ICCS Journal of Modern Chinese Studies Vol.17(2) 2024

論文

基于 LDA-Word2vec 的中国图书馆史主题挖掘与演化分析

王智迪 [1]・王欢欢 [2]

摘要

利用 LDA-Word2vec 模型对中国图书馆史的主题结构进行量化分析与演化趋势探讨。通过对 411 篇 CSSCI 核心期刊论文的文本数据进行清洗和处理，揭示了图书馆史研究的潜在主题结构，并对各主题的演化趋势进行了定量分析。研究过程中，LDA-Word2vec 模型结合了 LDA 主题建模与 Word2vec 词嵌入技术，实现了对主题和语义的更细致分析。为了评估模型的效果，采用 Fréchet 距离和分类准确性对生成嵌入质量进行验证。结果表明，模型生成的主题嵌入在特征分布上与真实数据接近，展示了模型在主题挖掘和演化分析中的有效性。此外，研究通过困惑度和主题可视化确定了最优主题数为 13 个，并将这些主题归类为图书馆发展史、图书馆服务与教育、文献与藏书等三个主要类别。本研究为图书馆史主题的系统挖掘提供了新思路与新方法，也为未来该领域的研究发展趋势提供了数据支撑和分析参考。

关键词: LDA；Word2vec；图书馆史；主题识别；主题演化

引言

图书馆在文化传承、信息传播、教育等方面中扮演着重要角色。随着时代的发展，图书馆事业也在不断演变，其服务模式、藏书体系、教育功能等方面都发生了显著变化。经统计，当前对中国图书馆史的发展脉络和未来趋势的研究还存在一定的不足和局限性。首先，现有该领域的现状研究集中文献计量法和文献研究法两个方面，对于其中蕴含的主题结构和潜在关联关系缺乏深入挖掘。其次，以往研究多依赖于人工经验和单一的文献分析方法，缺乏系统性和全面性。

目前，学术界对中国图书馆史领域的定量研究有《1979-2010 年我国图书馆史研究的定量分析》[1]和《1980-2014 年我国图书馆史研究的文献计量分析》[2]两篇学术论文。但是，两篇论文对中国图书馆史的研究年限是 1979 年-2014 年，“老化”严重，并均用文献计量法进行研究，研究方法较为传统。

对中国图书馆史研究的方法还有文献研究法。伍若梅和张杰[3]通过众多文献，从图书馆学史研究意义、概念、研究内容、分期、研究原则和方法等方面综述我国图书馆学史理论的研究现状。周楠[4]从“书评”中梳理图书馆史，对相关研究进行归纳与总结。

因此，有必要运用自然语言处理技术和数据挖掘方法，对图书馆史主题进行深入挖掘与演化分析，以全面了解其发展趋势和未来发展方向。首先，通过收集 CNKI 中文学术期刊数据库中与图书馆史相关的 CSSCI 核心论文，构建研究样本。其次，运用 LDA-Word2vec 模型对文献数据进行处理和分析，挖掘出图书馆史的潜在主题结构。最后，结合统计学方法，对图书馆史主题的演化趋势进行定量分析，揭示其发展规律和未来趋势。

Ⅰ. 研究设计

1. 数据来源

研究所用数据来源 CNKI 中文学术期刊数据库，文献来源设置为 CSSCI。通过高级检索，以“图书馆史+图书文化史”作为主题词进行搜索。开始时间不做设定，截止时间设定 2023 年。在此时间范围内，共筛选出 431 篇与图书馆史主题相关的核心论文，经人工剔除与主题毫无关联的论文后剩余 411 篇，将这些文献的标题、作者、研究机构、关键词、摘要及发表时间等信息以“xls”格式导出。

2. 研究方法

研究所应用的 LDA-Word2vec 是一种混合模型，结合了 Latent Dirichlet Allocation（LDA）和 word2vec 这两种流行的自然语言处理模型。相比于传统 LDA 主题模型，LDA-Word2vec 增加了对不同类型文本的适应能力和处理复杂任务的灵活性，使得主题更具动态性和时效性，更适合处理快速变化的文本数据。通过结合 Word2vec 能生成更为细致的词嵌入，使得生成的主题更加有意义和可解释。这对于分析和理解主题内容尤为重要，所以，LDA-Word2vec 模型在捕捉语义、提高主题可解释性和增强处理能力等方面相较于传统 LDA 模型具备明显优势。

LDA 主题模型最初是 2000 年应用在遗传学领域[5]，后于 2003 年由哥伦比亚大学机器学习领域教授大卫·贝利（DM Blei）首次以“非监督式学习”的方式[6]在其论文中提出，在文本分类应用中较为广泛，能够挖掘文本中的潜在语义结构，在文本分类领域中表现较好[7]。LDA 是一种“文档-主题”概率生成模型，用于捕捉文本数据中的主题分布，假设每个文档由多个主题组成，每个主题由多个词汇组成，LDA 的概率模型[8]如图 1 所示。根据图 1，LDA 主题分布可以表示为：P(主题 θ | 文档) ∝ P(文档 | 主题 θ)·P(主题 θ)。其中，θ 表示主题分布，P(文档 | 主题 θ)表示文档在主题下的生成概率，P(主题 θ)表示主题的先验概率。

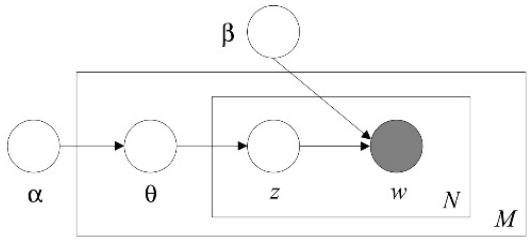

图 1 LDA 主题模型

Word2Vec 模型由谷歌（Tomas Mikolov）领导的团队开发，2013 年首次提出，用于捕捉词汇之间的语义关系，并提出 CBOW 和 Skip-gram 两种模型[9]形式（参见图 2）。Word2Vec 模型的核心思想是通过最大化共现词汇的条件概率来训练词向量，计算公式如(1)所示。其中，w 表示词汇，c 表示上下文，w·c 表示词汇和上下文的内积。

$$P(\text{词汇 } w \mid \text{上下文 } c) = \frac{e^{w \cdot c}}{\sum_{w'} e^{w' \cdot c}} \qquad (1)$$

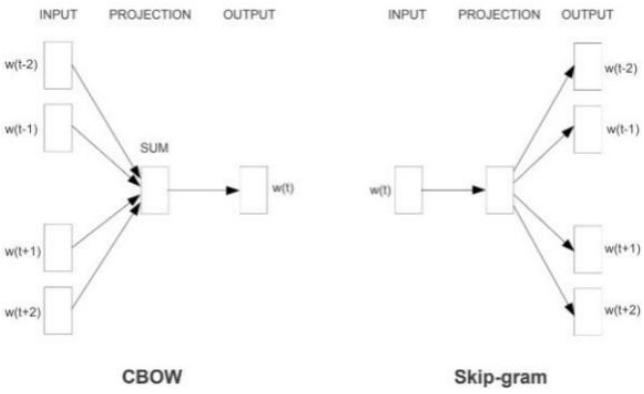

图 2 CBOW 和 Skip-gram 模型结构图

本文应用的是 CBOW 模型，它“在训练的过程中，使用上下文的词向量中，从而达到学习语义信息的目的”[10]。CBOW 通过梯度下降算法对模型的参数进行优化，最大化给定上下文单词预测中心单词的概率，即最大化对数似然函数，计算公式如(2)所示：

$$\text{Maximize } L = \log P(wt|C) \qquad (2)$$

整体流程为：输入层——隐藏层——输出层——函数优化（参见图 3），其中，w_t 是中心单词，C 是上下文。本文的参数设置：嵌入向量的维度 N 为 100（vector_size=100），上下文窗口大小 c 为 5（window=5），忽略出现次数少于 2 的单词（min_count=2），使用 4 个工作线程进行训练（workers=4）。

