小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0227
2026 年 5 月 第 5 期
Vol.47 No.5 2026

上下文自编码框架下的科研热点挖掘方法

王 睿[1,2,3,4],吕心诚[1],陆家豪[1],周永权[4]

[1](南京邮电大学 计算机学院、软件学院、网络空间安全学院,南京 210023)
[2](东南大学 东南大学新一代人工智能技术与交叉应用教育部重点实验室,南京 211189)
[3](智能互联系统安徽省实验室(合肥工业大学),合肥 230009)
[4](广西民族大学 广西混杂计算与集成电路设计分析重点实验室,南宁 530006)
E-mail:rui_wang@njupt.edu.cn

摘 要: 高效挖掘科研热点及其对应作者是学术研究领域的重要任务.针对传统作者主题模型忽略上下文语义、难以融合外部知识及缺乏背景主题建模的问题,本文提出了一种基于上下文的神经作者主题模型.该模型利用 Transformer 捕捉文本的上下文语义以提升主题推断准确性,将单词与作者的预训练嵌入引入解码过程并利用 vMF 分布对主题进行建模以提升主题质量,同时采用狄利克雷树分布作为先验以区分背景主题与热点主题.此外,本文提出两个量化研究热点与作者关联程度的指标.本文在构建的计算语言学、计算机视觉和数据挖掘 3 个数据集上进行实验,结果表明,本模型在主题一致性、多样性及作者-主题关联性指标上均优于对比方法,充分验证了其在科研热点挖掘上的优越性.

关 键 词: 科研热点挖掘;作者主题模型;von Mises-Fisher 分布;狄利克雷树分布
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2026)05-1089-10

Research Hotspot Mining Under a Contextual Autoencoding Framework

WANG Rui[1,2,3,4],LÜ Xincheng[1],LU Jiahao[1],ZHOU Yongquan[4]

[1](School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China)
[2](Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications,Southeast University,Ministry of Education,Nanjing 211189,China)
[3](Intelligent Interconnected Systems Laboratory of Anhui Province,Hefei University of Technology,Hefei 230009,China)
[4](Guangxi Key Laboratory of Hybrid Computation and IC Design Analysis,Guangxi Minzu University,Nanning 530006,China)

Abstract: Efficiently mining research hotspots and their corresponding authors is a critical task in academic research. To address the limitations of traditional author topic models, which often overlook contextual semantics, struggle to incorporate external knowledge, and fail to model background topics, this paper proposes a contextualized neural author topic model. The model utilizes Transformer to capture contextual semantics of text to improve the accuracy of topic inference, incorporates pre-trained word and author embeddings into the decoding process, and employs von Mises-Fisher distribution for topic modeling to improve topic quality. Meanwhile, it uses Dirichlet tree distribution as a prior to distinguish background topics from hotspot topics. Furthermore, the paper introduce two metrics to quantify the degree of association between research hotspots and authors. Experiments were conducted on three constructed datasets: Computational Linguistics, Computer Vision, and Data Mining. The results demonstrate that the model outperforms existing methods in topic coherence, diversity, and author-topic relevance, validating its superiority in mining research hotspots.

Keywords: research hotspot mining; author topic model; von Mises-Fisher distribution; dirichlet tree distribution

0 引 言

随着科学研究的发展,学术论文数量快速增长,不仅加大了研究者追踪领域动态的难度,也使得高效挖掘科研热点成为了一项重要任务.通过揭示文献中隐藏的主题及其与作者之间的关系,研究者可以更全面地了解某一领域的研究方向及代表性科研人员,这对科研管理与知识发现具有重要意义.

主题模型通过识别潜在主题结构为这一任务提供了新思路,然而现有方法仍然存在以下局限:1)传统主题模型[1]基于词袋(Bag-of-Words,BoW)[2]假设与概率图框架,忽略了词序和上下文信息,并且难以融入外部知识,限制了其在科研场景中的表现;2)神经主题模型(Neural Topic Models,NTMs)[3-5]利用神经网络的层次化特征提取能力,提高了主题推断的效率,但仍面临模式坍塌导致的主题多样性不足问题;3)现有模型未能融入作者信息,无法揭示科研热点与代表性科研人员之间的关联,限制了模型在科研热点分析中的应用

收稿日期:2025-06-09 收修改稿日期:2025-07-17 基金项目:国家自然科学基金青年基金项目(62102192)资助;中国博士后科学基金面上项目(2022M710071)资助;江苏省双创博士人才项目(JSSCBS20210530)资助;中央高校基本科研业务费专项资金项目(aiia-24-01,PA2025IISL0107)资助. 作者简介:**王 睿**,男,1990 年生,博士,讲师,研究方向为文本挖掘、自然语言处理、机器学习;**吕心诚**,男,2001 年生,硕士研究生,研究方向为自然语言处理;**陆家豪**,男,2001 年生,硕士研究生,研究方向为自然语言处理;**周永权**,男,1962 年生,博士,教授,研究方向为计算机软件与理论、计算智能及应用、神经网络.

价值.

针对现有主题模型缺乏作者信息的局限,研究者尝试通过作者-主题联合建模以更精准地挖掘科研热点并追踪学科前沿.基于概率图框架的方法[6]虽然可以同时建模作者与主题,但受限于词袋表示与复杂的推理过程,难以适应现代科研数据分析需求.近年来,有研究者尝试利用图神经网络建模作者-主题关系[7],但该方法在缺乏引用信息的场景下表现受限,且无法区分背景主题与热点主题.近期,有研究人员通过引入一个框架来整合作者信息[8],从而使主题模型能够对作者信息进行建模,不过该方法同样存在无法区分背景主题与热点主题的问题.综上所述,现有主题模型仍存在以下不足:1)未能利用文本的上下文信息,导致主题推断不准确;2)无法有效区分背景主题与热点主题,影响主题质量;3)评估体系不完善,对主题-作者关联度的评估多依赖主观分析,缺乏客观量化指标.

为解决以上问题,本文提出了一种基于上下文的神经作者主题模型(Contextualized Neural Author-Topic Model,CNATM),旨在提升作者-主题建模的语义捕捉能力和主题质量,并客观量化研究热点与作者的关联性.CNATM 首先利用 Transformer 生成文档嵌入,充分捕捉文档的上下文语义信息,从而克服了传统模型忽略词序和上下文的局限性.在解码阶段,CNATM 将主题建模为嵌入空间中的混合 von Mises-Fisher(vMF)分布,通过引入单词和作者的预训练嵌入作为外部语义知识,提升主题的语义一致性.此外,CNATM 利用狄利克雷树分布[9]作为主题分布的先验,从而在建模过程中将背景主题与任务主题区分开,以更准确地挖掘科研热点及对应的研究人员.

综上所述,本研究的主要贡献如下:

1)提出一种基于上下文的神经作者主题模型(Contextualized Neural Author-Topic Model,CNATM),该模型利用 vMF 建模主题并将单词、作者嵌入引入建模过程以提升主题质量.

2)CNATM 利用狄利克雷树分布[9]作为先验分布建模主题,避免背景词混入热点主题,提升热点挖掘的可解释性.

3)在计算语言学、计算机视觉、数据挖掘 3 个论文摘要数据集上的实验结果验证了 CNATM 的性能,实验结果表明,其在主题一致性、主题多样性及作者-主题关联性上均优于对比方法.

1 相关工作

1.1 传统的科研热点挖掘方法

科研热点挖掘旨在揭示学术领域的热门研究内容,帮助研究者理解不同领域的主导方向及其发展脉络.早期的热点挖掘方法主要基于文献计量学和引文分析,最典型的包括引文分析[10,11]、共引分析[12,13]和关键词分析[14,15].这些方法通过分析学术文献之间的引用关系、共同出现的关键词以及学者间的合作网络,揭示出学术领域的核心问题和研究热点.然而,这些方法在深入挖掘文献背后的语义信息方面存在明显限制,尤其是在处理大规模文献数据时.例如,引文分析和共引分析主要依赖于文献之间的引用关系,而忽略了文本内容的语义关联;相比之下,关键词分析虽然能够捕捉高频术语,但无法有效处理同义词和多义词问题.这些局限性促使研究者转向更先进的文本挖掘技术.

1.2 主题建模与作者主题建模研究进展

主题模型是文本挖掘中的一种重要方法,旨在从大规模语料库中提取潜在的主题结构.潜在狄利克雷分配[1](Latent Dirichlet Allocation,LDA)是其中最具代表性的模型,被广泛应用于学术文献分析以揭示研究热点,它将每篇文献看作是多个潜在主题的组合,每个主题通过一组单词概率分布来表示.然而,LDA 依赖于词袋模型,忽略了词序和上下文信息,难以捕捉复杂的语义关联,限制了其在复杂文本中的表现.

