arXiv:2302.01501v2 [cs.IR] 4 Jun 2023

ANTM：一种用于探索演化主题的对齐神经主题模型*

Hamed Rahimi†
索邦大学
巴黎，法国

Hubert Naacke
索邦大学
巴黎，法国

Camelia Constantin
索邦大学
巴黎，法国

Bernd Amann
索邦大学
巴黎，法国

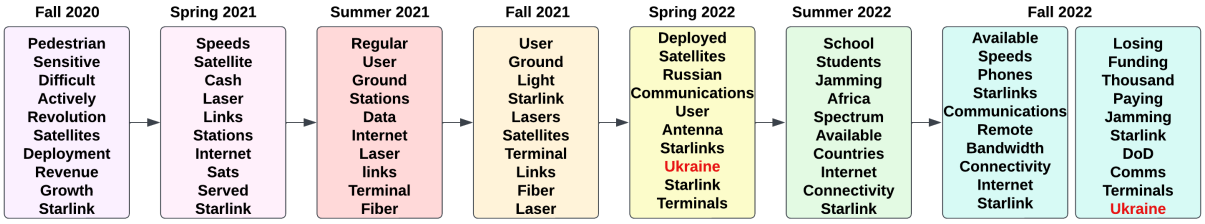


图 1：埃隆·马斯克关于星链的推文演变。“俄罗斯”和“乌克兰”这两个词出现在2022年春季和秋季，恰逢俄罗斯两次袭击乌克兰，以及埃隆·马斯克发推文支持乌克兰人通过星链免费使用互联网。

摘要

本文提出了一系列名为对齐神经主题模型（ANTM）的动态主题模型算法，它们结合了新颖的数据挖掘算法，构建了一个模块化框架，用于发现演化主题。ANTM 利用先进的预训练大语言模型（LLM）从文档中提取感知时间的特征，并采用重叠滑动窗口算法进行序列文档聚类，从而保持演化主题在时间上的连续性。该重叠滑动窗口算法能够识别每个时间段内数量不同的主题，并对齐不同时期语义相似的文档簇。此过程捕捉了不同时期涌现和消退的趋势，并使演化主题的表征更易于解读。在四个不同数据集上的实验表明，ANTM 在主题连贯性和多样性指标方面优于概率动态主题模型。此外，ANTM 易于使用且可适配不同类型的算法，因此提升了动态主题模型的可扩展性和灵活性。此外，我们还开发了一个 Python 软件包，供希望研究大规模文本数据中主题趋势和演化模式的研究人员与科学家使用。

CCS 概念

• 《**信息系统**》 → **文档主题模型**； • **计算方法** → 信息抽取。

关键词

动态主题模型、算法建模、演化主题

ACM参考文献格式：
Hamed Rahimi、HubertNaacke、Camelia Constantin 和 Bernd Amann。2023年。ANTM：用于探索演化主题的对齐神经主题模型。载于 *X* 论文集（*XXXX*）。ACM，纽约，NY，美国，11页。
https://doi.org/XXXXXXX.XXXXXXX

1 引言

主题建模是一种用于自然语言处理的统计技术，旨在从文本文档语料库中发现抽象主题[1,2]。这些模型广泛用于探索性数据分析，以组织、理解和总结大量文本数据 [3]。动态主题模型 [4, 5] 是适用于时间序列档案的一类主题模型。这些模型会随着新文档加入语料库而更新对潜在主题的估计，从而能够识别档案中的演变趋势和模式。主题演化分析 [6, 7] 已应用于各种场景，例如发现科学档案中研究主题和创新的演化过程 [8]，以及理解公众舆论对特定问题的趋势变化 [9]。

利用统计建模从数据中获取洞见，存在两种不同的文化方法[10]。因此，可以从两个视角研究主题模型：概率主题模型和算法主题模型。概率主题模型 [1, 11, 12]假设语料库中的每篇文档都是由多个主题混合而成的，而每个主题都是语料库中各个词语上的概率分布。算法主题模型 [13–16] 则结合多种算法，并

*https://github.com/hamedR96/ANTM
†hamed.rahim@sorbonne-universite.fr

允许为个人或课堂使用而免费制作本作品全部或部分内容的数字副本或纸质副本，前提是这些副本不得为盈利或商业利益而制作或分发，且副本须包含本声明及首页上的完整引文。对于本作品中由 ACM 以外的其他方拥有版权的组成部分，必须尊重其版权。注明出处后允许摘录。若要以其他方式复制、再版、发布到服务器或分发至列表，则须事先获得明确许可和/或支付费用。请向 permissions@acm.org 申请许可。*XXXX*，*2023*年，*X*© 2023年 计算机协会。
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM。.. $15.00
https://doi.org/XXXXXXX.XXXXXXX

利用数值优化技术，将主题表示为从语义相似文档聚类中提取的一组加权词语。本文旨在从算法视角研究动态主题模型，因此将动态主题模型分为两类进行探讨：概率动态主题模型（PDTM）和算法动态主题模型（ADTM）。