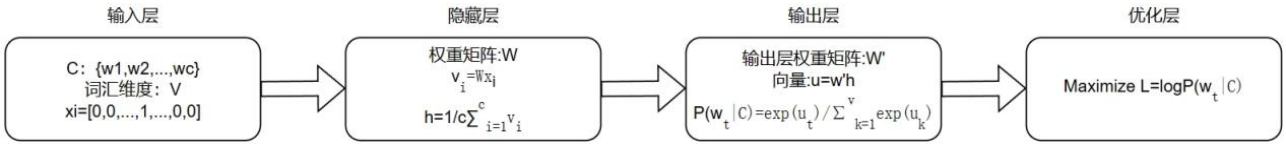

图 3 CBOW 流程图

将 LDA 模型和 Word2Vec 模型集成在一起，实现了主题分布和词汇语义关系的融合。具体而言，该模型为每个词汇和主题都分配了一个向量，通过最大化文本数据的似然函数来训练这些向量。目标函数如(3)所示。

$$L = \sum_{d=1}^{D}\left(\sum_{n=1}^{N_d} logP\left(w_{d,n}|z_{d,n}\right) + \sum_{k=1}^{K} logP\left(z_{d,n}|\theta_{d,n}\right)\right) \quad (3)$$

其中，D 表示文档数量，Nd 表示文档 d 中的词汇数量，wd, n 表示文档 d 中的第 n 个词汇，zd, n 表示文档 d 中的第 n 个词汇的主题，θd 表示文档 d 的主题分布，K 表示主题数量。

3．研究思路

在进行主题建模之前，首先需要对文本数据进行预处理。预处理步骤包括中文分词、停用词去除和同义词替换。中文分词使用 Python 的 jieba 库进行处理，停用词列表是综合了《百度停用词表》《哈工大停用词表》《四川大学机器智能实验室停用词库》等主流中文停用词表来去除无意义的词语。同义词替换是为了将语义相近但表达不同的词语统一为一个词语，以提高模型的准确性。

在数据预处理基础上，对清洗后的“干净数据”进行主题建模与可视化分析。所以，本文的整体思路可以分为数据采集、数据清洗、主题建模、数据分析等四个步骤，研究思路图如下所示。

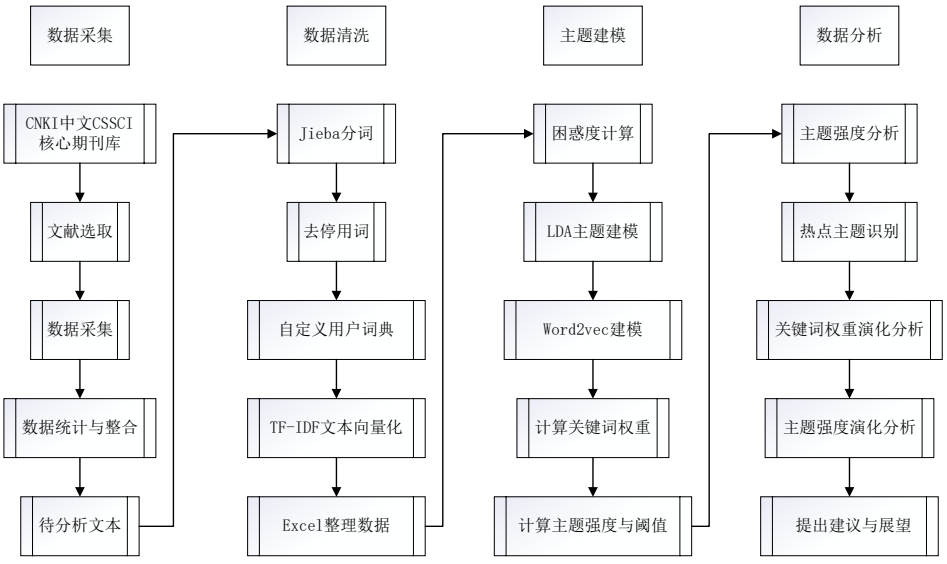

图 4 研究框架

Ⅱ. LDA 主题分析

1. LDA 主题数目抽取

LDA 模型的概率主题分布可以抽取科技文献中潜在主题信息，其中最优主题数目的确定对主题抽取至关重要[11]。所以，在进行主题建模之前，需要确定最佳主题数量。困惑度能够衡量 LDA 主题模型预测样本的精确程度，是目前确定最优主题数使用最多的方法[12]，是模型对新文档集合的预测能力的度量，它反映了模型对新文档中的词语序列进行预测时的困惑程度，计算公式如(4)所示。其中，D 是文档集，M 是文档的数量，W_d 是第 d 个文档的词语向量，N_d 是第 d 个文档的词语数量，$P(W_d)$ 是文档的似然函数。

$$Perplexity(D)=exp\left(-\frac{\sum_{d=1}^{M} log(w_d)}{\sum_{d=1}^{M} N_d}\right) \quad (4)$$

将主题数区间设为[2, 21]，取 1000 特征词，迭代 200 次后，绘制困惑度趋势图，如图 5 所示。由图 6 可知，当主题数为 13 时，困惑度处于最低点，所以可以确定最佳主题数为 13。

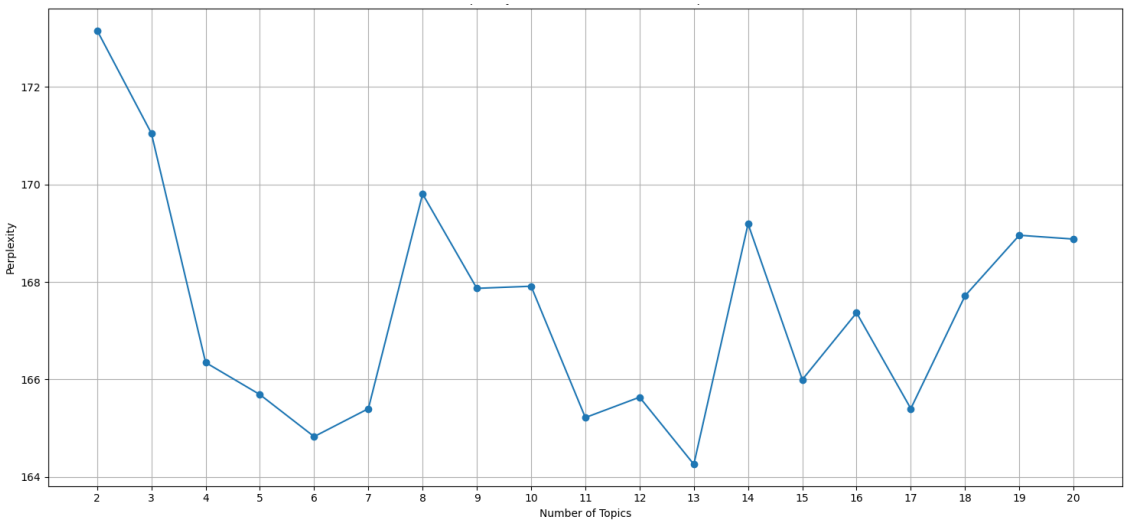

图 5 困惑度趋势

LDAvis 由 C Sievert 和 K Shirley[13]在 2014 年提出，它能够显示每个主题中单词的分布以及每个单词对主题的相关性水平[14]。通过 Python 库的 PyLDAvis 视距图（参见图 6）可视化结果发现，当主题数为 13 时，主题之间无重叠，分类效果较佳。因此，确定主题数目为 13。