近年来,深度学习在自然语言处理领域的广泛应用推动了主题建模方法的革新,基于神经网络的主题建模方法[3-5]逐渐成为研究热点.这类方法通过神经网络的层次化特征提取能力,不仅克服了传统 LDA 模型在高维稀疏数据上的计算效率瓶颈,还能更精准地捕捉词汇间的非线性语义关联.相较于传统概率图模型,神经主题模型实现了对文本潜在语义结构的深层解析,提升了主题的连贯性.然而,这类模型仍面临着挑战:基于 VAE 的神经主题模型受 KL 散度约束影响,可能出现后验坍塌现象,即解码器倾向于依赖输入噪声而非潜在变量,因而生成同质化的主题,降低了主题多样性;同时,传统主题模型及一些神经主题模型主要依赖词共现信息与词频统计,缺乏对领域知识库、语义本体等外部先验信息的融合,导致生成的主题在跨学科场景中容易混淆专业术语的语义边界.

为解决外部知识缺失问题,研究者引入了预训练词向量(如 Word2Vec[16],GloVe[17],JoSE[18]等)为词汇提供更加丰富的语义表示,引入这些词向量表示的主题模型可以生成语义上更相关的主题,从而弥补传统模型在复杂语义建模方面的不足.基于预训练模型的神经主题模型[19,20]也成为一种新的研究方向,通过 Transformer 得到的词向量可以更好地捕捉文档中的上下文信息,从而提高主题质量.近两年,大语言模型(Large Language Model,LLM)的兴起为主题建模提供了新的方向,研究者通过交互式提示引导大语言模型识别复杂语境下的主题语义结构[21],从而改善传统主题模型在捕捉复杂语义模式和生成高质量主题上的不足.

传统主题模型通常将文本视为独立于作者身份的语料集合,但在科研热点挖掘中,识别特定主题对应的作者及其研究方向对追踪学科前沿至关重要.为此,研究者开始探索作者与主题的联合建模方法,相较于主题建模的快速发展,两者联合建模的方法发展较为缓慢,相关研究相对较少.Rosen-Zvi 等人提出的作者主题模型(Author-Topic Model,ATM)[6]首次将作者信息引入主题建模过程,通过假设作者在每个主题上的偏好分布来实现作者和主题的关联建模,提升了对文献内容的理解和分析能力.但 ATM 基于 LDA 的概率图框架,仍然存在推理复杂度较高等问题,难以适应大规模数据集上的分析需求.近年提出的 VGATM 模型[7]利用图神经网络,通过捕捉文献间的引用关系和作者之间的合著关系来增强主题模型的表达能力.然而,VGATM 需依赖显式的引用与合著关系构建图结构,在数据稀疏或关系缺失的场景下,模型会因为无法有效构建文献或作者的邻接矩阵而失效.近期,Nagda 等人提出的 FANToM 框架通过整合作者信息提升了主题模型

的作者-主题对齐能力[8]，该框架不依赖于引用关系，能够高效的对作者信息进行建模，然而其在区分背景主题与热点主题方面仍存在不足. 此外，现有作者主题建模方法普遍缺乏量化指标来客观衡量主题与作者的关联程度，评价多为主观分析，通常仅展示模型抽取的主题词和作者，这也限制了其在科研热点挖掘中的应用.

1.3　vMF 分布及其应用

冯·米塞斯-费希尔(von Mises-Fisher，vMF)分布是一种定义在单位超球面上的概率分布，通过集中度参数(κ)和均值方向向量($\vec{\mu}$)刻画高维向量在球面上的方向性分布. 相较于传统欧式空间中的概率分布，vMF 分布更适合建模高维语义向量间的方向性关系，能够有效捕捉余弦相似度所反映的语义特征，适用于高维稀疏数据的场景.

vMF 分布已在多个领域展现其广泛的应用价值. 在无线通信领域，vMF 分布被用于替代高斯分布，提升三维空间中的定位鲁棒性[22]，在计算机视觉领域，vMF 分布常用于建模图像的方向性特征，广泛应用于人脸识别[23]、病理图像分析[24]、图像聚类[25]等任务；在自然语言处理领域，研究者利用 vMF 分布特性进行文本分类[26]和聚类任务[27]. 并在主题建模中通过 vMF 分布刻画嵌入向量的方向性关系[28].

2　基于上下文的神经作者-主题模型

如图 1 所示，本文提出的 CNATM 由 5 个部分构成：1)文档编码模块，以语料库中的文档 d 作为输入，通过 Transformer 编码器将文档转换为具有上下文语义信息的文档嵌入表示 $\hat{x}_c$，并通过文档推断网络将其转化为文档主题分布 $\hat{\theta}_d$；2)作者编码模块，以文档 d 对应的作者作为输入，将作者信息进行编码，通过作者推断网络得到作者对应的主题分布 $\hat{\theta}_a$；3)分布融合与先验匹配模块，将前两个模块得到的 $\tilde{\theta}_d$ 与 $\tilde{\theta}_a$ 按对应

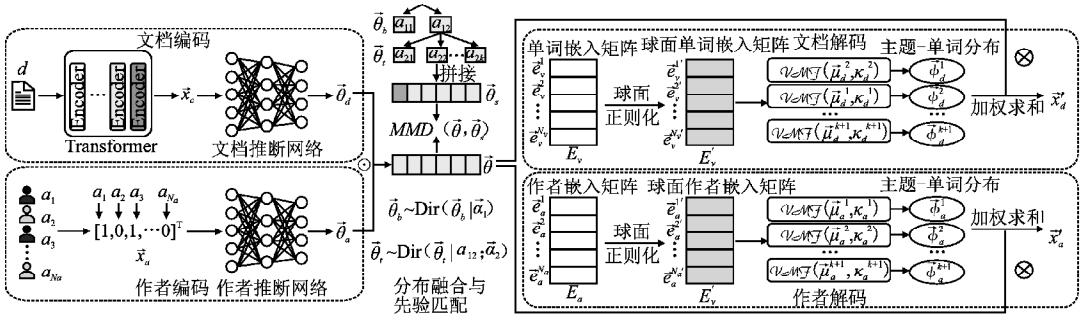

图 1　CNATM 模型结构

Fig. 1　Structure of CNATM model

元素相乘，并进行归一化，得到联合主题分布 $\hat{\theta}$，采用最大均值差异进行先验匹配；4)文档解码模块，融入词嵌入中的语义知识，在球面嵌入空间将每个主题建模成 vMF 分布，并由此得到主题-词分布 $\{\vec{\phi}_d^1,\vec{\phi}_d^2,\cdots,\vec{\phi}_d^{k+1}\}$，结合联合主题分布 $\hat{\theta}$，对文档进行重构；5)作者解码模块，融入作者嵌入中的语义知识，同样在球面嵌入空间将每个主题建模成 vMF 分布，在得到主题-作者分布 $\{\vec{\phi}_a^1,\vec{\phi}_a^2,\cdots,\vec{\phi}_a^{k+1}\}$ 后，结合联合主题分布 $\hat{\theta}$，对作者进行重构.

2.1　文档编码模块

为有效捕获文档中的语义特征，CNATM 在文档编码模块中采用预训练语言模型生成文档的语义表示，通过 Transformer 模型提取文档中每个单词的上下文嵌入向量，然后对所有单词的嵌入向量取平均值，生成文档级的语义表示，公式如下：

$$\hat{x}_c=\frac{1}{N}\sum_{i=1}^{N}\hat{e}_i \tag{1}$$

其中，$\hat{e}_i$ 为一篇文档中第 i 个单词的上下文嵌入向量，N 为该文档的单词数，$\hat{x}_c$ 表示文档嵌入. 为推断文档-主题分布，本文通过文档推断网络对其进行处理. 首先，对 $\hat{x}_c$ 进行线性变换，并通过谱归一化(Spectral Normalization，SN)稳定训练过程，经过 Softplus 激活函数处理后，得到隐藏层表示 $\vec{h}'$，随后，通过第 2 层线性变换和 Softplus 激活，将隐藏层输出映射为文档的最终语义表示 $\vec{s}'$，具体变换过程如下：

$$\vec{h}'=Softplus(SN(W_d^1\cdot\hat{x}_c+\vec{b}_d^1)) \tag{2}$$

$$\vec{s}'=Softplus(SN(W_d^2\cdot\vec{h}'+\vec{b}_d^2)) \tag{3}$$

最后，将 $\vec{s}'$ 映射为文档-主题分布 $\hat{\theta}_d$：

$$\hat{\theta}_d=Softmax(BN(W_D\cdot\vec{s}'+\vec{b}_D)) \tag{4}$$

其中 W_1^d、W_d^2 和 W_D 表示权重矩阵，$\vec{b}_d^1$、$\vec{b}_d^2$ 和 $\vec{b}_D$ 表示偏置项，推断得到的文档-主题分布 $\hat{\theta}_d$ 将用于后续的分布融合.