概率动态主题模型 [4, 6, 17–19]会为不同时期的不同词语和主题分配概率，从而能够推断数据中潜在的主题和模式如何随时间变化。另一方面，对齐动态主题模型 [14, 20, 21] 利用先进的数据挖掘技术，将词语和文档背后的概率分布形式化为定长向量表示空间。

概率动态主题模型虽广泛应用于各种研究中[22, 23], ，但面对词汇量极大的大型档案时，计算成本会变得很高 [24]。与对齐动态主题模型相比，其可扩展性较差，因为它们需要从复杂的后验分布中进行采样 [25]。此外，概率动态主题模型不适用于处理社交媒体监测（例如 Twitter）等应用中的短数据集 [26], ，而对齐动态主题模型则非常适合高效处理短文本数据 [26–28] ，因为它们通常会利用大语言模型进行特征提取。

然而，当前最先进的概率动态主题模型和对齐动态主题模型存在一些局限性。其中一个局限是，它们被限制为在每个时期维持相同数量的主题，且通常具有相同的全局参数化。因此，有时文档看似属于相同的全局主题，但实际上在特定时间段内属于不同的主题。该问题阻碍了对动态主题时间变化的捕捉，包括随着时间推移主题出现和消退时文档簇数量或每个簇规模的变化。此外，这一局限削弱了特定时期内演化主题的可解释性。这主要是因为主题表示仅源自给定时期内被分配到某个集合中的文档内容。该局限影响了对主题演化的观察，并降低了每个时期所呈现主题的质量。

为解决这一问题，我们提出了 ANTM，这是一种用于算法动态主题建模的新型架构（如图 2 所示），能够考虑动态主题的时间变化，并生成多样且高质量的主题。ANTM 通过一种用于时间文档聚类的重叠滑动窗口算法来发现演化主题。该算法称为对齐聚类，由将档案切分为相互重叠的片段、执行序贯密度聚类，以及对齐不同时间段内相邻且相似的文档簇组成。对齐聚类能够识别跨越多个时间段的一组相似主题，同时这些主题之间又存在足够差异，因而呈现出某种演化。ANTM 利用先进的预训练大语言模型（LLM） [29]，例如 Data2Vec [30],；后者采用标准 Transformer架构，以自监督方式预测文档的潜在表示。我们在四个数据集上开展的实验表明，与现有概率动态主题模型和对齐动态主题模型相比，ANTM 在一致性和多样性分数方面均有显著提升。此外，如图 1 所示，ANTM 能够有效应用于 Twitter 等短文本场景。

贡献。 我们提出 ANTM，这是一类新型动态主题模型，能够利用先进算法有效捕捉主题演化。其次，我们对 ANTM 生成的演化主题进行全面分析，并与最先进的动态主题模型（包括 DETM [19] 和 BERTopic [14],等近期进展，以及传统的概率动态主题模型 DTM [4]）进行比较，以评估其可解释性。第三，我们将展示，ANTM 在一致性和多样性方面优于现有动态主题模型，在处理短文本数据集时尤其出色。

组织结构。本文其余部分的组织如下。第2节将考察当前最先进的方法。第3节介绍 ANTM，第4节展示实验结果。最后，我们将在第6节总结全文。

2 相关工作

对于主题演化的定义存在多种假设，这些假设从统计学角度刻画了动态主题模型。DTM（动态主题模型） [4] 是首批将时间成分纳入主题演化研究的概率动态主题模型之一[17] 。DTM 是一种名为潜在狄利克雷分配（LDA） [11]的生成式概率模型的变体。DTM 创建的主题不仅包含一组词，还包含时间戳或时间范围。它采用贝叶斯方法，建模文档中各主题相对占比随时间的变化。此后又有多种新模型对 DTM 进行了扩展 [6, 18, 31–33]。例如， [6] 中的作者指出，时间本质上是连续的，因此他们提出的模型采用时间戳上的连续分布进行刻画。另一个例子是离散时间动态主题模型（dDTM） [32],，它要求数据具有离散的时间间隔；而连续时间动态主题模型（cDTM） [18]则能够处理任意时间点的数据，不受时间分辨率限制。

尽管 DTM 在当时是研究主题演化的创新工具，但它仍存在一些局限。首先是可扩展性问题：处理非常庞大的数据集时，DTM 的计算成本会很高，耗时也很长。此外，由于 DTM 假设主题会随时间保持相对稳定，因此无法充分捕捉主题随时间变化的多样性。因此，主题在整个数据集中保持一致，人们可能无法检测到主题的细微变化，也难以发现不同时期主题表示方式上的细微差异。此外，建模前必须预先确定主题数量。还有一些生成式概率主题模型 [34–37] 试图克服这些问题。例如，在动态嵌入主题模型（DETM） [19, 38] 中，DTM 与词嵌入相结合，通过提供信息更丰富的词表示来提升主题模型的性能。然而，这种方法仍然非常耗时，而且我们将证明，与对齐动态主题模型相比，其主题表示在多样性和一致性方面的质量较低。另一个例子是，为降低处理大型词汇表时的计算复杂度和变分参数数量， [39]提出了一种基于 DTM 的摊销变分推断方法。此外，也有研究致力于开发可扩展的概率动态主题模型，例如 [40]，将可处理先验的类别从维纳过程扩展到更一般的高斯过程类别。这