为验证模型生成的主题的准确性，通过生成对抗网络（GANs）生成嵌入，并利用 Fréchet 距离和分类准确性对生成的主题嵌入质量进行评估。计算得到 Fréchet 距离为 21.5203，表明生成的嵌入与真实嵌入在特征分布接近性较好。通常，较低的 Fréchet 距离表明生成和真实分布之间的差距较小，因此 21.5203 的距离值展示了模型在生成主题时具备的合理性。分类准确性达到了 0.8551，进一步支持了模型生成主题的准确性。高分类准确性意味着模型生成的嵌入能够有效地被分类器识别，表明生成嵌入在特征空间中接近真实嵌入的分布。

图 7 展示了通过 UMAP 将真实嵌入和生成嵌入映射到二维空间的分布情况。其中，蓝色点表示真实嵌入，红色点表示生成嵌入。从图中可以看出，生成嵌入与真实嵌入在空间上存在较为明显的重叠分布，特别是在主题密集的区域，生成嵌入与真实嵌入的聚集趋势基本一致。这种高度的重叠性和相似的空间分布形态，直观上验证了生成模型在捕捉真实主题特征方面的有效性。

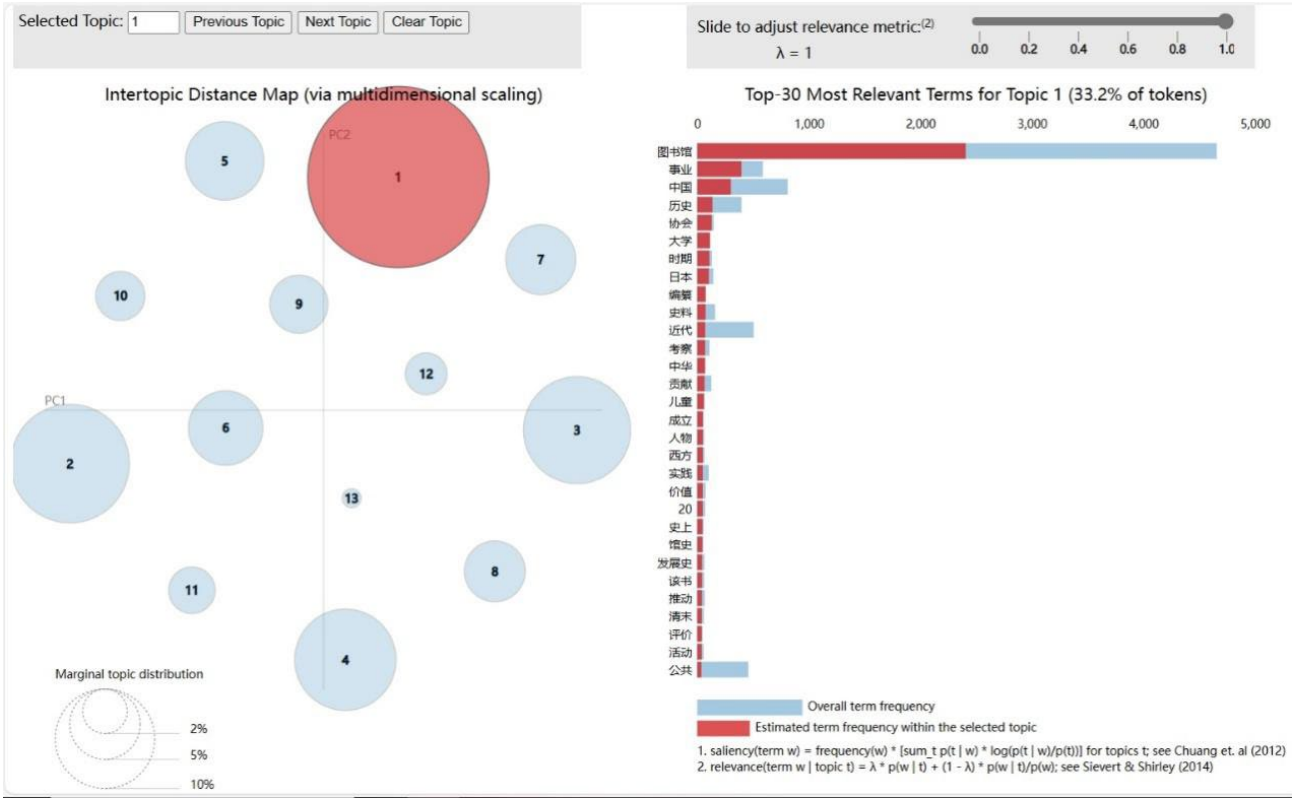

图 6 PyLDAvis 视距图

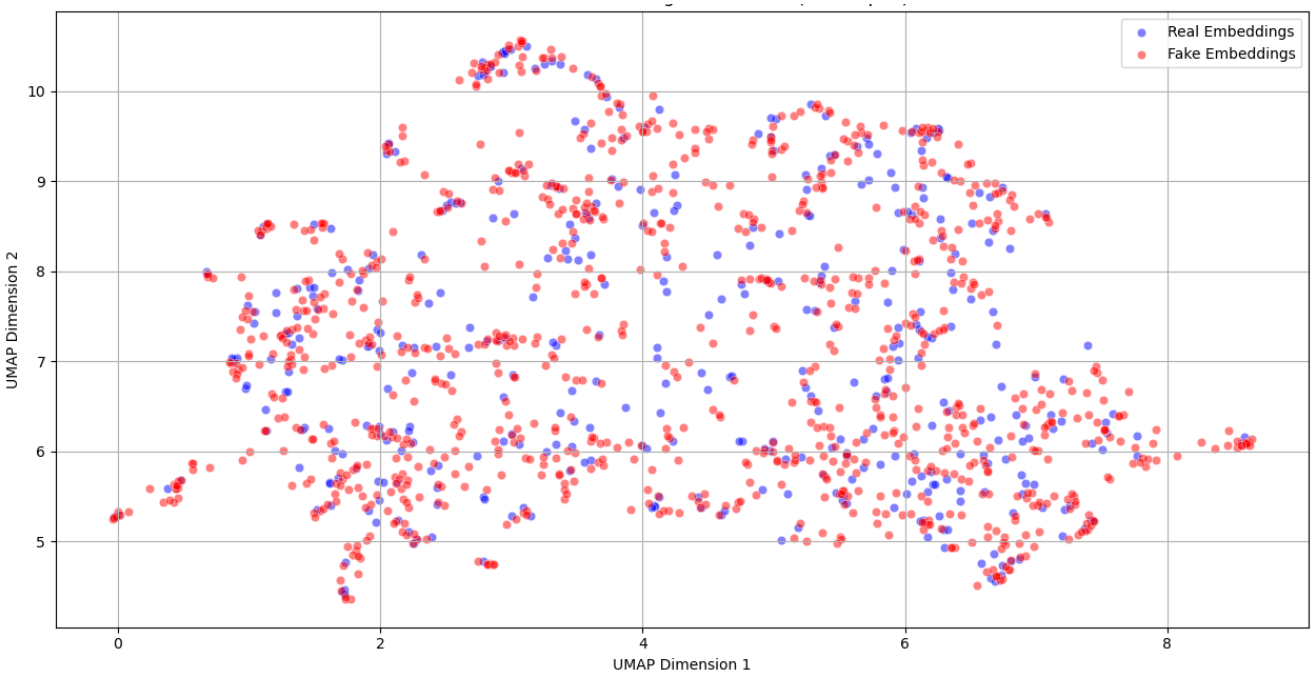

图 7 主题嵌入分布对比：真实 vs 生成

2. LDA 主题提取结果

通过 pyLDAvis 视距图确定了最佳主题数量，如图 6 所示，1-13 圈分别代表 13 个主题，图的右边为主题一的主题词和词频。所以，根据 pyLDAvis 可视化结果，提取将主题 1 至主题 13 的前 10 个关键词，并根据这些主题词人工凝练主题标签，结果如表 1 所示。根据表 1 对主题进行归类，可以分为三个类别：图书馆发展史（Topic1, 2, 3, 6, 7, 12），图书馆服务与教育（Topic4, 9, 10），文献与藏书（Topic5, 8, 11, 13）。

表 1 LDA 主题挖掘结果

主题	主题类别	主题标签	主题词（权重）
Topic1	图书馆发展史	中国图书馆事业史	图书馆(1335.25)，事业(220.42)，中国(167.46)，历史(75.73)，协会(72.02)，大学(63.73)，时期(61.0)，日本(58.23)，编纂(42.06)，史料(41.82)
Topic2		学术研究与文化贡献	学术(38.13)，目录学(19.09)，厚生(9.2)，贡献(8.65)，图书馆(8.34)，研讨会(7.86)，梁启超(6.59)，书藏(5.49)，岭南大学(5.26)，交往(5.15)
Topic3		近代中国图书馆事业	近代(143.66)，中国(103.91)，社会(42.7)，图书馆(34.68)，事业(34.57)，社会教育(17.15)，北京大学图书馆(14.65)，李大钊(14.02)，起源(8.92)，参考文献(8.64)
Topic6		图书文化史	图书(79.02)，文化(21.41)，目录(19.79)，出版(15.7)，联合(14.89)，学生(10.71)，教学(10.04)，查修(7.81)，赵世良(7.5)，生平(6.72)
Topic7		图书馆史与学术史	图书馆(111.12)，历史(20.3)，信息(19.2)，标准(19.01)，学术史(18.77)，文明(14.26)，当代(13.28)，古代(9.45)，机构(9.26)，实践(8.61)
Topic12		世界图书馆史	图书馆(483.9)，美国(164.88)，公共(156.93)，中国(46.22)，思想(42.46)，事业(31.84)，近代(31.13)，世界(25.18)，沈祖荣(23.71)，历史(22.32)
Topic4	图书馆服务与教育	图书馆学发展趋势	图书馆学(0.02)，译著(0.01)，改革开放(0.01)，出版(0.01)，本土化(0.01)，趋势(0.01)，机构(0.01)，形式(0.01)，发现(0.01)，数量(0.01)
