用: 陈婉铭, 刘媛华. 基于 LDA2vec 模型老年人主观幸福感主题时序演化分析[J]. 运筹与模糊学, 2025, 15(4): 85-97. DOI: 10.12677/orf.2025.154196

Abstract

The subjective well-being of older adults plays a significant role in actively addressing the challenges of aging. This study uses the LDA2vec topic model to perform topic extraction on articles related to elderly subjective well-being from the CNKI database. By applying the TF-IDF algorithm and combining the LDA model with the Word2Vec word vector model, the study delves into the core themes and evolution paths of elderly well-being from a temporal perspective. Four key themes were identified: "the relationship between elderly care models and social evolution," "social relationships and elderly individuals in remote areas," "elderly well-being from a cross-cultural perspective," and "health and digital aging." By calculating the topic popularity, the study presents the trend of topic popularity over the past five years. Additionally, the turning points and first publication times of each theme were discussed from a temporal dimension, and the evolution of topics across three time windows was visualized, offering an intuitive presentation of the topic structure and evolution trends in research on elderly subjective well-being. In terms of research hotspots, "mental health" and "social support" emerged as key themes in this field. Overall, the intersection and integration between topics have continued to evolve, with research themes gradually diversifying.

Keywords

LDA2vec Model, Topic Extraction, Temporal Topic Evolution, Elderly Subjective Well-Being

Copyright © 2025 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

根据国家统计局公布的人口普查数据,截至 2024 年底,我国 60 周岁及以上老年人口已达 3.1 亿,首次突破 3 亿人;其中 65 周岁及以上老年人口超过 2 亿,占比 15.6%。人口老龄化趋势日益加剧,因此积极应对老龄化具有重要的意义。主观幸福度被视为在积极老龄化中,探究老年人生活品质的关键指标[1],主观幸福感反映了个体对其生活体验的看法,侧重于个人对日常生活的满足感和自我评估的整体感受,它关注的是从老年群体自我生活和成长的视角出发,对老年人的心理健康与情感福利进行的重要研究。因此,改善老年人的生活水平,丰富他们的精神和文化活动,以进一步增强老年人的生活满意度和幸福度,已成为社会各领域需重点关注的问题[2]。近年来,我国学者对老年人主观幸福感研究开展了多方位、深层次的探索,然而缺少深入挖掘老年人幸福感背后的主题和演化规律。在文本挖掘和语言处理领域,主题模型作为一种统计方法,能够揭示文本中隐藏的主题信息。研究领域的热点分析对于发现学科及领域的发展路径和演化脉络至关重要,它为研究人员和政府部门提供了有价值的参考,对科学研究和学科发展具有显著意义[3]-[5]。作为科研成果的主要展示方式,期刊文章的内容分析和挖掘能帮助深入了解学科发展方向和研究现状[6]。学科主题的演化过程表现为复杂的非线性系统行为,系统在有序的动态平衡中逐步从低度有序过渡到高度有序演化状态,从而实现主题演变的良性循环[7]。通过对知识结构及其相互关系的综合分析,可以有效揭示研究热点和学科领域的演变趋势[5]。

LDA 模型作为广泛应用的主题建模工具,一直不断衍化和改进。同时,许多研究者使用不同的主题建模方法来分析各个领域的研究内容,应用场景也很广泛。在工业领域,Hu Z 等人[8]等人使用改进的加

权潜在狄利克雷分配模型和 Word2vec 方法从故障日志中学习维护经验, 改善了车辆维修诊断。施文等人[9]基于 SeqLDA (序贯潜在狄利克雷分配)分析消费者车辆缺陷投诉数据, 为汽车企业在召回操作和消费者权益保障提供理论依据和实际操作的支持。在医疗健康领域, Chen L 等人[10]对中国大陆报纸中对痴呆的描述进行了 LDA 主题建模分析, 以追踪中国相关政策中痴呆知识关键领域的变化。Pais N 等人[11]提出了基于随机特征选择的潜在狄利克雷分配(RFSLDA), 加入特征选择技术, 以提高整体分类性能, 以研究肠道微生物组对受试者健康状况的影响。陶成熙等人[12]利用 BERTopic 模型挖掘数据要素交易相关政策, 揭示政策核心主题以及不同时期各核心主题的演化趋向。

LDA2vec 是一种融合深度学习与主题建模的技术路径, 其基本理念是将潜在狄利克雷分配(LDA)模型与 Word2vec 词嵌入方法相结合, 通过神经网络优化机制, 实现对主题分布和词向量的联合表示学习。该方法在提升语义理解能力的同时, 也增强了模型在主题挖掘中的表达效果。探究领域的热点趋势能够把握学科研究方向, 但运用深度学习方法分析该领域热点趋势的研究还较少, 而且目前科学技术发展路线图中常常以文献数量的平均时间作为关键节点的参考, 但实际论文发表的增长趋势并非简单的线性上升, 依赖时间轴的平均时间可能并不够切合实际[13]。鉴于此, 本文将在运用 LDA2vec 主题模型基础上加入时间维度的考虑, 分析各主题首次出现时间和拐点的出现时间, 并可视化其结果挖掘其中的内在关联性和发展规律。以老年人主观幸福感为研究对象, 运用词频-逆文本频率(TF-IDF)算法、LDA 模型结合 Word2vec 词向量模型, 从时间维度上深入挖掘老年人幸福感的核心主题及其演变路径, 构建了一个从数据收集与处理、到主题挖掘与演化分析的全面研究框架, 旨在为主题挖掘和演化分析领域提供科学依据和参考。

2. 研究设计

2.1. 研究方法

2.1.1. LDA2vec 主题模型

隐含狄利克雷分布(LDA)主题模型由 Blei David M.、Ng Andrew Y.、Jordan 于 2003 年提出, 用来推测文档的主题分布[14]。LDA 作为一种机器学习模型, 在文本挖掘领域具有显著的应用潜力, 其优势在于能够有效地揭示文本数据中潜在的主题结构, 实现挖掘和分析出更加丰富、全面的内容信息[15]。LDA 模型能够捕获词汇之间的相关性, 然而却无法有效表达主题之间的相关性。