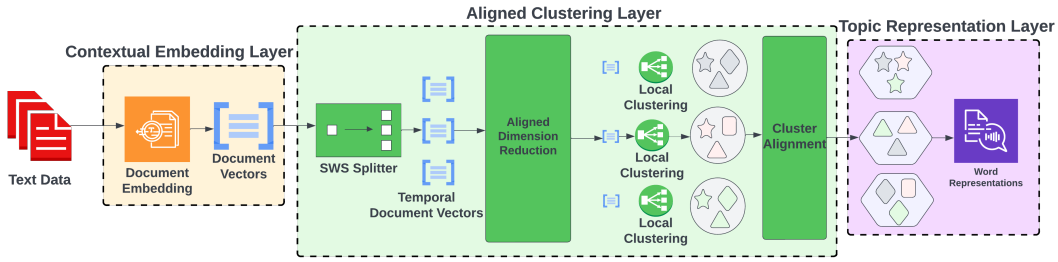


图 2：ANTM 的架构。 第一层使用先进的预训练大语言模型，为每篇文档提供时间感知向量表示。第二层将文档向量表示划分为一组时间帧，并采用重叠滑动窗口算法进行时间文档聚类。最后，第三层负责为每个对齐簇集合随时间提供词表示。

这种方法使模型能够应用于大规模文本集合，并探索在较长时期内平滑演化的主题。

尽管为解决概率动态主题模型的问题做出了所有这些努力，但研究仍表明，与对齐动态主题模型相比，概率动态主题模型在主题表示方面的一致性和多样性较低 [14, 20, 21]。对齐动态主题模型利用神经网络的进展，将词和文档所隐含的时间概率分布形式化为定长向量表示空间。这些方法在保留主题描述中重要词语的同时，生成密集且可解释的文档簇。最广泛使用的对齐动态主题模型包括 BERTopic[14], 。与概率动态主题模型不同，BERTopic 使用生成文档表示的大语言模型（LLM），将上下文嵌入生成的主题中。这些嵌入向量随后被划分为一组簇，用于描述档案中语义相似的一组文档主题。BERTopic 结合 BERT 和 c-TF-IDF [14]，在保留主题描述中重要词语的同时，生成密集且可解释的主题簇。通过在每个时间窗口中计算 c-TF-IDF 主题表示，BERTopic 能够进行动态主题建模。尽管 BERTopic 能够处理大型短文本数据集（得益于大语言模型），但它仍会在整个数据集上生成静态簇，仅以动态方式表示这些簇。这一过程忽略了动态主题模型在各个时期的时间结构。此外，它还限制了主题模型，使其在每个时期都只能包含相同数量的主题。事实上，我们将展示，在整个语料库中静态地对文档进行分组，并基于全局文档簇对其进行动态表示，会影响对主题演化的观察，并降低各时期所表示主题的质量。本文提出的方法比概率动态主题模型更快、可扩展性更强。此外，该方法能够考虑 BERTopic 中讨论的局限性，并生成质量更高的词表示，从而更细致地探索主题演化。

3 ANTM

如图 2 所示，ANTM 由三层组成：第一层使用先进的预训练基于 Transformer 的大语言模型（例如 Data2Vec），为每篇文档提供时间感知向量表示，以反映其内容与其他文档的关系。第二层将文档向量表示划分为一组时间帧，并采用重叠滑动窗口算法（基于 AlignedUMAP[41] 和一种名为 HDBSCAN 的层次化基于密度的聚类算法）进行时间文档聚类。最后，第三层负责为随时间变化的每组对齐簇提供词表示。该层采用基于类别的 TF-IDF 方法。

3.1 上下文嵌入层（CEL）

上下文嵌入层（**CEL**）负责为语料库 D中的文档 d 提供向量表示 y。更正式地说，文档 d 的嵌入 y 是一个映射**CEL**：$d \mapsto y \in \mathrm{R}^z$，用于捕捉语料库 D中的上下文和语义信息。文档嵌入在低维特征向量空间中表示词语和文档，其中嵌入维度 z 预计远小于词汇表的大小（即 D中唯一词语的数量）。**CEL** 利用在大型文本数据语料库上训练的预训练的基于 Transformer 的大语言模型（例如 Data2Vec、GPT4），为每篇文档计算时间感知向量表示。基于 Transformer 的模型通过使用注意力机制考量文档中词语的上下文，来捕捉文档的含义和语境。给定输入文档 d、 d中的词元数量 u，以及$\mathbf{h}j$（ d中第 j个词元的隐藏状态向量），即可通过对 d中所有词元的隐藏状态向量取均值，得到整个文档 d 的嵌入 $y = \mathbf{CEL}(d)$。Transformer 的注意力机制使模型能够关注输入文本的不同部分，并根据每个词语或短语在整篇文档语境中的重要性赋予相应权重。这使**CEL** 能够捕捉文档的含义及其时间维度。