2.2　作者编码模块

为有效捕捉作者信息，CNATM 在作者编码模块中对每个文档的作者进行二进制编码. 假设数据集总共有 N_a 位不同的作者 $\{a_1,a_2,\cdots,a_{N_a}\}$，文档 d 的作者集合为 $\{a_1',a_2',\cdots,a_k'\}$，则可以通过一个长度为 N 的二进制向量表示文档的作者信息. 这种编码方式如下所示：

$$\hat{x}_a=[x_{a_1},x_{a_2},\cdots,x_{aN_a}] \tag{5}$$

其中，$\hat{x}_a$ 表示文档 d 的作者编码向量，如果作者 a_j($j\in\{1,2,\cdots,N_a\}$)参与撰写了文档 d，则 $x_{a_j}=1$，否则 $x_{a_j}=0$.

随后，通过作者推断网络推断作者-主题分布 $\hat{\theta}_a$，先利用两个非线性层提取其语义表示：

$$\vec{o}_1=LeakyReLU(BN(W_a^1\cdot\hat{x}_a+\vec{b}_a^1)) \tag{6}$$

$$\vec{o}_2=LeakyReLU(BN(W_a^2\cdot\vec{o}_1+\vec{b}_a^2)) \tag{7}$$

其中 BN 表示批归一化，W_a^1 和 W_a^2 为隐藏层的权重矩阵，$\vec{b}_a^1$ 和 $\vec{b}_a^2$ 为偏置项，$\vec{o}_1$ 和 $\vec{o}_2$ 表示隐藏层的输出信号.

最后，需要将 $\vec{o}_2$ 映射到主题空间中，从而得到作者-主题分布，变换过程如下：

$$\hat{\theta}_a=Softmax(BN(W_A\cdot\vec{o}_2+\vec{b}_A)) \tag{8}$$

其中，W_A 和 $\vec{b}_A$ 分别表示输出层的权重矩阵和偏置项，$\hat{\theta}_a$ 为作者-主题分布.

2.3 分布融合与先验匹配

在完成文档编码和作者编码后,模型得到了文档-主题分布 $\check{\theta}_d$ 与作者-主题分布 $\check{\theta}_a$,为更好地捕捉科研文献中研究热点与作者研究兴趣的协同作用,CNATM 采用以下方式融合文档和作者主题分布:

$$\check{\theta} = Softmax(\check{\theta}_d \odot \check{\theta}_a) \tag{9}$$

其中⊙表示向量对应元素相乘,通过 *Softmax* 归一化确保融合后的分布满足概率特性,$\check{\theta}$ 表示文档和作者的联合主题分布.

为建模文献摘要中的背景词以提升研究热点主题抽取质量,本文引入了狄利克雷树分布作为先验,狄利克雷树分布[9]的核心思想是将多项式分布的叶节点概率替换为路径上各分支概率的乘积表示. 如图 1 中上部的树形结构所示,CNATM 在建模过程中采用两层的狄利克雷树分布作为先验,第 1 层有两个分支,第 2 个分支对应的节点 a_{11} 用于采样背景-主题分布 $\check{\theta}_b$,表示通用背景主题的概率,另一个分支对应的节点 a_{12} 与第 2 层相连,第 2 层包含 k 个叶节点(a_{21} ~ a_{2k}),采样热点-主题分布 $\check{\theta}_t$,表示 k 个研究热点主题对应的概率$[p(a_{11}),p(a_{21}),\cdots,p(a_{2k})]$,各叶节点的概率可由以下公式计算得到:

$$p(a_{ij}|S,T) = \prod_{(\text{nodes } j)} \prod_{(\text{branches } c)} b_{jc}^{\delta_{jc}(a_{ij})} \tag{10}$$

其中,S 表示树中所有分支的概率参数集合,T 表示树的拓扑结构,$\delta_{jc}(\cdot)$为指示函数,其值依赖于分支 jc 是否与节点 a_{ij} 相连:

$$\delta_{jc}(a_{ij}) = \begin{cases} 1 & \text{如果分支 } jc \text{ 与节点 } a_{ij} \text{相连} \\ 0 & \text{其他} \end{cases} \tag{11}$$

在树结构中,分支概率 $b_j = \{b_{jc}\}$ 的先验分布由狄利克雷分布定义:

$$p(b_j|\alpha) \sim Dirichlet(\alpha_{jc}) \tag{12}$$

其中 α_{jc} 为狄利克雷分布的超参数,CNATM 通过两层树结构分别设置超参数 α_1 和 α_2,从而实现背景主题与热点主题的精细区分,第 1 层超参数 α_1 控制根节点到 a_{11} 和 a_{12} 的分支概率 b_{11}、b_{12},使背景-主题分布 $\check{\theta}_b$ 倾向于生成通用背景主题,第 2 层超参数 α_2 控制 a_{12} 到 a_{21} ~ a_{2k} 的 k 个分支概率 b_{21} ~ b_{2k},用于调节热点-主题分布 $\check{\theta}_t$ 中 k 个热点主题的区分度,这种分层设计相比传统单一狄利克雷分布更具灵活性,通过第 1 层超参数 α_1 降低背景主题的影响,第 2 层超参数 α_2 用来聚焦热点主题,从而提升主题抽取的可解释性与质量. 最终,背景-主题分布 $\check{\theta}_b = [p(a_{11})]$ 和热点-主题分布 $\check{\theta}_t = [p(a_{21}),p(a_{22}),\cdots,p(a_{2k})]$ 拼接得到 $\check{\theta}_s = [p(a_{11}),p(a_{21}),\cdots,p(a_{2k})]$,即经狄利克雷树分布采样得到的文档-主题分布. 先验匹配时,使用最大均值差异(Maximum Mean Discrepancy,MMD[29])来约束联合主题分布 $\check{\theta}$,使其向采样得到的 $\check{\theta}_s$ 靠近,相较于 KL 散度,MMD 不需要显式计算分布的概率密度,更适合高维潜在空间的分布对齐(MMD 的计算详见本文 2.6 节).

2.4 文档解码模块

为融合词嵌入的语义知识,本文采用 vMF 解码器对文档的主题进行建模,具体而言,每个主题被建模成一个多维 vMF 分布,对于预训练的词嵌入矩阵 E_v,首先需要将其映射到单位球面上,映射变化如下:

$$\check{e}_v^{i'} = \frac{\check{e}_v^i}{\| \check{e}_v^i \|_2} \tag{13}$$

其中 $\check{e}_v^i$ 表示词表中第 i 个单词的预训练词嵌入向量,CNATM 使用的是 GloVe 模型在计算语言学、计算机视觉和数据挖掘 3 个论文数据集上训练出的 300 维的嵌入向量,$\check{e}_v^{i'}$ 为球面归一化后的词嵌入向量,利用 vMF 分布 $v\text{MF}_d(\vec{\mu}_d^k,\kappa_d^k)$ 来建模第 k 个热点主题,$\vec{\mu}_d^k$ 与 κ_d^k 分别表示分布的平均方向和集中参数. 对于词表中的一个单词 $v_i(i \in \{1,2,\cdots,N_v\})$,在第 $k(k \in \{1,2,\cdots,K+1\})$ 个主题的概率 ϕ_{k,v_i} 可由 vMF 分布的概率密度函数计算得到:

$$p(\check{e}_v^{i'}|topic=k) = C_N(\kappa_d^k)\exp(\kappa_d^k \vec{\mu}_d^{kT} \check{e}_v^{i'}) \tag{14}$$

$$\phi_{k,v_i} = \frac{p(\check{e}_v^{i'}|topic=k)}{\sum_{i=1}^{N_v} p(\check{e}_v^{i'}|topic=k)} \tag{15}$$

上述公式中 $C_N(\kappa_d^k)$ 表示归一化项,其中 $\kappa_d^k \geqslant 0$. 经过计算得到热点-单词分布 $\{\vec{\phi}_d^1,\vec{\phi}_d^2,\cdots,\vec{\phi}_d^{K+1}\}$ 后,结合之前得到的联合主题分布 $\check{\theta}$,通过以下公式对文档进行重构:

$$\check{x}_d' = \sum_{k=1}^{K+1} \vec{\phi}_d^k \cdot \theta_k \tag{16}$$

其中 $\vec{\phi}_d^k$ 表示一篇文档第 k 个主题对应的热点-单词分布,θ_k 为 $\check{\theta}$ 第 k 维对应的概率,$\check{x}_d'$表示重构后的文档表示.