LDA2vec 主题模型是 2016 年由克里斯·穆迪(Chris Moody)等人提出的一种主题模型, 它将 LDA 主题模型与 Word2vec 词向量模型结合, 利用深度学习技术增强了主题分析的效果。LDA 主题模型对于解释性强的文档主题提取效果表现良好。与此同时, Word2vec 通过 Skip-gram 和 CBOW 模型对词向量进行训练, 并被广泛应用于改进文本表示效果。Word2vec 模型是一种基于人工神经网络的词嵌入方法, 由输入层、输出层和隐藏层组成[16]。Word2vec 词向量模型算法可以将词转换为低维实数向量, 从而在向量空间中对其相似性进行量化。因此, 在这种转化后的向量空间中, 距离较近的词表明它们在语义上具有更高的相似度[17]。通过学者的实验, 基于 Word2vec 词向量的特征选择分类效果优于传统的特征选择[18]。采用 TF-IDF 度量对 Word2vec 模型进行加权。在主题提取方面, 运用加权词方案优于非加权方案[19]。

LDA2vec 主题模型结合了 LDA 和 Word2vec 两种方法的优势, 能够优化组合词向量、主题向量和文档向量, 通过引入神经网络的深度学习理念, 更准确地表达文档主题的聚类 and 主题词的关系[20]。LDA2vec 模型通过引入词向量, 使主题建模过程不仅依赖词频共现关系, 更融合了词语之间的分布式语义信息。在传统 LDA 中, 不同词项只根据文档级共现统计进行主题归类, 容易忽视词语间的深层语义关联, 导致出现语义相近词分布于不同主题的情况, 从而影响主题聚合的准确性与一致性。而在 LDA2vec 中, 每个

词被表示为一个稠密的词向量, 能够捕捉其在上下文中出现的语义偏好。例如, “满足” “快乐” “幸福” 在词向量空间中具有较高的余弦相似度, 即使它们在文中出现频次不同、共现结构不完全一致, 也可被聚合进表达 “主观幸福感” 的同一主题。这种语义邻近机制有效弥补了 LDA 在语义识别能力上的不足。此外, 词向量还能加强对多义词与同义词的建模能力。例如, “养老” 既可以指政策, 也可以指照护服务, 而词向量根据上下文能学习到其具体语义倾向, 从而提升主题词聚合的语义准确性。本研究选用了 Word2vec 词嵌入方法中的连续词袋模型(CBOW), 通过上下文词语的词向量预测目标词的表示向量, 其网络结构如图 1 所示。

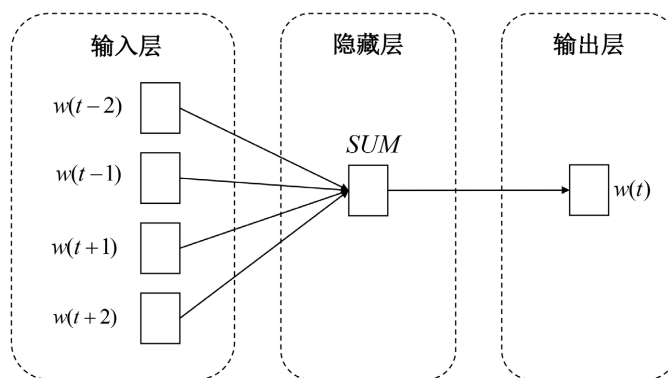


Figure 1. CBOW network architecture diagram
图 1. CBOW 网络结构图

$w(t)$ 为当前的目标词, 在构建 N 维词向量的前提下, 模型利用其上下文词汇作为输入对神经网络进行训练, 通过逐层调整参数矩阵, 实现目标词对应的隐藏层向量表示的学习。

LDA2vec 的核心思想是将文档表示为主题的加权组合, 并将每个主题表示为多个词嵌入向量的加权平均。其训练目标是在预测上下文词时最大化目标词和上下文向量的点积, 通过 Skip-gram 的方式学习词向量, 同时保留每个文档的主题分布信息。模型训练过程中, 联合优化词嵌入向量、主题向量和每个文档的主题分布三类参数。这种结构使得主题建模过程不仅依赖于词频共现信息, 也充分利用了词语的上下文语义特征, 从而缓解了传统 LDA 主题过于粗糙或语义碎片化的问题。

2.1.2. 最优主题数目的选择

在 LDA 模型中, 主题分布遵循 Dirichlet 分布, 该分布假设各主题的出现独立于其他主题的出现情况。这种独立性假设与实际数据存在矛盾, 导致 LDA 对主题数目 K 的变化极为敏感。大量实证研究表明, LDA 主题抽取的效果与潜在主题数目 K 值直接相关[21], 故确定最佳 K 值的方法显得尤为关键。

LDA 需要手动设定最佳主题数目 K 。有的学者用的是困惑度计算方法来判断最优主题数目。困惑度是文档内句子相似度几何平均数的倒数, 其值会随着句子相似性增加而逐步递减[20]。本文运用主题一致性也称为主题连贯度来选取最优主题数目。主题一致性是指各主题中词汇之间的联系程度。主题一致性是通过计算主题中重要词汇的语义相似性来评定该主题的得分。一致值越高, 聚类能力越好。计算公式为式(1)和(2) [22]:

$$coherence(t_h) = \frac{1}{\binom{t}{2}} \sum_{j=2}^t \sum_{i=1}^{j-1} \cos(wv_i, wv_j) \quad (1)$$

$$coherence(T) = \frac{1}{k} \sum_{h=1}^k coherence(t_h) \quad (2)$$

从 1 到 14 个主题的数据集对 LDA 模型进行评估, 并获得每个主题类别的连贯性。当一个主题内的词汇之间的语义关联更加紧密时, 该主题的一致性和模型的可解释性更强。