3.2 对齐聚类层（ACL）

虽然我们认同时间本质上是连续的，但这一层的设计基于这样一种观点：正如 [42] 所解释的，主题演化是周期性发生的。因此，我们将这些周期划分为彼此交叠的时间段，与 [4, 7]，类似；但我们使用时间感知算法来考虑时间的连续性，并对随时间出现和消退的演化主题进行建模。在对语料库 D进行基于 Transformer 的文档嵌入后，我们可以通过滑动窗口分割（**SWS**）过程，获得 D 的动态向量表示。该过程将 D 划分为一系列彼此交叠的时间

段 $\{W^1, ..., W^n\}$。随后，我们将时间段 W^t 内发布的文档集合记为 $D^t \subseteq D$，并将 D^t 的嵌入集合记为 $Y^t = \{\mathbf{CEL}(d) | d \in D^t\}$。这一层旨在通过依次将相似文档分组，发现演化文档簇。此过程称为分区聚类（**PC**）。

定义 *3.1*。设 D^t 为文档语料库 D 中第 t 个时间段内的一组文档，并令 Y^t 表示 D^t 的嵌入集合。分区聚类 **PC**：$D^t \mapsto \{D_i^t\}$ 根据文档 D^t 的嵌入 Y_t 对其进行聚类，并返回一组从 $i = 1$ 到 k_t 的局部文档簇 $D_i^t \subseteq D$，其中所有文档嵌入 Y_i^t 彼此相似。

如图 3 和图 4 所示，根据文档嵌入的相似度对文档进行聚类，可以揭示数据中的演化模式和趋势，而仅根据文档内容对其分组时，这些模式和趋势可能并不明显。

为了呈现这些簇的时间演化，我们执行最后一步对齐操作，根据嵌入相似度生成文档簇。结果是一组演化文档簇，其中的文档不仅与同一时间段内的其他文档在语义上相似，也与先前或后续时间段中的其他文档相似：

定义 *3.2.* 设 $\{D_i^t\}$ 为通过 **PC** 获得的局部文档簇集合。簇对齐 **CA**： $D \mapsto \{D_k\}$ 生成从 $k = 1$ 到 m 的 m 个文档子集 $D_k \subseteq D$，其中 D_k 是由一组具有相似嵌入簇 Y_i^t 的局部文档簇（主题） D_i^t 组成的并集。我们将 D_k 称为演化簇（EC），将 D_i^t 称为 D_k 的局部文档簇。我们可以使用不同的簇链接度量（单链接、平均链接、质心链接、全链接）来估计嵌入簇的相似度 Y_i^t（另请注意，由于 D_k 的局部文档簇的嵌入相似，因此 D_k 中的所有文档也相似）。图 5 展示了两个数据集上使用两种不同嵌入方法得到的演化簇示例。

3.2.1 实现。与 ANTM 的另外两个层类似，本层具有很高的灵活性和可扩展性，可采用多种技术和算法实现。如图 2 所示，我们使用基于密度的聚类技术 [43] 实现了所提出的重叠滑动算法。基于密度的聚类算法根据特征空间中点的密度识别数据集中的簇。这些算法特别适用于识别密度和形状各异的簇。它们还容易受到用于计算密度的参数以及数据中噪声或离群点的影响。此外，当簇形状任意且簇的数量事先未知时，这些算法也很有用。与要求输入固定数量主题的非参数PDTM相比，这一特性具有优势。对齐聚类按以下三个步骤构建。

对齐降维。为了克服维数灾难对基于密度的聚类算法造成的限制，有必要对文档嵌入进行降维。所提出的解决方案是采用对齐 UMAP [44], , 一种基于数据帧之间的关系对 UMAP [45] 嵌入序列进行对齐的方法。这使得能够对每个时间段内的文档向量进行局部降维，同时考虑其语义模式，从而确保它们始终处于同一向量空间中并保持连续性。

序列局部聚类。随后，对每个数据帧执行基于密度的聚类算法，将它们依次划分为一组簇。一种有效的方法是 HBD-SCAN [46],，它通过将 DBSCAN [47] 转换为层次聚类算法，再利用一种基于簇的稳定性提取扁平聚类的技术。如图 3 和图 4 所示，这些簇描述了某个时间段内的一个概念。

簇对齐。最后，此步骤将上一步得到的连续簇对齐。由于所有时间段的文档嵌入向量都处于同一空间中，因此我们可以通过考量各簇的质心来对其进行聚合。随后，我们使用 HDBSCAN 聚类技术作为连接方法 [48] ，通过识别质心的簇来对齐不同时期的文档簇。图 5 展示了此步骤的输出示例。