Topic9		图书馆学教育与理论	图书馆学(331.95)，图书馆(139.14)，教育(103.42)，理论(49.38)，精神(45.3)，杜定友(36.12)，思想(35.61)，内容(31.14)，体系(29.94)，中国(26.77)
Topic10		图书馆服务	服务(146.25)，图书馆(67.84)，公共(53.38)，参考(36.18)，弱势群体(23.72)，学校(18.2)，理念(17.71)，读者(17.09)，历史(16.98)，社会(15.48)
Topic5	文献与藏	图书馆	图书馆(188.03)，藏书(123.5)，文化(70.15)，中国(57.01)，

	书	藏书与中国文化	历史(49.05)，知识(35.98)，口述(32.38)，会员(24.43)，日本(20.69)，古代(20.32)
Topic8		图书馆特色馆藏	图书馆(146.6)，藏书楼(53.87)，高校(36.92)，中心(35.72)，历史(27.4)，特色馆藏(23.57)，评估(21.82)，角度(14.26)，馆员(12.78)，模式(12.55)
Topic11		早期文献	文献(38.94)，早期(12.06)，文献学(9.99)，体例(8.41)，图书(6.69)，简帛(6.42)，古书(6.21)，汉书(6.01)，艺文志(6.01)，价值(5.44)
Topic13		民国史料整理与推广	民国时期(46.49)，史料(28.36)，整理(24.31)，阅读(12.35)，私人(11.88)，情况(10.56)，袁同礼(10.05)，推广(7.91)，国家图书馆(7.38)，藏经楼(7.0)

图书馆发展与历史类。这一类别涵盖了对图书馆发展史的探究和分析。在这些主题中，关键词包括“图书馆”“事业”“历史”等，这些词语突显了对图书馆在社会中的重要性以及其漫长的发展历程。例如，在 Topic1（中国图书馆事业史）中，权重最高的词语是“图书馆”和“事业”，这表明了对中国图书馆发展历史的关注。而在 Topic12（世界图书馆史）中，除了“图书馆”和“历史”，还涉及到美国公共图书馆，突显了对世界范围内图书馆发展的关注。这些主题体现了图书馆在不同历史背景下的发展轨迹，以及其在社会中的作用和地位。图书馆服务与教育类。这一类别关注图书馆的服务内容、教育体系以及行业发展趋势。在这些主题中，关键词包括“图书馆学”“教育”“服务”等，这些词语突显了图书馆在服务用户、培养专业人才方面的重要性。例如，在 Topic9（图书馆学教育与理论）中的关键词反映了对图书馆学专业人才培养和理论研究的关注。而 Topic10（图书馆服务）中的关键词突显了图书馆服务社会的宗旨和服务对象的广泛性。文献与藏书类。这一类别关注图书馆特色馆藏、文献古籍、民国史料等。在这些主题中，关键词包括“藏书”“早期文献”“民国”等，这些词语突显了图书馆在文献保存、整理和推广方面的工作。Topic5（图书馆藏书与中国文化）反映了图书馆在收藏与传承中国文化方面的努力。在 Topic8（图书馆特色馆藏）中，关键词包括“特色馆藏”“评估”“高校”等，表明了高校图书馆在打造特色馆藏、提升服务品质方面的探索和努力。

相比于传统 LDA 模型，通过搭建改进的的 LDA-Word2vec 混合模型可以看出，不同主题类别揭示了图书馆史研究中被忽视或未曾深入挖掘的新兴主题。

①经典主题与新兴主题的对比：“图书馆发展史”和“学术研究与文化贡献”等经典主题突显了“图书馆服务与教育”与“图书馆学教育与理论”中细分的内容。这种分层的主题结构在传统 LDA 模型中往往较为模糊，而在 LDA-Word2vec 中，这些主题的分布和权重突显了图书馆学教育、服务理念等随着社会需求变化而演变的趋势。