2.1.3. 主题热度

主题热度可以展示科学知识的表现形式, 揭示各类领域知识在不同时间段的活跃度与关注度, 体现出领域知识结构的纵向变化[3]。主题的热度由 LDA 模型中的文档与主题之间的概率关系推算, 反映某一主题在不同时间段所占的比例, 如公式(3)所示:

$$\bar{\theta}_k^d = \frac{\sum_{d=1}^M \hat{\theta}_k^d}{M} \quad (3)$$

在时间窗口 k 内, 某主题的热度是文档中该主题 d 所占比例的总和除以文档的总数。

2.1.4. 余弦相似度

通过建立主题之间的相似度, 可以呈现主题属性随时间的演变过程, 揭示科学知识的发展情况[3]。主题演化能够揭示和描述研究主题的各种变化过程, 构建研究主题的时序演化路径, 从而识别并探索科学研究的前沿领域[5]。

为了进一步挖掘不同阶段主题的演化关系, 本文采用余弦相似度计算方法(4), 分析不同时间窗口中主题之间的关系, 从而衡量主题间的相似程度。余弦相似度值域为[0, 1], 越接近 1 表示两个主题之间相似程度高。

$$Sim(T_i^t, T_j^{t+1}) = \frac{\sum_{k=1}^n p_k(T_i^t) \times p_k(T_j^{t+1})}{\sqrt{\sum_{k=1}^n p_k^2(T_i^t)} \times \sqrt{\sum_{k=1}^n p_k^2(T_j^{t+1})}} \quad (4)$$

T_i^t 为时间窗口 t 下的主题; T_j^{t+1} 为时间窗口 $t+1$ 下的主题; $p(T_i^t)$ 和 $p(T_j^{t+1})$ 分别为 T_i^t 和 T_j^{t+1} 的概率值, 在主题模型结果中, 该概率值即为“主题一词”矩阵中的值。

2.2. 演化关系判别

主题演化过程中包括 5 种现象, 分为新生、增长、融合、分裂、收缩、消亡[23]。主题新生指的是当前时间窗口内的新主题, 这些主题在之前的时间窗口中并未出现; 主题增长表示当前主题是前一时间窗口主题的延续; 主题融合是指前一时间窗口中两个或多个主题在当前时间窗口合并成一个新的主题; 主题分裂意味着前一时间窗口中的某个主题在当前时间窗口分裂成两个或更多的新主题; 主题收缩指的是前一时间窗口中存在但规模缩小的主题; 主题消失则意味着前一时间窗口中的某个主题在当前时间窗口消失。

2.3. 研究框架设计

综上, 本文对中国知网数据库中老年人主观幸福感的文章进行分析, 研究框架见图 2, 主要研究思路如下:

1) 文本采集: 从中国知网数据库获取主题为“老年人主观幸福感”的文章, 并将其储存至 Excel 中。

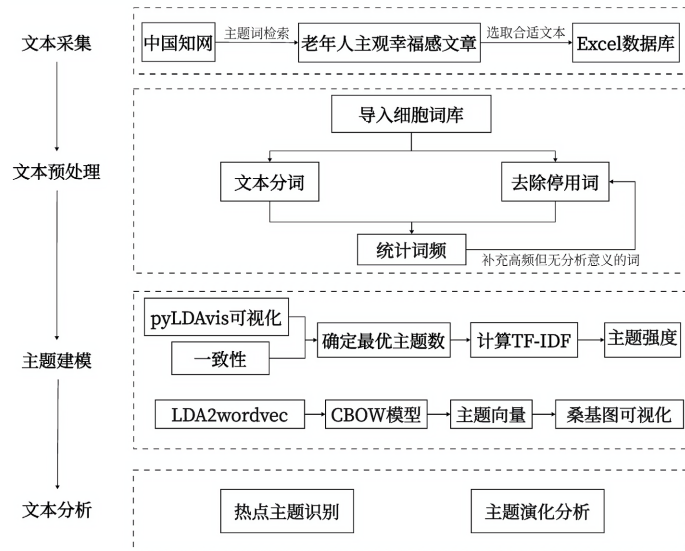


Figure 2. Research framework diagram of the theme mining of subjective well-being of the elderly
图 2. 老年人主观幸福感主题挖掘研究框架图

- 2) 文本预处理：对文本进行分词处理，并去除停用词，同时进行词频统计。
- 3) 主题建模：采用 LDA 模型对文本进行主题挖掘。
- 4) 文本分析：进行热点主题识别、时间点分析和主题演化分析，并结合主题内容提出相关建议。

3. 实证分析

3.1. 数据来源和文本预处理

本文采用主题词检索的方式，以中国知网数据库(CNKI)作为数据来源，以“老年人”和“主观幸福感”为主题词，从该领域出现首篇文献的年份开始，检索 1980 年到 2024 年期间发表的关于“老年人 + 主观幸福感”的文献数据，检索日期为 2024 年 3 月 2 日。CNKI 检索式为：SU = (‘老年人’ + ‘老人’ + ‘长者’ + ‘老年’)and SU=(‘主观幸福感’ + ‘生命满意度’ + ‘老年人幸福感’)。检索到中文文献 693 篇。本文所包含的文献仅限于期刊论文，排除了学位论文、会议论文、图书、专利、标准和研究成果等其他类型的文献。

目前，领域热点识别大多围绕着文章关键词分析，虽然关键词能够很好地反映一篇文章的中心主题，但关键词所表达的语义范围大小不一，这导致识别出的热点信息在语义表达上存在一定程度的不足 [23]。为了解决这个问题，本文从文章摘要作为出发点，不仅考虑文本结构和语义相关性，还考虑词语在上下文中的含义。