3.3 表示层

表示层负责为文档中每个时间段的每个簇提供词语表示。该层旨在识别能够表征文档集合中特定概念的最相关术语或短语。这些术语或短语可用于概括文档簇的主要思想或主题。

定义 *3.3*。主题表示层（**RL**）会计算一组 m 术语 $\{t_{ij}^r\}_{r=1}^m$，用于描述每个文档聚类 D_{ij} 的语义内容。

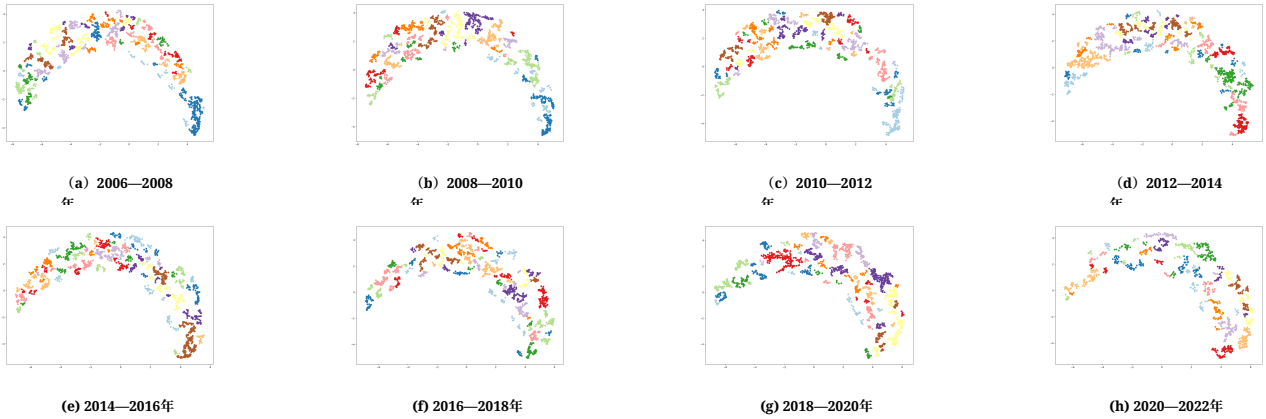
在提取出的词语集合中表示一组文档，有多种方法。一种方法是使用文档聚类中的词与文档联合嵌入，并找出与文档嵌入质心最接近的 m 词语。另一种方法是使用一种称为基于类别的词频-逆文档频率（c-TF-IDF） [14], 的方法。它是词频-逆文档频率（TF-IDF）的一种变体，会将文档的类别标签纳入考量。c-TF-IDF 不仅根据术语在文档中的频率加权，还会考虑这些术语与特定类别的相关性。c-TF-IDF 可确保主题表示中的每个词语至少出现在该主题的一篇文档中。此步骤的输出示例见图 6 和图 7。

4 实验

本实验旨在实现 ANTM，并找出能够生成高质量主题的最优超参数。随后，我们将 ANTM 与基线模型（DTM[4],、DETM[19] 和 BERTopic[14]）进行比较，并展示其在探索性演化分析中的能力。

4.1 数据集

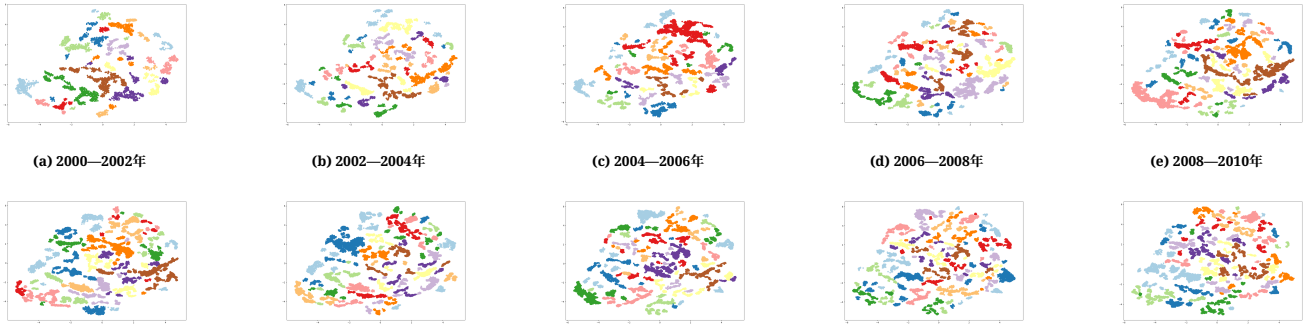
我们的实验使用了四个数据集，分别用于不同目的。第一个数据集是 DBLP [49] 档案库，其中包含 2000 年至 2020 年间发表的 16.8 万篇科学文章（标题和摘要）。我们使用 DBLP 将所提出的方法与 DTM [4] 和 BERTopic [14] 进行比较。第二个数据集是一个包含 1.6 万篇文档（标题和摘要）的样本