2.5 作者解码模块

作者解码的过程与文档解码类似,为融合作者嵌入中的语义知识,CNATM 同样采用 vMF 解码器对作者进行建模. 对于作者对应的预训练嵌入矩阵 E_a,首先将作者姓名去除空格并使用下划线连接(如“Qiang Ning”变为“Qiang_Ning”),然后将每篇文档的所有作者姓名拼接到摘要开头,形成新的文本序列,最后在整个语料库上使用 GloVe 模型进行训练,得到每个作者对应的 300 维嵌入向量,所有作者嵌入向量的堆叠构成了预训练嵌入矩阵 E_a. 对于其中每个作者的嵌入向量按照与文档解码模块相同的方式进行球面归一化,然后利用 vMF 分布 $v\text{MF}_a(\vec{\mu}_a^k,\kappa_a^k)$ 来建模第 k 个热点主题对应的作者,$\vec{\mu}_a^k$ 与 κ_a^k 分别表示分布的平均方向和集中参数. 一位作者 $a_i(i \in \{1,2,\cdots,N_a\})$ 在第 $k(k \in \{1,2,\cdots,K+1\})$ 个主题的概率 ϕ_{k,a_i} 可由 vMF 分布的概率密度函数计算得到:

$$p(\check{e}_a^{i'}|topic=k) = C_N(\kappa_a^k)\exp(\kappa_a^k \vec{\mu}_a^{kT} \check{e}_a^{i'}) \tag{17}$$

$$\phi_{k,a_i} = \frac{p(\check{e}_a^{i'}|topic=k)}{\sum_{i=1}^{N_v} p(\check{e}_a^{i'}|topic=k)} \tag{18}$$

其中 $C_N(\kappa_a^k)$ 表示归一化项,$\kappa_a^k \geqslant 0$,ϕ_{k,a_i} 表示作者 a_i 在第 k 个主题中的概率. 经过计算得到热点-作者分布 $\{\vec{\phi}_a^1,\vec{\phi}_a^2,\cdots,\vec{\phi}_a^{k+1}\}$ 后,结合联合主题分布 $\check{\theta}$,作者的重构方式如下:

$$\check{x}_a' = \sum_{k=1}^{K+1} \vec{\phi}_a^k \cdot \theta_k \tag{19}$$

其中 $\vec{\phi}_a^k$ 表示一篇文档第 k 个主题对应的热点-作者分布,θ_k 为 $\check{\theta}$ 第 k 维对应的概率,$\check{x}_a'$表示重构后的作者表示.

2.6 训练目标

在主题建模任务中,为使联合主题分布 $\check{\theta}$ 能够准确反映真实的分布,需要设计一个合适的训练目标来引导模型的学习过程. 为此,本文引入最大均值差异(Maximum Mean Discrepancy,MMD[29]),来衡量生成的分布与真实分布之间的差异. MMD 是一种基于核方法的统计量,可以有效地量化两个分布之间的距离. 其基本思想是通过计算生成分布与目标分布之间的距离来优化模型. 给定一批由模型生成的文档-主题

分布 $\Theta=\{\tilde{\theta}_1,\tilde{\theta}_2,\cdots,\tilde{\theta}_B\}$,作为近似后验分布 Q_Θ,以及对应从两层狄利克雷树分布的先验 P_θ 中采样得到的文档-主题分布 $\Theta'=\{\tilde{\theta}'_1,\tilde{\theta}'_2,\cdots,\tilde{\theta}'_B\}$,MMD 的定义如下:

$$MMD(Q_\Theta,P_\Theta)=\widehat{MMD}(\Theta,\Theta')=\frac{1}{B(B-1)}\sum_{m\neq n}k(\tilde{\theta}_m,\tilde{\theta}_n)+\frac{1}{B(B-1)}\sum_{m\neq n}k(\tilde{\theta}'_m,\tilde{\theta}'_n)-\frac{2}{B^2}\sum_{m,n}k(\tilde{\theta}_m,\tilde{\theta}'_n)\tag{20}$$

其中,B 表示批量大小,$k(\cdot,\cdot)$是核函数,通过最小化该 MMD 损失,模型能够逐步调整生成的文档-主题分布,使其与真实分布的差异最小化,从而提高模型对主题的捕捉能力,最终获得更精确的主题.

模型的训练目标为最小化总损失 $\mathcal{L}$,该损失由重构损失 $\mathcal{L}_r$ 和先验匹配损失 $\mathcal{L}_{MMD}$ 两部分组成,定义如下:

$$\mathcal{L}=\mathcal{L}_r+\mathcal{L}_{MMD}\tag{21}$$

其中,重构损失 $\mathcal{L}_r$ 由文档重构损失 $\mathcal{L}_{r\_doc}$ 和作者重构损失 $\mathcal{L}_{r\_aut}$ 两部分组成,均采用交叉熵损失函数计算,公式如下:

$$\begin{aligned}\mathcal{L}_r&=\mathcal{L}_{r\_doc}+\mathcal{L}_{r\_aut}\\&=\lambda_d\cdot\mathbb{E}_{P_{\tilde{x}_d}}\mathbb{E}_{Q_\Theta(\tilde{\theta}|\tilde{x}_d)}c(\tilde{x}_d,\tilde{x}'_d)+\lambda_a\cdot\mathbb{E}_{P_{\tilde{x}_a}}\mathbb{E}_{Q_\Theta(\tilde{\theta}|\tilde{x}_a)}c(\tilde{x}_a,\tilde{x}'_a)\\&=\left[-\lambda_d\cdot\frac{1}{B}\sum_{b=1}^{B}\sum_{v=1}^{N_v}x_v\log(\tilde{x}'_v)\right]+\left[-\lambda_a\cdot\frac{1}{B}\sum_{b=1}^{B}\sum_{w=1}^{N_a}x_w\log(\tilde{x}'_w)\right]\end{aligned}\tag{22}$$

文档重构损失 $\mathcal{L}_{r\_doc}$ 衡量输入文档的词袋表示 $\tilde{x}_d$ 与重构文档 $\tilde{x}'_d$ 之间的差异,作者重构损失 $\mathcal{L}_{r\_aut}$ 衡量作者的二进制编码 $\tilde{x}_a$ 与重构后的作者表示 $\tilde{x}'_a$ 之间的差异,$c(\cdot,\cdot)$表示交叉熵损失函数,λ_d 和 λ_a 为加权系数,其中 λ_d 设置为 $1/(l_d\log N_v)$,l_d 为文档的平均长度,N_v 为词表大小,而 λ_a 设置为 $1/(l_a\log N_a)$,l_a 表示文档的平均作者数量,N_a 为作者总数,这些系数的设定遵循 Nan 等人[30]提出的方案.

先验匹配损失 $\mathcal{L}_{MMD}$ 用于优化联合主题分布 $\tilde{\theta}$,使其与先验分布 $\tilde{\theta}_s$ 对齐,定义为:

$$\mathcal{L}_{MMD}=MMD(\tilde{\theta},\tilde{\theta}_s)\tag{23}$$

其计算方法已在公式(20)中给出.

3　实验

3.1　实验设置

本文基于从 DBLP 爬取的数据构建了 3 个科研论文数据集:计算语言学(CL)、计算机视觉(CV)和数据挖掘(DM),并在此基础上进行实验. 在构建过程中,本文参考了谷歌学术"热门出版物"中的子类别排名,筛选了排名前 20 的顶级期刊或会议的文章. 这种筛选方式确保所选文献在各自领域内具有较高的学术影响力,同时也保证了数据集中文本内容的质量和权威性.

所有数据集均经过一系列的文本预处理操作,包括拼写检查[1]、词形还原[2]、停用词移除以及低频词过滤. 此外,为提高数据质量,本文对摘要内容以句子为单位进行了针对性筛选,具体方法为:定义任务关键词列表和模型关键词列表,分别涵盖计算语言学、计算机视觉和数据挖掘领域中常见任务和模型的相关术语;对于摘要里的每个句子,计算其与两个关键词列表的余弦相似度,若句子与任务关键词列表的最大相似度高于与模型关键词列表的最大相似度,则保留该句子,否则过滤掉. 经过预处理后的数据集统计信息如表 1 所示.

表 1　数据集统计信息

Table 1　Statistics of the datasets

数据集	文档数	单词数	作者数
计算语言学(CL)	23056	4549	30343
计算机视觉(CV)	20195	4789	39584
数据挖掘(DM)	14994	5018	34128

本文选取了以下基准方法进行对比,并采用其官方实现或建议参数设置:

1)LDA[1]:一种经典的基于概率图的主题模型,假设文档由多个主题组成,每个主题由一组词的概率分布表示,而每个词的出现概率由主题分布和词分布共同决定,实验中使用基于 Gibbs 采样的实现方法[3].