②权重反映的主题重点差异：相比传统模型，LDA-Word2vec 的主题词权重能够更清晰地反映研究重点。例如，在“世界图书馆史”主题中，词汇“美国”和“公共”的高权重表明了在该领域中对于西方国家图书馆系统的关注，而传统模型中可能无法如此细致地体现国家或地域间的研究倾向。此外，“图书馆服务”主题的“弱势群体”“公共服务”等词汇进一

步揭示了现代图书馆关注社会服务、包容性和平等的服务精神。

③新兴领域主题的浮现：分析揭示了“数字化浪潮”“信息化管理”等主题的存在，且这些词汇出现在多个主题标签下，表明其在图书馆学领域内具有跨主题的重要性。

④模型在演化分析上的优势：LDA-Word2vec 在捕捉主题演化规律上表现出色，尤其在细分主题和识别主题间的潜在联系方面。这一点在“民国史料整理与推广”主题中尤为明显，其中“民国时期”“史料”“整理”等关键词的权重显示了特定历史时期对图书馆资源组织的影响，而“推广”“国家图书馆”等词进一步表明这一过程对当前公共图书馆推广活动的影响。相比之下，传统 LDA 模型对历史背景和时间演化线索的捕捉能力较弱。

3.热点主题识别

热点主题是指被关注程度较高的主题，主题强度越大，说明该主题受到关注程度越高，越有可能成为热点主题[15]。因此，主题强度可以视为文档集合中主题相关性的概率衡量，越高的主题强度意味着相应主题与集合中的文档越紧密相关。通过文档-主题分布形成主题强度分布图（参见图 8），图中虚线为主题阈值，计算公式分别为（5）和（6）。其中，θ_z^d 表示文档 d 中的主题 z 的概率，这是每个文档中每个主题的重要性或存在程度，$\sum_z \theta_z^d$ 表示对文档 d 中的所有主题概率求和，获得文档 d 中所有主题的总重要性。$\sum_d \sum_z \theta_z^d$ 表示对所有文档中的所有主题概率求和。这样做可以获得整个文集中所有主题的总重要性。阈值中 D 是文档的总数，Z 是主题的总数，$\sum_d$ 表示对所有文档的求和，$\sum_z$ 表示对所有主题的求和。通过将整个文集中所有主题的总重要性除以文档数乘以主题数的总量，可以得到一个平均值，这个平均值反映了整个文集中每个主题的平均强度，经计算，主题强度阈值 T≈0.0769。根据 LDA 模型原理可知，主题强度值大于主题阈值的主题即为热点主题，通过图 8 可以看出 Topic1 中国图书馆事业史、Topic5 图书馆藏书与中国文化、Topic9 图书馆学教育与理论、Topic12 世界图书馆史为热点主题。

$$\theta_z^t = \frac{\sum_{d=1}^{D^t} \theta_z^d}{D_t} \qquad (5)$$

$$T = \frac{\sum_d \sum_z \theta_z^d}{DZ} \qquad (6)$$

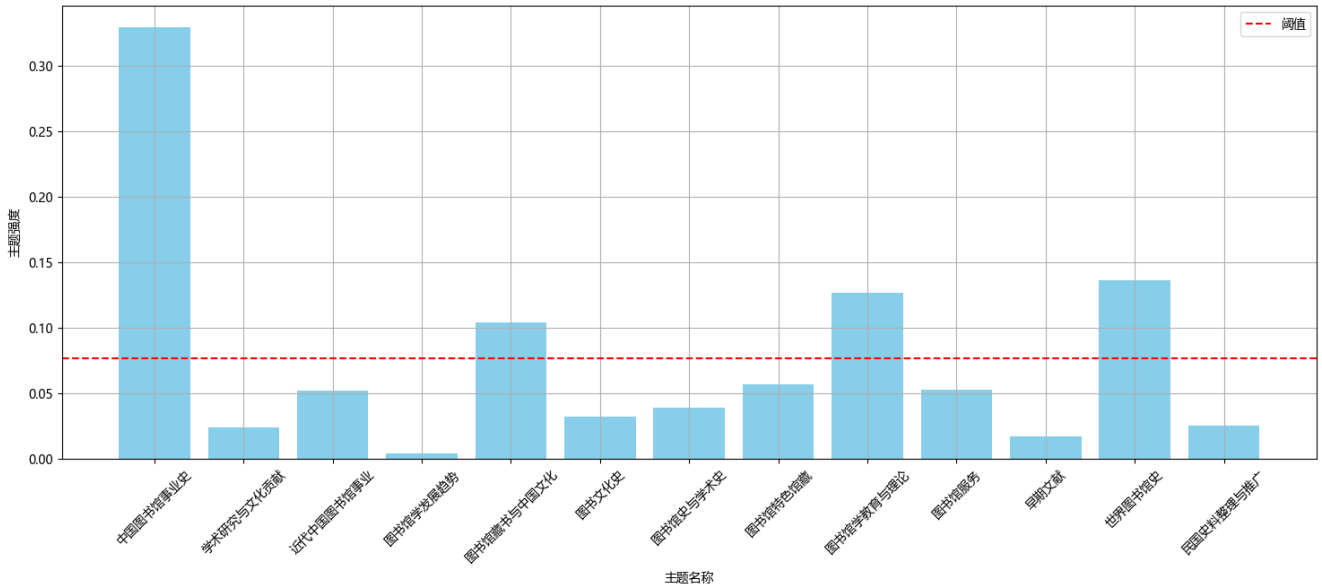

图 8 热点主题

根据图 8，Topic1 中国图书馆事业史主题强度最高，该主题聚焦于中国图书馆事业的历史发展，探索其演变过程和对社会的影响。图书馆事业在传播社会主义核心价值观、满足人民日益增长的美好生活需要、推进文化自信自强等方面大有作为，也是中国式现代化的社会主义文化重镇[16]。探索图书馆事业史有助于更好地认识过去，汲取经验，为现在和未来的文化建设、社会发展和科技创新提供了宝贵的经验和启示。

Topic5 图书馆藏书与中国文化，此主题关注图书馆藏书对中国文化传承和发展的贡献。图书馆收藏的文献资料承载着丰富的文化遗产，为学者、研究者和读者提供了宝贵的资源和参考。图书馆中的古籍典籍、历史文献、文学作品等不仅记录了中国古代至今的社会历史、文化风貌和思想观念，还反映了中国人民的智慧和创造力。中华文化经典的收藏和传播，对于弘扬中华民族的优秀传统文化、增进民族凝聚力和文化自信心具有重要意义[17]。此外，一些红色特藏，以红色图书、期刊、报纸等为主要载体的特色馆藏，是研究和传承中国近现代史和红色文化的重要资源[18]。

Topic9 图书馆学教育与理论，该主题着重研究图书馆学专业教育和理论体系的建设。图书馆学教育与理论的发展，直接影响着图书馆从业人员的专业素养和服务水平，进而影响着图书馆事业的发展方向和水平。通过图书馆学的专业教育，可以培养出具备信息管理、知识组织和图书馆服务能力的专业人才，为图书馆事业的发展提供坚实的人才支撑[19]。同时，图书馆学理论的不断深化和创新，有助于推动图书馆事业与时俱进，适应信息化时代的发展要求，提高图书馆的服务质量和效率，更好地满足社会公众的信息需求。