文本预处理对中文摘要进行分词、清洗掉无意义的无效字词和词频统计。文本分词和去除停用词采用的是 Python 中的 Jieba 模块精确模式，需要先建立自定义词库即专业术语词库有助于精确切分中文摘要文本，同时根据预分词词频统计结果补充停用词，例如：“老年人”、“老人”、“幸福感”等高频但无分析意义的词，从而减小对模型结果的干扰。

3.2. LDA2vec 建模

主题建模采用 Python 中的 Gensim 模块进行实验研究。本研究结合一致性计算结果和 pyLDAvis 模块得出的可视化聚类结果共同确定最佳主题数。

根据图 3 所示, 选择区间[1, 20]范围内的整数作为潜在主题数, 本文对不同主题数 K 下模型的一致性评估值(CV 值)进行了分析。结果显示, 当主题数 K 小于 5 时, 模型的 CV 值随着主题数的增加而逐渐上升; 然而, 当主题数 K 超过 5 时, 模型的 CV 值开始出现缓慢的波动下降趋势。特别值得注意的是, CV 值在主题数为 5 时达到峰值, 表明此时模型生成的主题更具有解释性和一致性。结合 pyLDavis 可视化见图 4, 当主题数为 4 时, 各主题之间兼具差异度以及特征词交叉范围小。因此本文选定了最优主题数为 4, 以确保生成的主题更加清晰易懂。在本研究中, LDA2vec 模型的关键参数设置如下: 主题数设置为 4, 经过多次实验与主题一致性评价选择最优值; 词向量维度设为 300, 与预训练 Word2vec 嵌入保持一致; 负采样数设为 15, 用于加快训练速度并提升词语区分性; 上下文窗口大小设为 5; 优化器采用 Adam 优化器, 学习率设为 0.001。此外, 为提升模型稳定性和主题聚合性, 本研究引入早停机制, 防止过拟合, 并在多轮实验基础上选取主题分布最清晰、语义最一致的模型结果。对中文文献进行全局主题建模, 得到 4 个主题, 主题关键词及主题归纳见表 1。

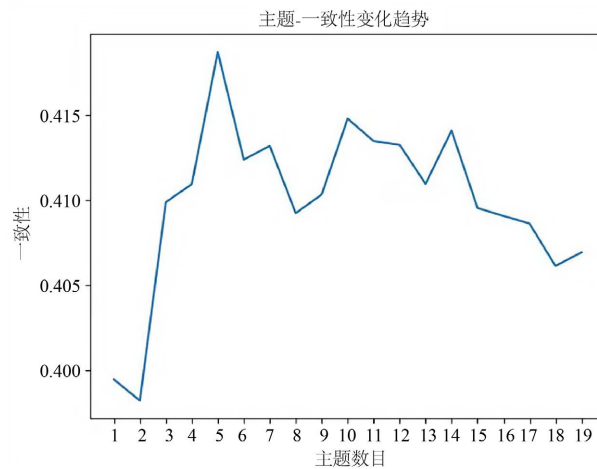


Figure 3. Trend of topic coherence from 2019 to 2024
图 3. 2019~2024 年的主题一致性变化趋势图

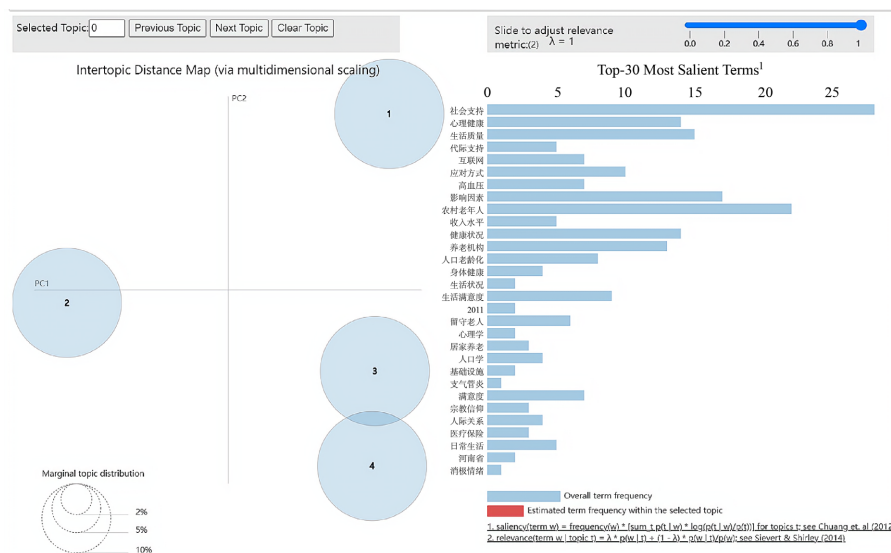


Figure 4. pyLDavis visualization
图 4. pyLDavis 可视化

Table 1. Keyword distribution results
表 1. 关键词分布结果

编号	主题标识	前 20 个关键词
主题 1	养老模式与社会演变的关系	社会支持、农村老年人、影响因素、心理健康、孤独感、体育锻炼、生活质量、养老机构、健康状况、生活满意度、应对方式、婚姻状况、慢性病、人口老龄化、互联网、身心健康、高血压、收入水平、糖尿病、随机抽样、日常生活、养老院、代际支持、性体验、社会保障、人口学、养老模式、满意度、CGSS、养老保险
主题 2	社会关系与偏远地区老年人	社会支持、农村老年人、影响因素、生活质量、健康状况、养老机构、婚姻状况、体育锻炼、慢性病、心理健康、应对方式、人口老龄化、满意度、生活满意度、孤独感、高血压、互联网、留守老人、社会保障、糖尿病、养老院、性体验、人际关系、追踪调查、身体健康、养老方式、身心健康、收入水平、身体状况、日常生活
主题 3	跨文化视角下的老年幸福感	社会支持、农村老年人、生活质量、心理健康、健康状况、影响因素、孤独感、养老机构、体育锻炼、应对方式、生活满意度、满意度、婚姻状况、人口老龄化、留守老人、糖尿病、互联网、高血压、社会保障、慢性病、身心健康、代际支持、养老院、幸福感、身体健康、国内外、日常生活、人口学、调节作用、纳西族
主题 4	健康与数字化老龄化	农村老年人、社会支持、影响因素、体育锻炼、健康状况、生活质量、养老机构、心理健康、婚姻状况、孤独感、应对方式、生活满意度、人口老龄化、慢性病、高血压、社会保障、满意度、糖尿病、日常生活、养老院、身心健康、性体验、幸福感、互联网、养老方式、代际支持、身体状况、经济支持、留守老人、调节作用