2)BAT[5]:一种基于双向对抗训练的神经主题模型,由生成器和判别器两部分组成,生成器学习生成符合真实文档分布的主题,判别器区分真实文档和生成文档的主题分布.

3)GBAT[5]:BAT 模型的一种扩展,在生成器中引入多元高斯分布对每个主题进行建模,从而更好地捕捉主题.

4)CTM[31]:一种基于 VAE 的神经主题模型,通过引入上下文嵌入来提升主题的一致性,利用变分推理优化主题分布的学习过程[4].

5)BERTopic[19]:一种结合预训练语言模型和聚类技术的神经主题模型,通过 BERT 生成文档嵌入,然后对文档嵌入进行降维处理,最后使用聚类算法对文档进行分组[5].

6)CTMNeg[32]:CTM 模型的一种扩展,通过引入负采样机制增强主题区分能力,提高了模型的鲁棒性[6].

7)vONTSS[28]:一种基于 VAE 的神经主题模型,利用冯·米塞斯-费舍尔分布对主题进行建模,能够更好地捕捉主题空间中的方向性特征[7].

8)ECRTM[20]:一种引入嵌入聚类正则化(ECR)机制的神经主题模型,在训练过程中对嵌入空间施加聚类约束,从而提升主题的多样性[8].

9)CWTM[33]:一种结合 BERT 上下文词嵌入的神经主题模型,利用 BERT 模型的上下文捕捉能力提升主题建模的

[1]https://github.com/barrust/pyspellchecker

[2]https://github.com/explosion/spaCy

[3]https://gibbslda.sourceforge.net

[4]https://github.com/MilaNLProc/contextualized-topic-models

[5]https://github.com/MaartenGr/BERTopic

[6]https://github.com/AdhyaSuman/CTMNeg

[7]https://github.com/xuweijieshuai/Neural-Topic-Modeling-vmf

[8]https://github.com/BobXWu/ECRTM

效果[9].

10) FASTopic[34]:一种结合最优传输理论和预训练词嵌入的神经主题模型,通过最优传输框架优化文档与主题之间的分布匹配,能够高效地进行主题建模[10].

11) BertSenClu[35]:一种基于 Sentence Transformer 的主题建模方法,利用其生成的句子嵌入,通过期望最大化(EM)算法进行主题推断,强调句子级语义信息的利用[11].

12) KeyNMF[36]:一种结合 Transformer 上下文嵌入的主题建模方法,通过非负矩阵分解在静态或动态语境下提取主题结构[12].

13) Semantic Signal Separation[37](在表 2 中模型名缩写为 SSS):一种基于独立成分分析的主题建模方法,通过将上下文嵌入分解为相互独立的语义成分以提取主题[13].

14) CASTer[38]:一种结合上下文嵌入与自相似性机制的主题建模方法,通过候选词在语料中的上下文一致性筛选主题词,从而提升主题质量[14].

15) TNTM[39]:一种基于 Transformer 语义嵌入的神经主题模型,采用变分自编码器进行主题建模,能够高效地进行主题推断[15].

16) ATM[6]:一种基于概率图的作者主题模型,假设每个作者与一组主题相关联,而文档的主题分布由其作者的主题偏好决定,实验中使用 gensim 库中的实现方法[16].

17) FANToM[8]:一种融合作者信息的神经主题模型,通过学习作者与主题的对齐关系,提升主题质量并挖掘作者的研究兴趣[17].

3.2 评价指标

为全面评估模型的性能,本文采用了以下评价指标:主题一致性、主题多样性以及主题-作者相关性.

主题一致性通过 4 个广泛使用的指标进行评估:C_P、C_A、*NPMI* 和 *UCI*,用于衡量主题的语义一致性. 所有主题一致性指标均通过 Palmetto 库[18]进行计算. 较高的指标值表示主题在语义上更为一致且质量更高.

主题多样性通过计算唯一术语(Unique Term, UT)来衡量,计算公式如下:

$$UT = \frac{N_U}{N_T} \tag{24}$$

其中,N_U 表示所有主题中唯一术语的数量,N_T 表示所有主题中的术语总数. 较高的 *UT* 值表示主题具有更高的多样性.

主题-作者相关性通过两个新指标评估:

1) BM25-MaxSim(Maximum Similarity based on BM25, BMS):该指标评估模型抽取的作者与热点主题的相关性,基于作者文章摘要与主题词集合之间的 BM25 相似度. 对于每位作者,选取其所有摘要与该作者对应的主题词集合的最大相似度得分作为该作者的相关性得分;然后,对每组作者的得分取平均,再对所有组取平均,得到整体相关性水平. 其计算公式如下:

$$BMS = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{1}{N_{A_i}}\sum_{a\in A_i,d\in Abs(a)}\max(BM25(d,Q_i))\right) \tag{25}$$

其中 K 表示主题数(行数),N_{A_i} 表示一行热点主题对应的作者数量,实验中设置为 10, A_i 表示第 i 个主题(第 i 行,$i\in[1,K]$)对应的作者集合,$Abs(a)$ 表示作者 a 对应的摘要集合,d 表示一篇摘要,Q_i 表示第 i 个主题对应的单词集合.

2) TFIDF-TermStrength(Term Strength based on TF-IDF, TTS):该指标评估模型抽取的作者与科研热点之间的相关性,通过累加作者文章摘要中每个主题词的 TF-IDF 值实现. 对于每位作者,针对其对应主题词集合中的每个词,选取其在所有摘要中的最大 TF-IDF 值并求和,得到该作者的相关性得分;然后,对所有作者和所有组的得分求总和,得到整体相关性强度. 其计算公式如下:

$$TTS = \sum_{i=1}^{K}\sum_{a\in A_i}\sum_{w\in Q_i}\max(TFIDF(w,d)) \tag{26}$$

其中 $TFIDF(w,d)$ 表示单词 w 在摘要 d 中的 TF-IDF 值,其他符号与公式(25)中的含义相同.

3.3 实验结果和分析

为全面评估模型的表现,本文在每个数据集上进行 5 组不同主题数(K = 20,30,50,75,100)的实验,并计算了主题一致性和主题多样性指标的平均值. 实验结果如表 2 所示.

从表 2 的实验结果可以看出,在 5 种不同主题数设置下,CNATM 在 CL、CV 和 DM3 个数据集上的主题一致性和主题多样性指标均优于对比方法,CNATM 的良好表现可能归因于:1)使用 Transformer 模型捕捉摘要上下文信息,增强了主题词的语义一致性;2)引入预训练词嵌入,通过外部语义知识提升了主题质量;3)使用狄利克雷树分布建模主题,能够有效地区分出背景词和主题词,从而提高了主题的可解释性和多样性.

为进一步展示主题一致性随主题数的变化趋势,图 2 以折线图的形式呈现了 CNATM 的表现,选取 ATM、FANToM、BERTopic、KeyNMF 以及 CWTM 进行对比,黑色折线表示 CNATM,灰色折线表示各对比模型,图例位于折线图下方. 之所以仅选取上述模型做折线图进行对比,原因如下:

1)模型的类型:ATM 和 FANToM 是对比模型中具备作者主题建模能力的两种方法,能够挖掘作者与主题的关联,因

[9] https://github.com/Fitz-like-coding/CWTM

[10] https://github.com/BobXWu/FASTopic

[11] https://github.com/JohnTailor/BertSenClu

[12] https://x-tabdeveloping.github.io/turftopic/KeyNMF/

[13] https://x-tabdeveloping.github.io/turftopic/s3/

[14] https://github.com/yananma1029/CAST

[15] https://github.com/ArikReuter/TNTM

[16] https://radimrehurek.com/gensim/models/atmodel.html

[17] https://github.com/mayanknagda/fantom

[18] https://github.com/dice-group/Palmetto

表 2　实验结果

Table 2　Experimental results

Dataset	Model	C_P	C_A	*NPMI*	*UCI*	*UT*
CL	LDA	0.181	0.154	0.007	−0.650	0.649
	BAT	0.033	0.146	−0.042	−1.794	0.606
	GBAT	−0.202	0.104	−0.061	−2.201	0.554
	CTM	0.093	0.153	−0.025	−1.460	0.555
	BERTopic	0.057	0.154	−0.036	−1.726	0.753
	CTMNeg	0.079	0.152	−0.017	−1.165	0.643
	vONTSS	0.042	0.127	−0.029	−1.220	0.914
	ECRTM	−0.123	0.136	−0.058	−2.235	0.805
	CWTM	0.215	0.159	0.015	−0.545	0.622
	FASTopic	−0.199	0.138	−0.077	−2.650	0.852
	BertSenClu	−0.361	0.122	−0.104	−3.264	0.908
	KeyNMF	0.062	0.150	−0.026	−1.576	0.593
	SSS	−0.242	0.142	−0.083	−2.802	0.879
	CAST	−0.408	0.117	−0.072	−2.335	0.682
	TNTM	−0.093	0.122	−0.051	−1.792	0.908
	ATM	0.042	0.141	−0.043	−1.602	0.479
	FANToM	−0.076	0.128	−0.075	−2.328	0.770
	CNATM	**0.227**	**0.163**	**0.044**	**−0.038**	**0.945**
CV	LDA	0.183	0.150	0.003	−0.680	0.647
	BAT	0.117	0.152	−0.028	−1.472	0.673
	GBAT	0.194	0.148	0.023	−0.296	0.605
	CTM	0.095	0.138	−0.031	−1.529	0.579
	BERTopic	0.134	0.161	−0.018	−1.346	0.736
	CTMNeg	0.091	0.139	−0.028	−1.464	0.698
	vONTSS	0.021	0.121	−0.061	−2.092	0.855
	ECRTM	−0.068	0.139	−0.071	−2.598	0.825
	CWTM	0.226	0.154	0.026	−0.189	0.598
	FASTopic	−0.123	0.130	−0.086	−2.921	0.898
	BertSenClu	−0.281	0.132	−0.123	−3.928	0.896
	KeyNMF	0.103	0.144	−0.034	−1.712	0.590
	SSS	−0.170	0.145	−0.095	−3.204	0.881
	CAST	−0.357	0.093	−0.112	−3.298	0.665
	TNTM	−0.078	0.124	−0.070	−2.270	0.895
	ATM	0.121	0.136	−0.011	−0.759	0.455
	FANToM	−0.001	0.131	−0.073	−2.335	0.745
	CNATM	**0.270**	**0.166**	**0.050**	**0.092**	**0.941**
DM	LDA	0.192	0.159	0.009	−0.642	0.725
	BAT	0.192	0.170	−0.004	−0.982	0.652
	GBAT	0.169	0.155	0.005	−0.797	0.661
	CTM	0.161	0.170	−0.005	−1.049	0.544
	BERTopic	0.136	0.172	−0.015	−1.304	0.685
	CTMNeg	0.162	0.157	−0.001	−0.847	0.694
	vONTSS	0.045	0.123	−0.045	−1.633	0.773
	ECRTM	0.021	0.156	−0.048	−2.109	0.741
	CWTM	0.200	0.159	−0.018	−1.346	0.544
	FASTopic	0.023	0.145	−0.061	−2.167	0.803
	BertSenClu	−0.258	0.137	−0.128	−4.107	0.903
	KeyNMF	0.165	0.161	0.009	−0.823	0.673
	SSS	−0.262	0.165	−0.074	−2.504	0.883
	CAST	−0.286	0.155	−0.060	−1.985	0.567
	TNTM	−0.170	0.131	−0.087	−2.734	0.866
	ATM	0.187	0.167	0.005	−0.644	0.652
	FANToM	0.062	0.134	−0.044	−1.615	0.774
	CNATM	**0.284**	**0.176**	**0.057**	**0.223**	**0.916**

此与同样支持作者主题建模的 CNATM 可进行直接对比.

2)模型的性能:根据表 2 的实验结果,BERTopic、KeyNMF 和 CWTM 这 3 个模型的总体表现较好,同时也是当前神经主题模型中最具代表性的方法,适合作为基准进行评估.

3)可视化效果:若将表 2 中的所有对比模型纳入折线图,过多的折线会导致图例过于复杂且线条重叠,难以分辨,降低折线图的可读性.

因此,仅选取了上述 5 个模型,以确保图 2 清晰直观.

从图 2 的折线图中可以看出,在所有数据集和不同主题数的条件下,CNATM 在 4 个主题一致性指标上均优于各对比模型,进一步验证了 CNATM 在主题一致性方面的优越性.

在主题-作者相关性的评估中,考虑到 ATM 和 FANToM 是对比模型中具备作者建模能力的方法,因此选择这两个模型作为强对比模型,以衡量模型在挖掘作者与科研任务相关性方面的表现. 表 3 展示了 ATM、FANToM 与 CNATM 在 BMS 和 TTS 这两个相关性指标上的对比结果.

表 3　作者与科研热点的相关程度

Table 3　Relevance of the author to the research hotspot

		CL		CV		DM	
		BMS	TTS	BMS	TTS	BMS	TTS
20	ATM	5.15	60.495	5.19	120.082	4.12	107.594
	FANToM	2.23	104.625	2.50	163.032	2.22	139.596
	CNATM	**5.19**	**234.353**	**5.76**	**229.191**	**6.09**	**228.017**
30	ATM	4.38	187.234	4.30	154.117	3.17	149.751
	FANToM	2.42	178.624	2.23	221.927	1.82	173.459
	CNATM	**5.13**	**328.072**	**5.68**	**324.705**	**6.22**	**318.797**
50	ATM	3.10	115.080	2.96	216.692	2.08	198.001
	FANToM	2.25	279.179	2.26	330.374	2.01	301.338
	CNATM	**4.25**	**439.682**	**5.18**	**418.715**	**5.39**	**437.782**
75	ATM	2.39	177.660	2.35	306.931	1.50	203.354
	FANToM	2.25	408.786	2.18	463.456	1.86	424.139
	CNATM	**3.73**	**505.593**	**4.92**	**557.089**	**5.15**	**566.604**
100	ATM	2.33	268.915	2.29	290.453	1.42	262.273
	FANToM	2.32	568.961	2.45	519.374	2.02	560.036
	CNATM	**3.65**	**662.912**	**4.37**	**621.124**	**5.02**	**723.536**

从表 3 的实验结果可以看出,CNATM 在这两个相关性指标上的表现均优于 ATM 与 FANToM,表明其在主题-作者相关性方面表现更好.

为更直观地比较 CNATM 与 ATM 以及 FANToM 抽取出的主题及对应作者,表 4 展示了 3 个模型在问答系统(QA)、句法分析(Syntactic Parsing)和机器翻译(Machine Translation)3 个主题上的抽取结果.

可以看到 ATM 与 FANToM 抽取的主题词中均包含背景词,且部分作者与主题关联性较低,而 CNATM 在主题词聚焦度和作者匹配精度上均表现得更好.

3.4　消融实验

为验证狄利克雷树分布和 Transformer 编码器生成的文档表示对模型性能的影响,本文设置了以下对比模型进行消融实验:

1)w/o dirichlet tree distribution:在 CNATM 基础上改用狄利克雷分布作为先验进行建模;

2)w/o Transformer encoder:去除 Transformer 编码器,改

用文档的词袋表示并通过多层感知机获取文档主题分布.

这部分消融实验在 CL 数据集上进行,基于 5 种主题数($K=20,30,50,75,100$)的指标均值结果如表 5 所示.

表 5 的结果显示,去除狄利克雷树分布或 Transformer 编码器均会导致模型性能的下降,说明两者对科研热点挖掘任务有积极影响,去除狄利克雷树分布会导致模型难以区分背景词和主题词,降低主题聚焦度;去除 Transformer 编码器则因无法捕捉上下文信息而影响主题质量.综上所述,狄利克雷树分布和 Transformer 编码器的引入有效提升了模型性能.