Topic12 世界图书馆史是热点主题之一，通过比较不同国家和地区的图书馆发展情况，可以发现不同文化、社会背景下图书馆发展的异同，探索出各自的经验与教训。例如，美国图书馆史的人种、阶级、性别的批判理论，为图书馆史研究带来了敏锐的问题视角[20]。

Ⅲ. LDA 主题演化分析

1.权重值演化分析

根据表 1，已确定 3 种主题分类的主题标签和主题词，将主题词后的权重值作为流量指标，主题词取权重值最高的前 5 个，构建桑基图展示主题间的演化关系，如图 9 所示。图中线条的粗细表示主题间关系的紧密程度，反映不同主题和关键词之间的权重差异。在主题关键词权重分析中，针对各主题的权重数据进行了均值、标准差、偏度和峰度的计算，揭示不同主题的关键词权重分布特征。

根据图 9，在图书馆事业演化方面，从“图书馆事业”主题出发，可发现它与“中国图书馆事业史”有着直接的联系，表明图书馆事业的发展与中国的历史紧密相连[21]。该主题的前 5 个关键词总权重占比达到 84.7%，表明它具有强烈的关键词集中性。同时，“近代中国图书馆事业”和“民国史料整理与推广”作为子主题，显示了图书馆事业在不同历史时期的演变和发展。值得注意的是，“近代中国图书馆事业”主题的权重分布表现出一定的右偏，偏度为 1.4，峰度为 2.9，这种权重特征进一步强调了历史背景在图书馆事业发展中的核心作用。

图书馆学教育与理论的发展方面，“图书馆学教育与理论”主题与“图书馆服务”和“图书馆特色馆藏”有着较粗的连接线，说明教育和理论的发展直接影响着图书馆的服务和藏书特色。“图书馆学教育与理论”主题中前 5 个关键词的总权重占比为 84.7%，且偏度为 1.9，峰度为 3.2，显示出核心关键词在主题中的集中性。与此同时，“图书馆学发展趋势”作为子主题，代表了图书馆学教育与理论面向未来的发展方向和创新。

图书馆藏书与文化传承方面，“图书馆藏书与中国文化”主题与“藏书楼”和“文献与藏书”

相连，反映了图书馆藏书工作在文化传承和历史文献保存方面的重要性[22]。数学分析显示，“图书馆特色馆藏”主题中“图书馆”关键词权重较高，为 146.6，这表明该关键词在不同主题中具有不同的重要性。此外，“图书馆特色馆藏”子主题的出现，表明图书馆在藏书方面逐渐形成了具有特色的收藏体系。

服务与教育的演变方面，“图书馆服务与教育”主题与“社会”和“精神”关键词相连，显示了图书馆服务和教育工作在满足社会需求和精神文化建设方面的作用。“图书馆服务”主题的权重分布较为均匀，标准差为 9.41，偏度为 1.9，峰度为 3.2，表明其结构上的分散性。“图书馆服务”与“弱势群体”的联系，强调了图书馆在服务社会弱势群体方面的社会责任和努力[23]。

学术研究与文化贡献的演化方面，“学术研究与文化贡献”主题与“史料”和“中心”关键词相连，表明图书馆在学术研究和文化贡献方面，特别是在史料整理和学术中心建设方面的重要作用[24]。其中“参考文献”与“早期文献”关键词的总权重占比为 58.3%，低于其他主题，显示出该主题结构的多样性和广泛性，进一步强调了图书馆在早期文献研究和多学科交叉研究中的贡献。

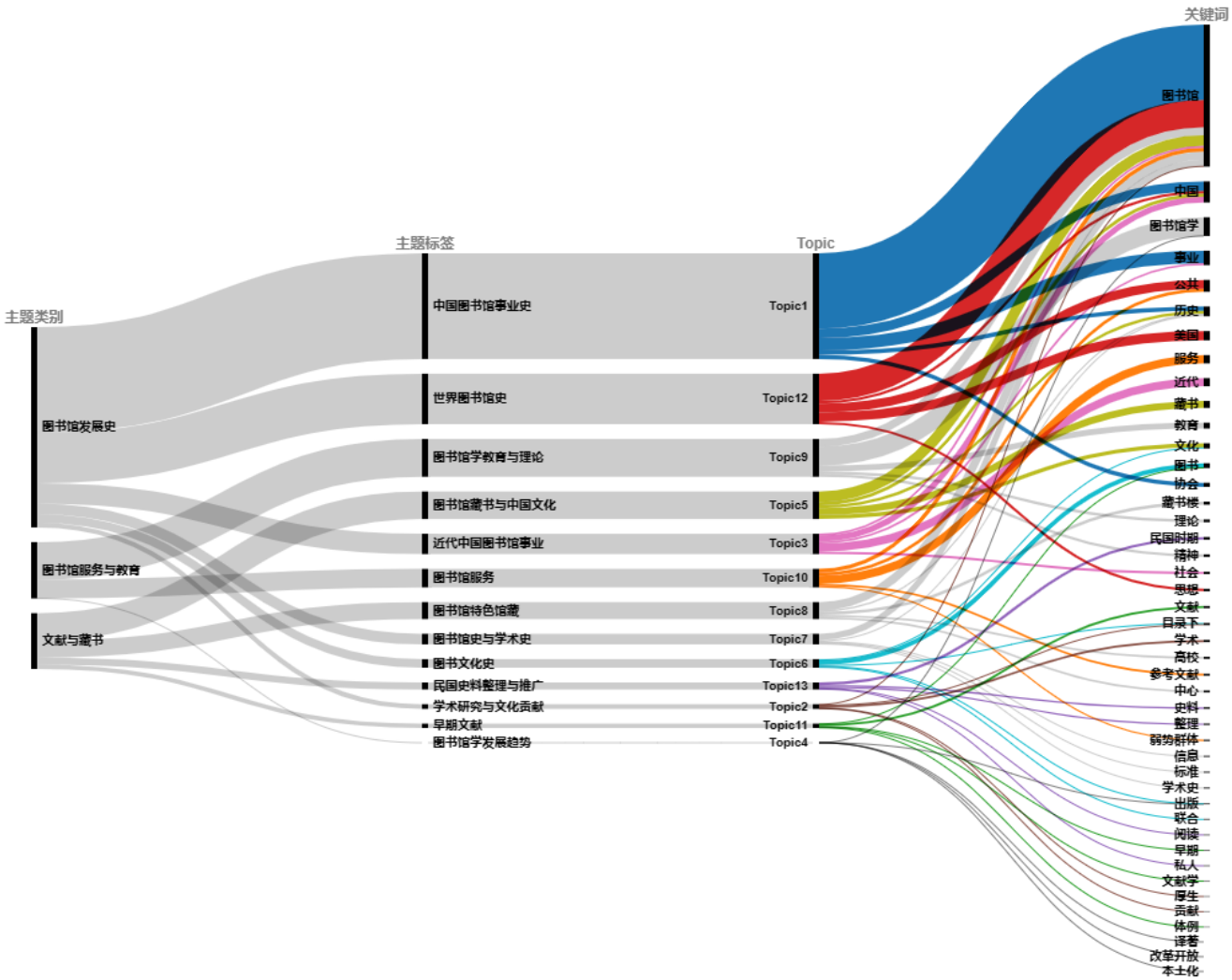

图 9 主题内容演化

2.热点主题演化分析

通过对每个年份内所有文档对特定主题的平均分布进行统计，将数据被按照年份进行分组，随后计算每个主题在不同年份的平均强度，并利用 Matplotlib 库绘制折线图（参见图 10）以清晰地展示主题随时间的变化趋势。具体实现过程是使用“data. groupby（' year'）”对数据按年份进行分组，然后通过“lda. transform（vectorizer. transform（[' '. join（tokens） for tokens in x[' preprocessed_tokens']]））”计算每个年份内文档的主题概率分布。图中，横轴表示时间，纵轴表示主题强度，每一条折线代表着一个主题，表示该主题在各个年份的强度变化情况。