主题 1 涉及关键词“养老机构”、“人口老龄化”、“互联网”，这些词汇反映了现代社会在面对老龄化挑战时的演变和应对策略，家庭结构变化、子女外出务工与社会保障制度对养老选择的深层影响。随着农村劳动力外流，留守老人的养老需求越来越依赖社区与政府资源，这表明社会结构变迁已深刻重塑了养老方式的选择机制。随着社会结构的变化和人口老龄化的不断加剧，传统的家庭养老模式逐渐受到挑战，新型养老方式如养老机构备受关注。同时，互联网的普及也为老年人提供了更多社会参与的机会与途径。因此，主题 1 聚焦于老年人的养老模式与社会变革之间的关系，包括养老方式的演进、社会对老年人的支持以及服务体系的变迁等方面。

主题 2 涉及关键词“留守老人”、“人际关系”、“追踪调查”，与社会关系紧密相关。留守老人通常指因子女外出务工、离婚等原因而独居的老年人，他们往往缺乏充分的社会支持和陪伴，缺乏社会支持与陪伴感成为老年人幸福感下降的重要原因。而偏远地区的老年人则可能面临更多资源匮乏、交通不便等问题，社会关系对其影响更为重要。因此，主题 2 聚焦于这些社会关系与偏远地区老年人之间的联系，探讨如何改善他们的社会关系以提升生活质量与幸福感。

主题 3 涉及关键词“幸福感”、“国内外”、“纳西族”的跨文化视角。这些词汇表明研究超越单一文化或地域范围，从更广泛、跨文化的角度探讨老年人的幸福感。以纳西族为例，作为特定文化群体，其价值观、生活方式及幸福感定义可能与其他文化有异。同时，融合国内外研究可提供更全面、多元的视角理解老年人的幸福感。

主题 4 涉及关键词“健康”、“数字化”、“老龄化”，重点研究数字化技术在老龄化社会中对健康的影响。随着数字化技术的普及与发展，其在老年人健康管理、医疗服务及健康信息获取等方面发挥愈加重要的作用。例如，远程医疗、智能健康监测设备等可为老年人提供更便捷、高效的医疗服务，有助于提高他们的健康状况和生活质量。同时，老年人还存在数字素养不足的问题，对健康服务获取形成阻碍，进而影响主观幸福感。该主题突显了数字化技术对老年人健康的重要性，以及其在老龄化社会中的潜在作用。

社会支持在所有主题中都占据了重要地位, 这表明了老年人幸福感与社会支持之间密切的关系。社会支持是指个体主观上感受到的帮助和支持, 虽然可能不一定反映客观事实, 但它可以作为精神支柱, 提升个体的主观幸福感。此外, 健康状况在各主题中也得到了充分的关注, 显示了老年人主观幸福感与健康之间密切相关的普遍趋势。家庭养老、机构养老与社区养老模式的选择往往受社会支持网络结构的制约, 农村地区因子女外出、代际支持削弱而更多依赖外部照护资源。社会支持水平直接影响孤独感、应对方式与心理健康状况, 是决定老年人主观幸福感的重要中介变量。健康状况差的老年人更倾向选择专业照护型养老模式, 对医疗可及性与机构服务质量敏感度更高。互联网使用能力不仅影响健康服务的获取, 也决定老年人与外部社会的连接程度, 在各类养老模式与社会支持中扮演关键调节因子。每个主题都深入分析了不同因素对老年人主观幸福感的影响, 说明幸福感是一个受多种因素综合影响的复杂问题。

3.3. 主题热点趋势

计算四个主题在 2019~2023 近 5 年每个年限的主题热度, 将时间作为横轴, 主题热度作为纵轴, 绘制了主题热度趋势变化图见图 5。通过这种可视化方法, 可以清晰地观察不同年份的热点主题以及这些主题的热度变化情况, 从而更好地把握研究领域的动态变化。