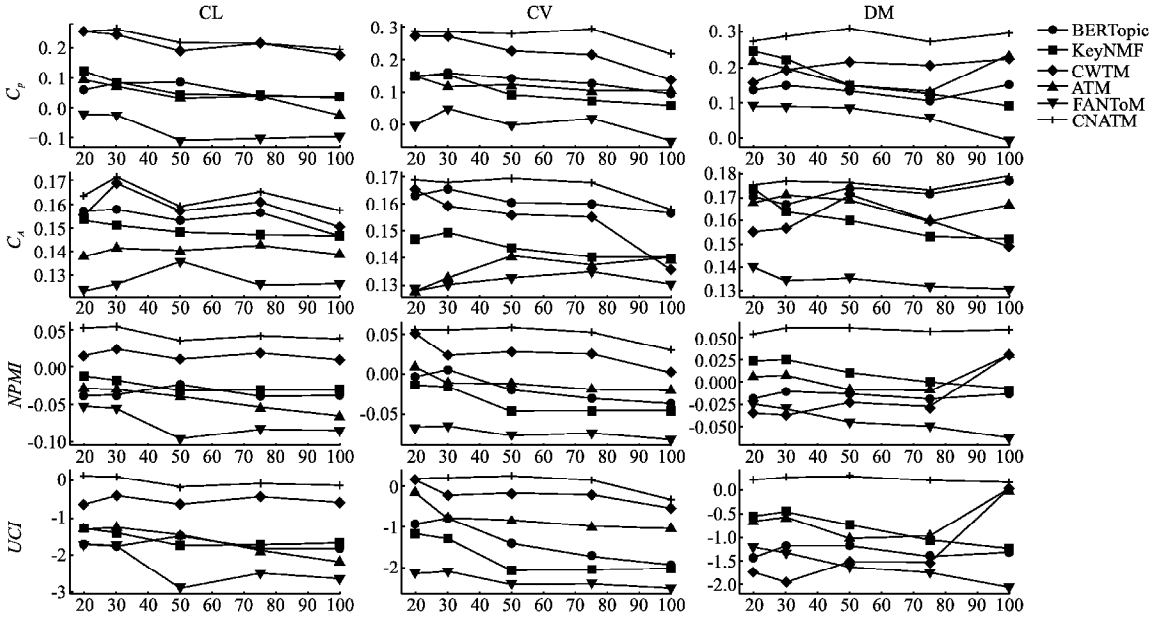

图 2 主题一致性实验结果

Fig. 2 Experiment results about topic coherence

表 4 主题及对应作者抽取样例

Table 4 Topic and corresponding author extraction examples

ATM	问答系统(QA)	question answer reasoning summarization retrieval summary require prompt explanation ability
	对应作者	Hannaneh_Hajishirzi Peter_Clark Nan_Duan Lei_Li M_Nagata Matt_Gardner Mark_Hopkins Alvaro_Herrasti Cristian_Petrescu-Prahova Chitta_Baral
	句法分析(Syntactic Parsing)	bert syntactic annotation bias contextualize parser transfer evaluate parse abstract
	对应作者	Baobao_Chang Yulia_Tsvetkov Nathan_Schneider Man_Lan David_Jurgens Nicholas_Asher Djamé_Seddah Swabha_Swayamdipta Shashi_Narayan Kalina_Bontcheva
	机器翻译(Machine Translation)	machine transfer score translate evaluate alignment test automatic reference monolingual
	对应作者	M_Utiyama B_Haddow Alexander_M_Fraser Chris_Callison-Burch Marcin_Junczys-Dowmunt Jianfeng_Gao Ondrej_Bojar Michael_Auli Benjamin_Marie Matt_Post
FANToM	问答系统(QA)	sense comprehension question answer wordnet understand choice analogy example meaning
	对应作者	Shizhu_He Karthik_Gopalakrishnan A_Korhonen Pei_Zhou Rahul_Khanna Mohammad_Taher_Pilehvar Nanyun_Peng Tejas_Gokhale Wee_Chung_Gan J_Malmaud
	句法分析(Syntactic Parsing)	parser segmentation feature translate number switch conll rely datum syntactic
	对应作者	Suyoun_Kim Rosana_Ardila Lillian_Lee Jennifer_Foster M_Kohler Irshad_Ahmad_Bhat J_Hansen M_Erp Kim_Gerdes Alankar_Jain
	机器翻译(Machine Translation)	decode encoder attention translation relevant give syntax agent sentence parameter
	对应作者	Yuan-Fang_Li Xing_Wang Tao_Qin Weihua_Luo Nancy_F_Chen Hannaneh_Hajishirzi Tsendsuren_Munkhdalai Xinyu_Hua S_Iida Lifeng_Shang
CNATM	问答系统(QA)	question answer challenge address strategy argument consider issue response call
	对应作者	Ting_Liu Mohit_Bansal Preslav_Nakov Luke_Zettlemoyer Xu_Sun Maosong_Sun Nanyun_Peng Yoav_Goldberg Bing_Qin Xiang_Ren
	句法分析(Syntactic Parsing)	semantic annotation syntactic similarity lexical textual utterance contextual paraphrase parse
	对应作者	Ting_Liu Mohit_Bansal Luke_Zettlemoyer Christopher_D_Manning Hannaneh_Hajishirzi Wanxiang_Che Jianfeng_Gao Kentaro_Inui M_Surdeanu Noah_A_Smith
	机器翻译(Machine Translation)	translation english multilingual linguistic german arabic bilingual lexicon grammar french
	对应作者	Xuanjing_Huang M_Zhou Pascale_Fung Qi_Zhang Xipeng_Qiu D_Klakow Nan_Duan Furu_Wei D_Gildea Tao_Gui

本文在 CNATM 模型中使用的 Transformer 模型如下:

1)BERT-base-nli-mean-tokens[19](缩写为 MEAN),一个用于句子级语义建模的 BERT 模型.

2)Sentence-T5-base[20](缩写为 T5),一个基于 T5 架构的句子嵌入模型.

表 5　不同消融模型的实验结果

Table 5　Experimental results of ablation experiments

模型名	C_P	C_A	*NPMI*	*UCI*	*UT*
w/o dirichlet tree distribution	0.216	0.162	0.042	−0.056	0.944
w/o Transformer encoder	0.202	0.163	0.038	−0.157	0.939
CNATM	**0.227**	**0.163**	**0.044**	**−0.038**	**0.945**

3)All-mpnet-base-v2[21](缩写为 MPNET),一个高性能的通用句子嵌入模型.

4)All-MiniLM-L6-v2[22](缩写为 MINILM),一个轻量级句子嵌入模型.

5)BERT-base-uncased[23](缩写为 BASE),基础的 BERT 模型.

6)MiniCPM-Embedding[24](缩写为 CPM),一个中英双语句子嵌入模型,具备跨语言检索能力.

7)Qwen3-Embedding-0.6B[25](缩写为 QWEN),一个多语言句子嵌入模型,具备文本检索、聚类与排序能力.

为评估不同 Transformer 模型对科研热点挖掘任务的影响,本文在 CL 数据集上进行了消融实验,在 5 组不同主题数(K=20,30,50,75,100)下各指标的平均结果如表 6 所示.

表 6　不同 Transformer 模型的消融实验结果

Table 6　Experimental results of ablation using different transformer models

模型名	C_P	C_A	*NPMI*	*UCI*	*UT*
CNATM + MEAN	**0.227**	0.163	0.044	**−0.038**	0.945
CNATM + T5	0.217	**0.167**	**0.045**	−0.041	0.934
CNATM + MPNET	0.211	0.158	0.038	−0.127	**0.955**
CNATM + MINILM	0.213	0.159	0.040	−0.123	0.945
CNATM + BASE	0.199	0.159	0.037	−0.190	0.941
CNATM + CPM	0.219	0.160	0.041	−0.056	0.949
CNATM + QWEN	0.223	0.162	0.041	−0.106	0.954

根据表 6 可以观察到,CNATM + T5 的组合在各项指标上表现均衡,而 CNATM + BASE 的表现相对较差,主要是因为其作为基础的 BERT 模型,缺乏句子级语义建模能力. 对于 CL 数据集,Sentence-T5-base(CNATM + T5)是较优选择;若针对其他数据集建模,则需根据模型的具体表现选择合适的预训练模型.

3.5　参数分析

为评估超参数对模型性能的影响,本文在 CL 数据集上调节学习率(Learning rate)、第 1 层和第 2 层狄利克雷树分布的超参数(α_1,α_2)进行参数分析,主题数设置为 30,实验结果如表 7 所示.

表 7　参数分析

Table 7　Parameter analysis

参数名	值	C_P	C_A	*NPMI*	*UCI*	*UT*
学习率	0.001	0.234	0.160	**0.050**	**0.149**	0.960
	0.003	**0.250**	**0.163**	0.049	0.052	0.940
	0.005	0.203	0.152	0.032	-0.208	**0.977**
	0.007	0.153	0.138	0.016	-0.414	0.950
	0.009	0.173	0.151	0.021	-0.445	0.963
	0.01	0.169	0.142	0.015	-0.512	0.950
α_1-第 1 层狄利克雷树分布的超参数	2	0.212	0.168	0.048	0.018	0.960
	1	0.197	0.165	0.040	-0.196	0.937
	0.5	0.248	0.167	**0.050**	0.066	0.933
	0.1	0.217	0.156	0.048	**0.129**	**0.963**
	0.01	**0.237**	**0.169**	0.046	-0.098	0.950
	0.001	0.207	0.158	0.040	-0.135	0.953
α_2-第 2 层狄利克雷树分布的超参数	2	0.207	0.158	0.045	-0.024	0.950
	1	0.246	0.166	0.045	-0.021	0.930
	0.5	0.218	0.164	0.044	-0.017	0.953
	0.1	0.255	0.167	**0.053**	**0.121**	0.940
	0.01	**0.258**	**0.171**	0.051	0.096	0.947
	0.001	0.207	0.166	0.050	0.099	**0.967**

表 7 表明,学习率增大时,主题一致性逐渐下降,而对主题多样性的影响较小,学习率为 0.001 时表现最优;较大的 α_1 能够更好的区分背景主题,α_1=2 时模型的表现最为均衡;较大的 α_2 下模型在主题一致性和多样性上表现一般,α_2=0.01能在两者间实现最佳平衡. 综上,α_1=2 和 α_2=0.01 的组合使 CNATM 表现最优.