(a) 2006—2008年 (b) 2008—2010年 (c) 2010—2012年 (d) 2012—2014年

(e) 2014—2016年 (f) 2016—2018年 (g) 2018—2020年 (h) 2020—2022年

图 3：分区聚类。 arXiv 数据集中的文档使用 Data2Vec 进行嵌入，并在每个 ti 中依次聚类 me 框架。这些聚类依次对齐，以构建演化主题，如图5(a)所示。



(a) 2000—2002年 (b) 2002—2004年 (c) 2004—2006年 (d) 2006—2008年 (e) 2008—2010年

(f) 2010—2012年 (g) 2012—2014年 (h) 2014—2016年 (i) 2016—2018年 (j) 2018—2020年

图4：分区聚类。DBLP数据集中的文档使用 BERT 进行嵌入，并在每个时间段内依次进行聚类。如图5(b)所示，这些簇会依次对齐，从而形成演化主题。

，这些文档从 arXiv [50] 中提取，发表于 2006年年至 2022年年。我们使用该数据集，将采用两种设置（Data2Vec 和 BERT）的 ANTM 与 DETM [19] 和 BERTopic 进行比较。第三个数据集包含埃隆·马斯克于 2017年至 2022 年底发布的 1.9 万条推文 [51]。最后一个数据集包含《纽约时报》于 2012年至 2022年间发表的 21 万篇文章 [52]。我们使用埃隆·马斯克的推文和《纽约时报》数据集，基于所提出的方法对不断演变的主题进行探索性演化分析。数据集的统计信息汇总于表1。

表1：实验所用数据集的统计信息

数据集	文档	词元	词汇表	日期范围
DBLP	168K	18.5M	144K	2000-2020
arXiv	16K	1.17M	12.6K	2006—2022年
EM 的推文	19K	91.5K	4K	2017—2022年
《纽约时报》新闻	210K	5M	61K	2012—2022年

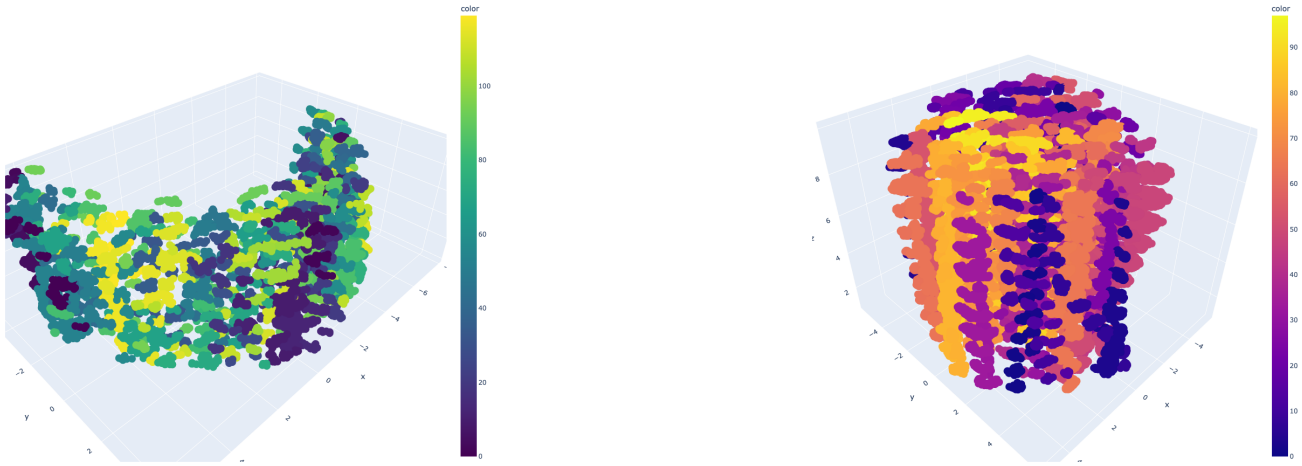
4.2 基线模型

我们将 ANTM 的性能与三种主要的动态主题模型进行比较。第一个模型是 DTM[4],，该模型已广泛应用于各种场景，例如舆情追踪[53]。第二个模型是 DETM [19],，它通过引入词嵌入对 DTM 进行了扩展，以提升主题质量和性能。最后一个模型是 BERTopic[14],，这是一种基于大语言模型的主题模型，采用静态过程对文档嵌入进行聚类，同时动态地表示主题。

4.3 评估指标

由于动态主题模型生成的主题数量并不能充分体现其质量，也不能作为其优于其他模型的充分指标，因此我们使用主题连贯性（TC） [54–56] 和主题多样性（TD） [19, 57]，从人工可解释性和多样性两个方面评估主题模型的性能。

主题连贯性（*TC*）。该指标用于衡量一个包含 m 个词的主题的可解释性，并通过估计一个主题是否高度



(a) Data2Vec | arXiv (b) BERT | DBLP

图 5：演化簇：这些簇（颜色相同）是对每个时间段内的文档进行顺序聚类后得到的结果，并与其相邻的时间段对齐。如图 1、图 7 和图 6 所示，这些演化簇的表示有助于探索主题随时间的演变。