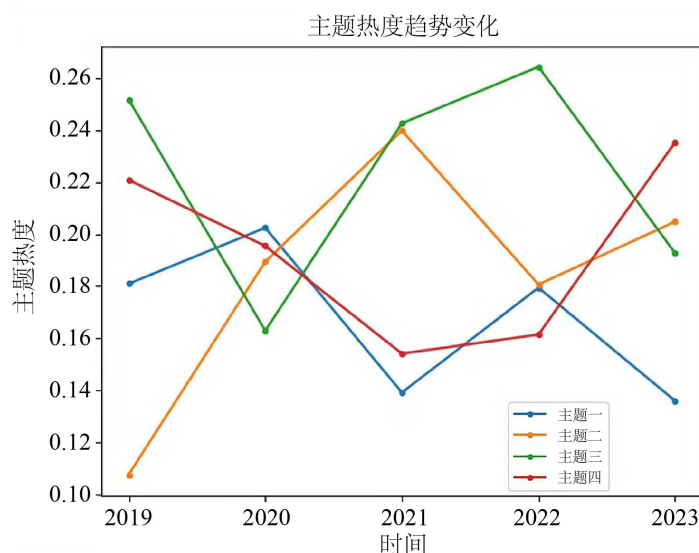


Figure 5. Trend of topic popularity from 2019 to 2024

图 5. 2019~2024 年的主题热度趋势变化图

在当前学术研究中, 养老模式与社会演变的关系这个主题的主题热度在近 5 年似乎呈下降趋势。这一现象可能归因于多重因素, 原因可能是目前我国农村养老模式研究已经取得比较丰富的成果, 其次也在一定程度上反映出农村养老模式受学界重视的程度相对下降, 这种下降趋势在一定程度上反映了学术界对农村养老模式的研究重心的转移。偏远地区老年人主题总体呈上升趋势, 说明仍是研究热点, 要保障和改善偏远地区老人生活问题。当代社会随着数字化技术的持续发展, 数字化老龄化成为备受关注的焦点。同时跨文化视角下老年人主观幸福感的研究也备受重视。

3.4. 主题在时间维度上的动态变化分析

通常情况下, 主题在时间轴上先呈现出逐渐增加的趋势, 并最终趋于稳定状态。拐点是函数凹面变

化的点, 意味着斜率从增加转变为减少。在拐点附近, 主题正处于快速增长的阶段, 该时间段主题为研究的热点, 随后进入平稳期成为核心成熟主题。

为了分析主题发表论文的动态变化, 首先需要绘制主题相关论文的数量随时间变化的散点图, 根据散点图进行曲线拟合同时获得重要时间点, 包括首次出现时间和拐点时间。以年份为 x 轴, 论文篇数为 y 轴绘制的散点图见图 6, 通过曲线拟合可以直观地观察主题发展趋势。利用一元三次函数拟合曲线, 若拟合曲线 $f(x)$ 的三次导数不为 0, 则判断该点为拐点, 并在图像中标记出拐点的位置。

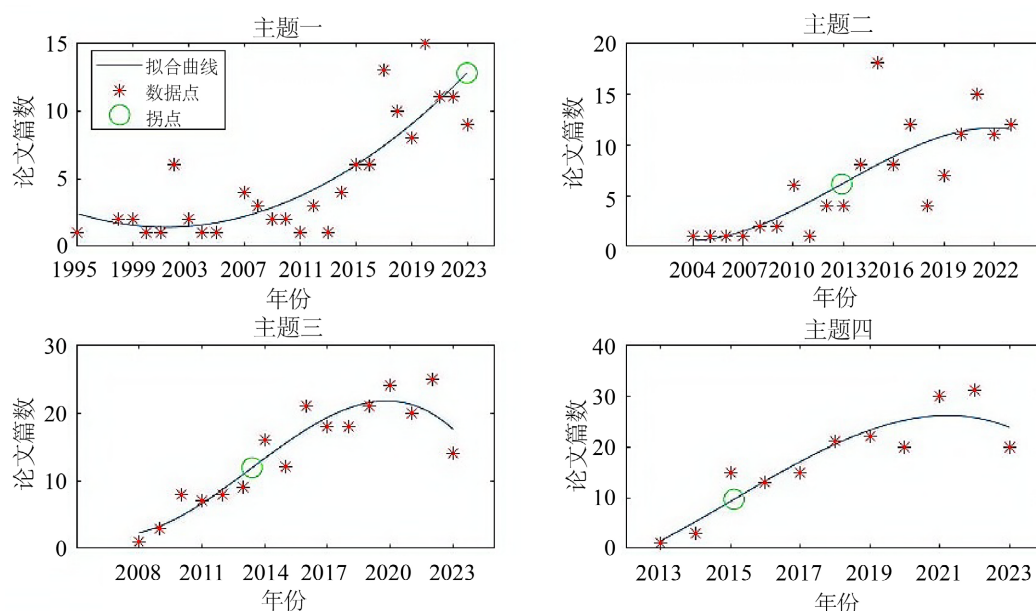


Figure 6. Key time points for different topics

图 6. 不同主题的关键时间点

在时间维度上, 主题 1 首次出现的时间最早, 发文点为 1995 年, 主题 2 的首次发文点为 2004 年, 主题 3 的首次发文点为 2008 年。而主题 4 的首次发文点为 2013 年, 相对于其他主题而言, 其出现较为晚期, 代表了当时的新兴主题。通过观察图表可知, 主题 2 和主题 3 的拐点大致出现在 2013 年左右, 表明在此时间段内这些主题已经成为具有影响力的研究方向, 并且目前已经达到了稳定状态。与主题 2 和主题 3 相比, 主题 4 的拐点则在 2015 年后出现, 结合其首次发文时间为 2013 年, 可见主题 4 的发展速度较快, 并且年均发文量较大, 备受研究者的重视。主题 1 的拐点出现在 2023 年, 这在时间序列中较晚, 考虑到其首次出现时间为 1995 年, 而且该主题涉及养老模式, 这表明在社会不断演化的过程中, 养老模式也在不断演变并对老年人的主观幸福感产生影响。

3.5. 主题演化分析

本文以年度发文量和文献增长率对检索年限进行时间切片, 研究范畴聚焦于老年人主观幸福感领域, 将相关文献划分为三个时间窗口: 第 1 阶段为萌芽期(1995~2011)总体发文量较为稀少, 第 2 阶段为成长期(2012~2017)发文量呈显著增加趋势, 第 3 阶段为成熟期(2018~2023)发文量在一定水平上波动, 显示出研究领域的稳定与成熟。通过对这三个阶段的观察, 可以更清晰地了解老年人主观幸福感研究领域的发展历程和趋势。