4　结束语

本文提出的 CNATM 融合了预训练语言模型和球面主题建模的优势,有效解决了传统主题模型在语境建模与主题区分上的局限性. 通过引入 Transformer 模型生成的上下文嵌入表示,结合 von Mises-Fisher 分布进行主题建模,不仅提升了文档的语义表示能力,还有效地捕捉了不同主题间的方向性关系. 此外,狄利克雷树分布的引入使得模型能够区分背景主题与任务主题,进一步提升了主题一致性和多样性. 实验结果表明,CNATM 在主题一致性、主题多样性以及作者-主题关联性等方面表现出色.

CNATM 的创新在于将预训练语言模型、球面主题建模与作者信息相结合,为科研热点挖掘提供了更精确的工具. 未来研究将探索动态建模,以适应科研领域的快速变化,追踪热点演变,及时捕捉新兴趋势和创新点. 随着技术不断完善,

[19] https://huggingface.co/sentence-transformers/bert-base-nli-mean-tokens

[20] https://huggingface.co/sentence-transformers/sentence-t5-base

[21] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[22] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[23] https://huggingface.co/google-bert/bert-base-uncased

[24] https://huggingface.co/openbmb/MiniCPM-Embedding

[25] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B

CNATM 有望为科研人员提供更有价值的工具,推动科研管理和知识发现领域的进一步发展.

References:

[1] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.

[2] Zhang Y, Jin R, Zhou Z H. Understanding bag-of-words model: a statistical framework[J]. International Journal of Machine Learning and Cybernetics, 2010, 1(1): 43-52.

[3] Miao Y, Yu L, Blunsom P. Neural variational inference for text processing[C]//International Conference on Machine Learning, 2016: 1727-1736.

[4] Wang R, Zhou D, He Y. Atm: adversarial-neural topic model[J]. Information Processing & Management, 2019, 56(6), doi: 10.48550/arXiv.1811.00265.

[5] Wang R, Hu X, Zhou D, et al. Neural topic modeling with bidirectional adversarial training[C]//58th Annual Meeting of the Association for Computational Linguistics, 2020: 340-350.

[6] Rosen Zvi M, Griffiths T, Steyvers M, et al. The author-topic model for authors and documents[C]//Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, 2004: 487-494.

[7] Zhang D C, Lauw H W. Variational graph author topic modeling [C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022: 2429-2438.

[8] Nagda M, Ostheimer P, Fellenz S. Tethering broken themes: aligning neural topic models with labels and authors[J]. arXiv preprint arXiv: 2410.18140, 2024.

[9] Minka T. The dirichlet-tree distribution[EB/OL]. https://tminka.github.io/papers/dirichlet/minka-dirtree.pdf, 1999.

[10] Waltman L, Van Eck N J. A new methodology for constructing a publication-level classification system of science[J]. Journal of the American Society for Information Science and Technology, 2012, 63(12): 2378-2392.

[11] Ding Y, Zhang G, Chambers T, et al. Content-based citation analysis: the next generation of citation analysis[J]. Journal of the Association for Information Science and Technology, 2014, 65(9): 1820-1833.

[12] Chen C. Science mapping: a systematic review of the literature[J]. Journal of Data and Information Science, 2017, 2(2): 1-40.

[13] Hou J, Yang X, Chen C. Emerging trends and new developments in information science: a document co-citation analysis (2009-2016) [J]. Scientometrics, 2018, 115(2): 869-892.

[14] Zhang D, Zhang Z, Managi S. A bibliometric analysis on green finance: current status, development, and future directions[J]. Finance Research Letters, 2019, 29(C): 425-430, doi: 10.1016/j.frl.2019.02.003.

[15] Pesta B, Fuerst J, Kirkegaard E O W. Bibliometric keyword analysis across seventeen years (2000-2016) of intelligence articles[J]. Journal of Intelligence, 2018, 6(4): 46, doi: 10.3390/jintelligence6040046.

[16] Church K W. Word2Vec[J]. Natural Language Engineering, 2017, 23(1): 155-162.

[17] Pennington J, Socher R, Manning C D. Glove: global vectors for word representation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2014: 1532-1543.

[18] Meng Y, Huang J, Wang G, et al. Spherical text embedding[J]. Advances in Neural Information Processing Systems, 2019, 32, doi: 10.48550/arXiv.1911.01196.

[19] Grootendorst M. BERTopic: neural topic modeling with a class-based TF-IDF procedure[J]. arXiv preprint arXiv: 2203.05794, 2022.

[20] Wu X, Dong X, Nguyen T T, et al. Effective neural topic modeling with embedding clustering regularization[C]//International Conference on Machine Learning, 2023: 37335-37357.

[21] Pham C, Hoyle A, Sun S, et al. TopicGPT: a prompt-based topic modeling framework[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2024: 2956-2984.

[22] Nurminen H, Suomalainen L, Ali Loytty S, et al. 3D angle-of-arrival positioning using von Mises-Fisher distribution[C]//21st International Conference on Information Fusion, 2018: 2036-2041.

[23] Conti J R, Noiry N, Clemencon S, et al. Mitigating gender bias in face recognition using the von mises-fisher mixture model[C]//International Conference on Machine Learning, 2022: 4344-4369.

[24] Alirezazadeh P, Dornaika F, Charafeddine J. Mises-Fisher similarity-based boosted additive angular margin loss for breast cancer classification[J]. Artificial Intelligence Review, 2024, 57(12): 326, doi: 10.1007/s10462-024-10963-4.

[25] Wang P, Wu D, Chen C, et al. Deep adaptive graph clustering via von Mises-Fisher distributions[J]. ACM Transactions on the Web, 2024, 18(2): 1-21.

[26] Chikhi N F. Scientific publications clustering using textual and citation information[J]. Expert Systems with Applications, 2024, 248: 123319, doi: 10.1016/j.eswa.2024.123319.

[27] Zhang R, Guo J, Lan Y, et al. Aggregating neural word embeddings for document representation[C]//Advances in Information Retrieval: 40th European Conference on IR Research, 2018: 303-315.

[28] Xu W, Jiang X, Rao S S H, et al. vONTSS: vMF based semi-supervised neural topic modeling with optimal transport[C]//Findings of the Association for Computational Linguistics, 2023: 4433-4457.

[29] Gretton A, Borgwardt K M, Rasch M J, et al. A kernel two-sample test [J]. Journal of Machine Learning Research, 2012, 13(1): 723-773.

[30] Nan F, Ding R, Nallapati R, et al. Topic modeling with wasserstein autoencoders[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 6345-6381.

[31] Bianchi F, Terragni S, Hovy D. Pre-training is a hot topic: contextualized document embeddings improve topic coherence[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021: 759-766.

[32] Adhya S, Lahiri A, Sanyal D K, et al. Improving contextualized topic models with negative sampling[C]//19th International Conference on Natural Language Processing, 2022: 128-138.

[33] Fang Z, He Y, Procter R. CWTM: leveraging contextualized word embeddings from bert for neural topic modeling[C]//Proceedings of the Joint International Conference on Computational Linguistics, 2024: 4273-4286.

[34] Wu X, Nguyen T, Zhang D, et al. Fastopic: pretrained transformer is a fast, adaptive, stable, and transferable topic model[J]. Advances in Neural Information Processing Systems, 2024, 37: 84447-84481, doi: 10.48550/arXiv.2405.17978.

[35] Schneider J. Efficient and flexible topic modeling using pretrained embeddings and bag of sentences[C]//International Conference on Agents and Artificial Intelligence, 2024, doi: 10.5220/0012404000003636.

[36] Kristensen McLachlan R D, Hicke R M M, Kardos M, et al. Context is key (NMF): modelling topical information dynamics in chinese diaspora media[J]. arXiv preprint arXiv: 2410.12791, 2024.

[37] Kardos M, Kostkan J, Vermillet A Q, et al. Semantic signal separation[J]. arXiv preprint arXiv: 2406.09556, 2024.

[38] Ma Y, Xiao C, Yuan C, et al. CAST: corpus-aware self-similarity enhanced topic modelling[J]. arXiv preprint arXiv: 2410.15136, 2024.

[39] Reuter A, Thielmann A, Weisser C, et al. Probabilistic topic modelling with transformer representations[J]. arXiv: 2403.03737, 2024.