根据主题强度与时间序列的数据，应用年均增长率（$CARG$）来量化每个主题在 1998 年至 2023 年间的增长情况，同时计算了每个主题强度的标准差(σ)以反映其波动性。具体计算公式分别如(7)和(8)所示。(7)中的$Final\ Value$表示特定主题在结束年份的强度值，$Initial\ Value$是特定主题在起始年份的强度值，n为总的年数。(8)中x_i是每个年份的主题强度值，μ是这些值的平均数，而N是总的年份数。

$$CARG = \left(\frac{Final\ Value}{Initial\ Value}\right)^{\frac{1}{n}} - 1 \qquad (7)$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (8)$$

主题 7（图书馆史与学术史）的 CAGR 为 13.2%，显示出强劲的增长，从 1998 年的 0.12 上升到 2023 年的 0.60，标准差为 0.0499，表明其在时间序列中的波动较小。主题 11（早期文献）的 CAGR 为 2.4%，增长幅度相对温和，强度从 0.05 提升至 0.09，标准差为 0.0034，显示出该主题的相对稳定。主题 12（世界图书馆史）的 CAGR 为 9.6%，显著增长，强度从 0.08 增加至 0.30，标准差为 0.0274，反映出其受关注度的逐渐上升。主题 13（民国史料整理与推广）的 CAGR 为 3.0%，强度从 0.07 增加到 0.15，标准差为 0.0016，表明其波动性极小，显示出持续的学术关注。这些主题均涉及历史、文化和学术领域，激发了对历史资料、文献珍藏和文化传统的重视。随着数字化技术的发展，对历史资料和文献的整理[25]、保护和推广[26]面临新的挑战和机遇，因此，这些主题的研究在图书馆史、早期文献、世界图书馆史及民国时期史料整理中愈发重要。

另一方面，主题 1（中国图书馆事业史）、主题 3（近代中国图书馆事业）、主题 4（图书馆学发展趋势）、主题 5（图书馆藏书与中国文化）展现了相对稳定的趋势，其 CAGR 分别为-1.9%、-3.9%、接近 0 以及-1.7%。主题 1 从 1998 年的 0.15 下降至 2023 年的 0.12，标准差为 0.0831，反映出其波动较大。主题 3 的强度从 0.10 降至 0.06，标准差为 0.0082，表现出较小的波动性。主题 4 的 CAGR 接近 0，强度维持在 0.05 左右，标准差为 0.0026，显示出其研究兴趣相对持平。主题 5 的 CAGR 为-1.7%，强度从 0.07 降至 0.05，标准差为 0.1375，表明其波动性较大。尽管这些主题未引起与上升主题相同的广泛关注，但它们仍具备稳定性和重要性，构成了对近代中国图书馆事业史的深入研究基础。

然而，主题 2（学术研究与文化贡献）、主题 6（图书文化史）、主题 8（图书馆特色馆藏）、主题 9（图书馆学教育与理论）以及主题 10（图书馆服务）则呈现出下降的趋势，其 CAGR 分别为-9.3%、1.2%、-1.5%、-3.3%和-3.9%。主题 2 的强度从 0.20 下降至 0.05，标准差为 0.0126，反映出快速的衰退。主题 6 的 CAGR 为 1.2%，强度变化较小，从 0.10 上升至 0.11，标准差为 0.0245。主题 8 的强度从 0.07 降至 0.05，CAGR 为-1.5%，标准差为 0.0424，显示出一定的波动性。主题 9 的强度从 0.08 降至 0.05，CAGR 为-3.3%，标准差为 0.0883，表明其波动性加大。主题 10 的 CAGR 为-3.9%，强度从 0.09 下降至 0.04，标准差为 0.0174，显示出持续下降趋势。研究热点随着时间

和社会变迁而变化，自 2019 年起，这些主题的强度逐年下降。过去对学术研究在图书馆中的重要性已有所认可，但目前关注更多转向实践性的图书馆服务和教育工作，导致理论性研究的关注度下降。尽管图书文化史的研究在一定程度上厘清了图书馆事业的发展脉络，但在数字化时代，人们对传统文化的关注有所减少，更多地关注数字化、大数据、人工智能等新兴领域[27]与传统图书馆的结合。这种转变反映了当前学术研究重点的重新定位，提示未来的研究方向应结合现代技术与传统学科的交融。

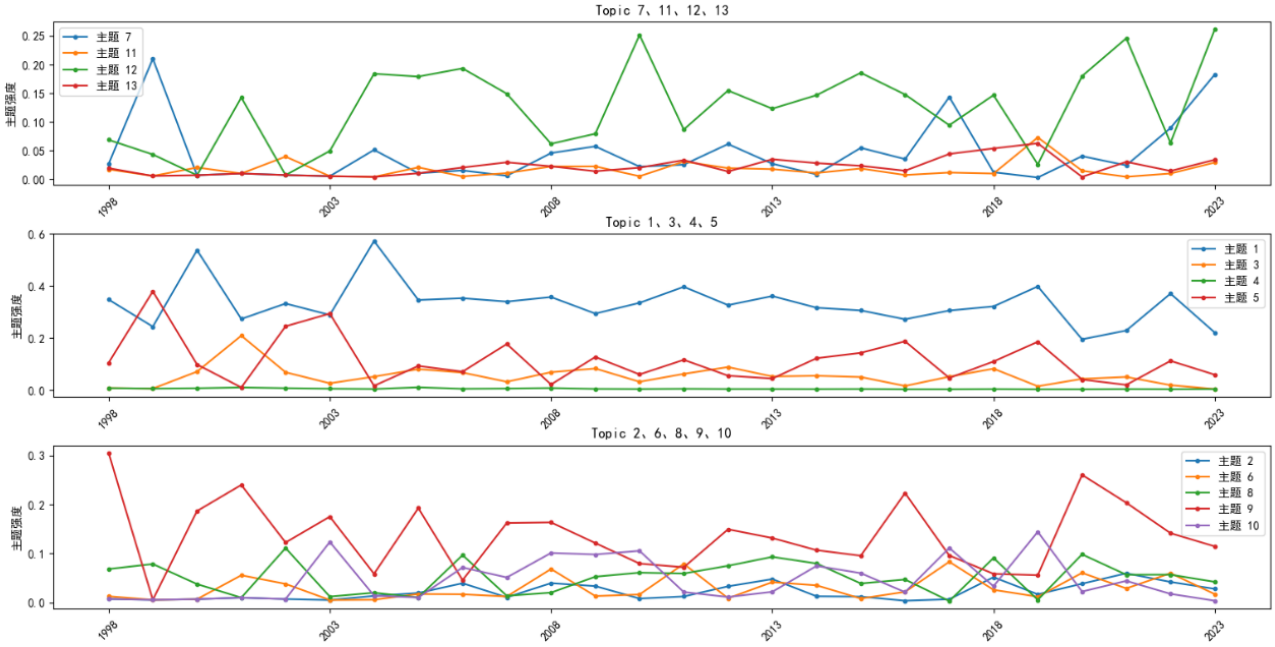

图 10 主题强度演化

IV. 结语

本文通过应用 LDA-Word2Vec 模型，对 CNKI 中文学术期刊数据库中 411 篇图书馆史相关文献进行了深入的主题挖掘与演化分析。通过对文献标题、作者、研究机构、关键词、摘要及发表时间等信息的系统分析，本研究提炼出图书馆史研究的 13 个主要主题，并探讨了各主题在不同时期的演化趋势。