根据困惑度计算结果和 pyLDAvis 模块的可视化输出, 选择各个阶段的主题数目, 以此建立 LDA 模

型。首先采用 Python 中的 Gensim 模块中的 Word2vec 训练预处理后语料库的词向量, 然后用训练好的 Word2vec 模型生成各时间窗口对应主题的词向量, 再运用余弦相似度计算相邻时间窗口的主题相似度。为了确保主题间演化关系的准确性, 剔除无关的主题联系, 本文将相似度阈值设置为 0.3。即过滤各时间窗口主题之间低于阈值的主题, 以便更准确地挖掘各个时间窗口主题之间的关系。图中的同一列表示相同时间段的主题, 每条连线表示相邻时间窗口之间的主题关联, 线条的粗细反映了主题间的相似度, 得到老年人主观幸福感领域 1980~2023 年三个阶段的主题路径演化图见图 7。

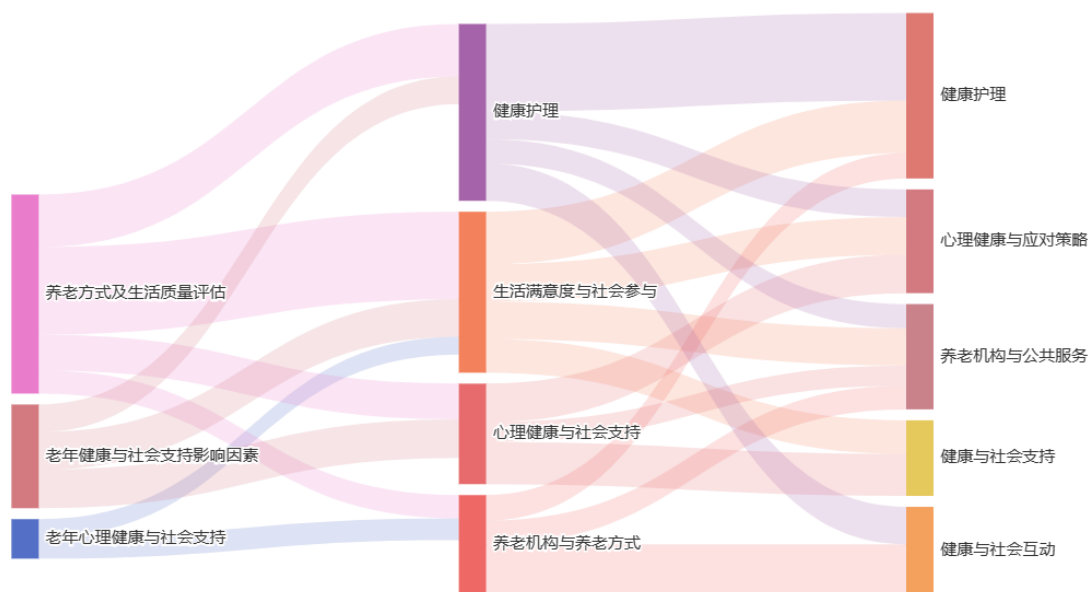


Figure 7. Thematic evolution roadmap in the field of subjective well-being of the elderly

图 7. 老年人主观幸福感领域主题演化路径图

从 1995 年到 2011 年, “养老方式及生活质量评估”, “老年健康与社会支持影响因素”以及“老年心理健康与社会支持”这三个主题逐渐细化出多条路径, 这些路径相互交叉和融合。2011 年之后, “养老方式及生活质量评估”逐步演化为“健康护理”, “生活满意度与社会参与”以及“心理健康与社会支持”这三类新主题; 2017 年到 2023 年之后, 这些主题进一步演化为“健康护理”, “心理健康与应对政策”, “养老机构与公共服务”, “健康与社会支持”以及“健康与社会互助”这五条主题路径。可以看出, 这种演化呈现出持续性和逻辑性。

从整体上看, 在 1980~2023 年期间, 主题间的交叉融合不断发展, 相关主题数量逐渐增加, 研究的主题逐渐多样化。由演化路径图可知, 大多数主题合并通常伴随新主题的出现。主题的合并代表着主题相关的知识体系成长, 是科学研究的焦点。主题的分化一般是核心成熟主题分化出多个分支主题, 表明主题的发展趋于稳定。

从研究热点看, 一些研究(如健康、社会支持)在不同的发展时期一直保持着良好的发展趋势。随着时间的推移, 研究主题在知识领域中的位置可能会发生变化。以心理健康这一主题为例, 在初始阶段出现时是一个普通但潜在的边缘成熟主题, 虽然具备较高的新颖性, 但尚未引起广泛关注。在受到科研人员的研究重视后, 在第二和第三阶段成为核心成熟主题, 这也证明了演化路径图能够直观地展示主题发展过程。

演化路径中表现出来的大多都是主题的新生、增长和融合现象, 对于主题消亡的情况相对较少。这

在一定程度上表明了领域发展的多样化。

4. 结论与展望

本研究基于主题模型对中国知网数据库中关于老年人主观幸福感的文章进行了全面的主题挖掘和主题演化分析。通过一致性计算结果和 pyLDAvis 可视化工具确定了最佳的主题数目, 进行主题挖掘并提取了各主题下的关键词, 结合研究文献对主题进行了深入分析, 得到“养老模式与社会演化的关系”、“社会关系与偏远地区老年人”、“跨文化视角下的老年人幸福感”和“健康与数字老龄化”四个主题。然后计算了各主题在不同年份的热度, 分析主题热度的变化趋势。发现随着当前数字化技术的不断进步和普及, 数字化老龄化已逐渐成为一个备受关注的重要议题, 尤其是在面对老龄化社会挑战的背景下, 数字化老龄化的研究和应用受到了越来越多的关注。并从时间维度上深入分析了各主题的关键时间点。最后, 将文章分为三个阶段, 并利用 LDA2vec 模型对各阶段文章进行了主题提取和演化分析, 以可视化方式展现了演化趋势。通过对 1980~2023 年期间各研究主题的演化分析, 整体上主题间的交叉与融合逐步发展, 相关主题数量不断增加, 且研究领域呈现出逐渐多样化的趋势。大多数主题的合并通常伴随新主题的出现, 反映了科学研究焦点的演变。在研究热点方面, “心理健康”与“社会支持”在不同时间段内持续保持较好的发展趋势。而随着时间的推移, 心理健康主题在初期作为一个普通但潜在的边缘性主题, 但随着科研人员逐步重视该领域的研究, 心理健康逐步从边缘主题发展为核心成熟主题。