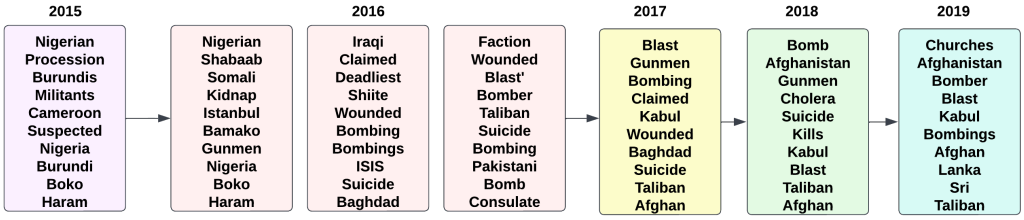


图 6：《纽约时报》关于外国恐怖主义活动的新闻演变。新闻主题从 2015年尼日利亚的博科、哈拉姆，转变为 2016年伊拉克巴格达的伊斯兰国，随后在 2017年到 2019年转向阿富汗喀布尔的塔利班，这一转变颇为有趣。

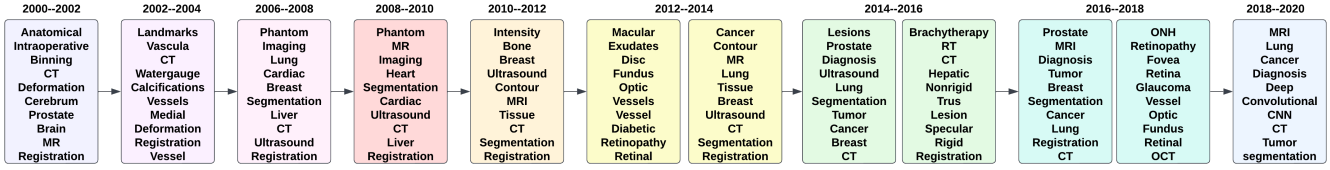


图 7：基于 DBLP 文档的医学计算机科学研究演变。“卷积神经网络”“肺”和“诊断”等词出现在 2018 年至 2020年期间，这与疫情以及计算机视觉研究的进展相吻合。

连贯来进行判断：如果一个主题由倾向于出现在同一文档中的词语表示，那么该主题就是高度连贯的。TC 的定义见 [55]，使用共现值 C_V，是归一化点互信息（NPMI）[58]的一种变体，计算主题 i 中每对词语（t_i^r 和 t_i^s）之间的平均共现值，并对 N 个主题取平均。

主题多样性（*TD*）。主题多样性用于估算给定主题集合中主题表示的多样性。对于此指标，我们采用 建议的唯一词比例（PUW）方法 其定义为词汇量除以主题集合中的词语总数。低多样性主题集合中的主题共享许多词语，而高多样性主题集合中的主题则很少有共同词语。

4.4 实验设置

ANTM 的上下文嵌入层（**CEL**）采用两种嵌入配置：Data2Vec（ANTM-Data2Vec）和 BERT（ANTM-BERT）。为了评估 ANTM-Data2Vec，我们

在 arXiv 数据集的文档摘要上运行了预训练的 Data2Vec（*facebook/data2vec-text-base*[59]）。同样，我们在 DBLP 的标题和摘要、埃隆·马斯克推文的文本内容，以及《纽约时报》文章的描述上运行了 BERT（*all-mpnet-base-v2* [60]），以获取有关数据集语义上下文的文档嵌入，用于评估 ANTM-BERT。在这两种配置中，每个文档都被映射到一个 768 维的稠密向量空间。随后，我们将嵌入向量划分为一系列时间段，以探索文档内容的动态变化。分段步骤的设置信息汇总于表 2。正如 [61], 所建议的，这些分段值全面呈现了数据，并确保分析能够捕捉随时间发生的变化。

表2：分段设置摘要

数据集与时间段大小	长度	重叠	#帧
DBLP 文档	3 年	1 年	10
arXiv 文档	3 年	1年	8
《纽约时报》文章	3年	1年	10
埃隆·马斯克的推文	6个月	2个月	16

下一步是为降维（使用 AlignedUMAP）和序列文档聚类（使用 **PC**）寻找最优超参数。

为此，我们针对 ANTM 的多个参数开展网格搜索研究，以评估每个参数对 DBLP数据集整体性能的独立贡献。在这项研究中，我们谨慎选取以下范围来探索不同方面：UMAP维度设为 [2, 3, 4, 5], , UMAP邻居数范围为 [10, 15, 20, 50, 100], , 而 **PC** 中的最小簇大小则在 [10, 15, 20, 50, 100] 范围内变化。如此全面的选取使我们能够彻底考察这些参数的影响。对于每种组合，我们根据主题质量（主题连贯性×主题多样性）以及生成的簇数量对结果进行评分，因为仅凭主题质量不足以找到最优配置：在少数情况下，某些配置只包含一个多样性和一致性都很高的大型主题。为避免这种情况，我们通过将主题质量按主题数量归一化来定义主题 *score*，如下所示。

$$\text{Score} = \frac{1}{n}\sum_{i=0}^{n-1} \text{TC}_i \times \text{TD}_i \times \frac{T_i}{\text{T}_i^{max}} \tag{1}$$