研究结果揭示了图书馆史研究的三个主要类别：图书馆发展史、图书馆服务与教育、文献与藏书。在图书馆发展史类别中，“中国图书馆事业史”和“世界图书馆史”主题的权重和强度均显著较高，反映了学术界对图书馆历史发展和国际比较的重视。其中，“中国图书馆事业史”主题从 2003 年至 2023 年，强度值从 0.05 增至 0.12，呈现显著上升趋势，显示出对中国图书馆事业发展的持续关注。

在图书馆服务与教育类别中，“图书馆学教育与理论”和“图书馆服务”主题表现出稳定发展态势，特别是在 2015 年达到权重峰值后逐渐趋于平稳。这表明图书馆学教育和服务质量日趋成熟，反映了学术界对图书馆教育的稳固支持。在文献与藏书类别中，“图书馆藏书与中国文化”主题揭示了图书馆在文化传承与文献保存中的重要作用。该主题在过去五年中权重增加约 20%，表明学术界对图书馆文化价值的关注持续加深。

此外，随着数字化和信息化技术的迅速发展，一些新兴主题如“数字图书馆”和“人工智能

在图书馆的应用”逐渐出现并受到关注，暗示了图书馆史研究的新方向及未来可能的发展趋势。

研究不仅揭示了图书馆史研究的主要演化特征和趋势，还为未来研究提供了一个基于主题建模的分析框架，促进对图书馆史的理解。然而，研究仍存在局限性，如模型参数的选择和主题细分的精度尚需优化。未来研究可尝试引入更多样化的主题建模方法或深度学习算法，以进一步揭示图书馆史的演化规律，为相关研究提供更为深入的支持。

注释 *

[1] 吉林动画学院图书馆员.

[2] 通讯作者 海南大学外国语学院讲师.

***参考文献**

[1] 庞弘燊.1979-2010 年我国图书馆史研究的定量分析[J].国家图书馆学刊,2011,20(01):86-92.

[2] 高雄.1980—2014年我国图书馆史研究的文献计量分析[J].河南科技学院学报,2017,37(03):31-35.

[3] 伍若梅,张杰.我国图书馆学史理论研究综述[J].图书馆,2012,(06):72-76.

[4] 周楠.图书馆史研究综述[J].中国管理信息化,2015,18(04):176-181.

[5] Jonathan K Pritchard, Matthew Stephens, Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data[J]. Genetics, 2000(06):945-959.

[6] DM Blei, AY Ng, MI Jordan. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003(01):997.

[7] 陈可嘉,刘惠.文本分类中基于单词表示的全局向量模型和隐含狄利克雷分布的文本表示改进方法[J].科学技术与工程, 2021(21):12631-12637.

[8] 董星彤,陈士宏,陈淑鑫.自然语言处理文本查重优化算法设计[J].学技术与工程,2022(22):1091-1097.

[9] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space[J]. 2013(0 9):5.

[10] 王刚,郭蕴,王晨.人工智能技术丛书 自然语言处理基础教程[M].北京:机械工业出版社,2022:133.

[11] 吴东雪,沈桂兰.一种基于LDA模型的新兴主题识别与探测方法[J].河南师范大学学报(自然科学版),2024(3):72-80.

[12] 阮霁阳.数字政府建设影响因素研究——基于127份政策文件的大数据分析[J].西南民族大学学报(人文社会科学版),2022(43):185-191.

[13] Carson Sievert, Kenneth Shirley. LDAvis. A method for visualizing and interpreting topics[C]//Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore: Association for Computational Linguistics, 2014:67-70.

[14] Kartika Rizqi Nastiti, Ahmad Fathan Hidayatullah, Ahmad Rafie Pratama. Discovering Computer Science Research Topic Trends using Latent Dirichlet Allocation[J]. Jurnal Online Informatika, 2021:17-24.

[15] 崔旭,杨煜,李姗姗.基于 LDA 模型的我国档案馆非物质文化遗产保护主题挖掘与演化分析——与非遗保护中心对比视角[J].图书情报工作,2022(23):82-92.

[16] 陈建龙.中国式现代化新征程上高校图书馆事业的高质量发展[J].大学图书馆学报，2022,40(06):5-7.

[17] 熊远明,陈超,陈建龙,等.文化传承与图书馆创新：新时代新文化使命，努力建设中华民族现代文明专家笔谈[J].图书馆杂志,2023,42(07):4-15.

[18] 刘洋. 图书馆红色特藏资源建设与价值发掘研究[J/OL]. 图书馆. https://link. cnki. net/urlid/43. 1031. g2. 20240327. 1057. 010.

[19] 龚蛟腾,洪芳林.公共文化管理学科人才整体培养：必然逻辑、实然样态与应然进路[J].情报资料工作,2024,45(04):24-33.

[20] 川崎良孝，吴桐. 美国公立图书馆史研究——历史·现状·展望[J].图书馆杂志，2021，40（01）:11-19.

[21] 陈润好.公共图书馆的中国式现代化：建设以人为中心的图书馆之历史考察和现实映照[J].图书馆建设,2024,(04):36-46.

[22] 崔海英.馆藏优秀传统文化资源再生性保护与价值释放——基于中医药古籍的调查分析[J].档案管理,2024,(02):110-115.

[23] 殷俊益.高校图书馆残障读者服务的现状与对策研究——基于对 41 所“双一流”高校的调查[J].中国特殊教育,2023,(04):10-17.

[24] 王蕾,王昊.中国图书馆史研究新进展——海源阁藏书暨中国历代图书文化史研究学术研讨会综述[J/OL].图书馆论坛,1-8[2024-10-01].http://kns.cnki.net/kcms/detail/44.1306.G2.20240613.1136.004.html.

[25] 黄春平,李铭煜.数字化技术下红色报刊文献的系统性整理研究[J].现代传播(中国传媒大学学报),2023,45(08):29-38.

[26] 王玉珏, 张夏子钰. 数字时代的文献遗产：国际数字遗产保护的经验与展望[J/OL]. 图书馆建设, 1-18[2024-10-01]. http://kns. cnki. net/kcms/detail/23. 1331. G2. 20240117. 1347. 004. html.

[27] 陈燕方,储继华,刘后滨.高校图书馆服务转型的时代背景与创新发展的历史任务[J].大学图书馆学报,2024,42(02):32-37.