本文更深入地考虑了时间维度中拐点时间和首次出现时间, 以更准确地分析各主题在学科领域演化过程中的作用, 并揭示学科主题的演化趋势, 从而全面了解学科主题的状态整体情况。然而该方法存在一些不足之处, 即缺乏相关领域专家的知识背景对结果进行主题概括和总结。未来研究可进一步考虑纳入更多类型的文献数据, 例如政策报告、学术专著等, 以更全面地反映当前研究的知识基础和发展脉络。同时, 可开展更细致的时间序列分析, 以捕捉主题演化过程中的动态细节和潜在拐点, 增强研究的解释力与前瞻性。

基金项目

国家自然科学基金(72071130)。

参考文献

- [1] 吴捷. 老年人社会支持、孤独感与主观幸福感的关系[J]. 心理科学, 2008(4): 984-986+1004.
- [2] 陈艳艳, 王鹏飞, 魏翔. 休闲模式与老年人主观幸福感: 作用机制及实证检验[J]. 统计与决策, 2023, 39(24): 69-73.
- [3] 黄菡, 王晓光, 何静, 等. 基于矩阵相似度的主题演化路径判别研究[J]. 情报学报, 2023, 42(11): 1265-1275.
- [4] 程秀峰, 邹晶晶, 叶光辉, 等. 融合 Word2Vec 的半积累引用共词网络的领域主题演化研究[J]. 情报学报, 2023, 42(7): 801-815.
- [5] 颜端武, 苏琼, 张馨月. 基于时序主题关联演化的科学领域前沿探测研究[J]. 情报理论与实践, 2019, 42(7): 144-150.
- [6] 阮光册, 夏磊. 基于 Doc2Vec 的期刊论文热点选题识别[J]. 情报理论与实践, 2019, 42(4): 107-111+106.
- [7] 王莉亚. 基于离群数据的主题演化规律分析[J]. 情报杂志, 2013, 32(6): 59-63.
- [8] Hu, Z., Zhang, X. and Xiong, H. (2024) Two-Stage Attention Network for Fault Diagnosis and Retrieval of Fault Logs. *Expert Systems with Applications*, 249, Article ID: 123365. <https://doi.org/10.1016/j.eswa.2024.123365>
- [9] 施文, 渠玉杰, 蒋国银. 基于随机 Kriging 的汽车品牌质量序贯主题比较研究[J]. 中国管理科学, 2023, 31(11): 114-127.
- [10] Chen, L., Tian, F.F., Fu, Y. and Kahana, E. (2024) Dementia Knowledge in Chinese Newspapers (2005-2020): A Topic Modeling Analysis. *Journal of Aging & Social Policy*, 1-17. <https://doi.org/10.1080/08959420.2024.2348966>

-
- [11] Pais, N., Ravishanker, N., Rajasekaran, S., Weinstock, G. and Tran, D. (2024) Randomized Feature Selection Based Semi-Supervised Latent Dirichlet Allocation for Microbiome Analysis. *Scientific Reports*, **14**, Article No. 8855. <https://doi.org/10.1038/s41598-024-59682-4>
- [12] 陶成煦, 吴江, 税典程, 于洋. 取向与趋向: 数据要素交易政策主题挖掘与演化研究[J]. 情报理论与实践, 2024, 47(6): 39-48.
- [13] 许海云, 张慧玲, 武华维, 等. 新兴研究主题在演化路径上的关键时间点研究[J]. 图书情报工作, 2021, 65(8): 51-64.
- [14] Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003) Latent Dirichlet Allocation. *Machine Learning Research Archive*, **3**, 993-1022.
- [15] 李秀霞, 程结晶, 韩霞. 发文趋势与引文趋势融合的学科研究主题优先级排序——以我国情报学学科主题为例[J]. 图书情报工作, 2019, 63(11): 88-95.
- [16] Onan, A. (2020) Sentiment Analysis on Product Reviews Based on Weighted Word Embeddings and Deep Neural Networks. *Concurrency and Computation: Practice and Experience*, **33**, e5909. <https://doi.org/10.1002/cpe.5909>
- [17] 唐明, 朱磊, 邹显春. 基于 Word2Vec 的一种文档向量表示[J]. 计算机科学, 2016, 43(6): 214-217+269.
- [18] 陈磊, 李俊. 基于词向量的文本特征选择方法研究[J]. 小型微型计算机系统, 2018, 39(5): 991-994.
- [19] Curiskis, S.A., Drake, B., Osborn, T.R. and Kennedy, P.J. (2020) An Evaluation of Document Clustering and Topic Modelling in Two Online Social Networks: Twitter and Reddit. *Information Processing & Management*, **57**, Article ID: 102034. <https://doi.org/10.1016/j.ipm.2019.04.002>
- [20] 祁瑞华, 付豪. “一带一路”智库报告主题挖掘与演化研究[J]. 智库理论与实践, 2022, 7(5): 11-19.
- [21] 关鹏, 王曰芬. 科技情报分析中 LDA 主题模型最优主题数确定方法研究[J]. 现代图书情报技术, 2016(9): 42-50.
- [22] 张绍武, 邵华, 林鸿飞, 等. 基于主题模型的新疆暴恐舆情分析[J]. 中文信息学报, 2018, 32(5): 105-113.
- [23] Palla, G., Barabási, A. and Vicsek, T. (2007) Quantifying Social Group Evolution. *Nature*, **446**, 664-667. <https://doi.org/10.1038/nature05670>