其中，Score、TC_i和TD_i分别表示时间段 i内的质量分数、主题连贯性和主题多样性。此外，令 T_i表示时间段 i内的主题数量，T_i^{max}表示通过对所有设置组合进行实验所得到的最大主题数量。

图 8(a) 和图 8(b) 分别展示了箱线图，用于说明在 DBLP数据集样本上每个时间段的簇数量（T_i）和主题质量（Q_i）的分布情况（共 4 个切片，每年 5000 份文档，时间范围为 2005 年至 2020年）。最后，我们计算所有时间段的平均质量分数，并根据这一聚合值按升序排列。表 3 展示了本次在 DBLP数据集上开展的研究中得到的得分最高的前 5 组参数设置，揭示了这些参数的重要性。

结果表明，改变 UMAP维度（2、3、4、5）并未对结果产生明显影响。然而，使用 UMAP 大小为 10 且最小聚类大小为 10 的设置获得了最高分数。

因此，我们按如下方式选择参数。对于 AlignedUMAP 的降维设置，我们选择了余弦相似度度量，输出维度为 5。随后，我们在每个时间段上使用欧几里得距离运行 HDBSCAN，并将每个聚类的最小文档数设为 10，从而为各个时期生成一组语义相似的文档簇（图 3 和图 4）。由于所有文档嵌入都位于同一向量空间中，我们可以进一步使用欧几里得距离，对所有文档簇的质心再次运行 HDBSCAN，并将最小文档数设为 2，以便为每个数据集获取演化簇（图 5）。最后，我们使用 c-TF-IDF 为每个簇中的文档表示一组 $m = 10$ 词（图 1、图 6 和图 7）。

基线方法。DTM在DBLP数据集的标题和摘要上分别配置了50个和100个主题。此外，DETM在arXiv数据集的摘要上配置了100个主题和50个训练轮次。DTM和DETM的主题数量根据ANTM以无监督方式生成的主题数量确定。此外，我们在DBLP和arXiv数据集上运行了BERTopic，并使用与所提模型相同的句子转换器模型。对于BERTopic的其余超参数，我们采用默认配置。最后，每个主题模型的词表示数量设为：arXiv 文档为20个，其余数据集为10个。

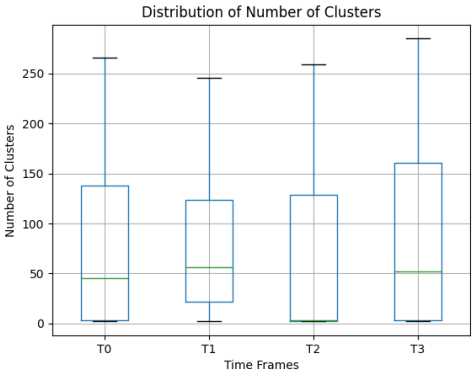
5 结果

基于BERT文档嵌入的所提模型在DBLP数据集上生成了100个演化主题。对于arXiv数据集，分别基于Data2Vec和BERT的所提模型生成了130个和110个演化主题。另一方面，BERTopic在DBLP数据集上生成了500个全局主题，在arXiv数据集上生成了120个全局主题。

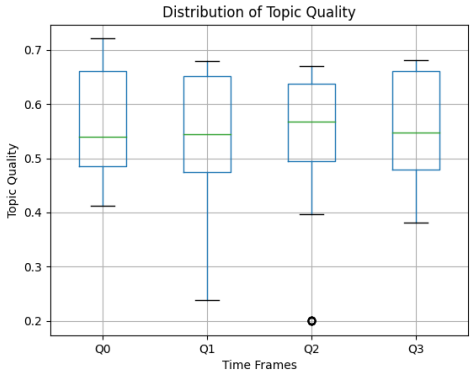
使用主题连贯性和主题多样性指标，我们可以分析所提出模型得到的结果，并从两个角度评估这些模型。首先，我们可以按时间段逐期考察所有主题的时间质量，分析其词表示的一致性和多样性。这项分析比较动态主题模型以多样且连贯的方式描述时间主题的能力。第二个角度是分析各主题所呈现的词表示的质量。这项分析比较动态主题模型随时间探索各动态主题演变的能力。

5.1 分时段质量分析

按时间段分析的目的是确定：对每个时间段分别应用对齐文档聚类，是否比采用全局文档聚类流程的竞争模型 BERTopic 能获得更好的主题质量。如图 9(a) 至 9(c) 所示，首次按时间段进行的评估基于 DBLP数据集；在我们的数据集中，该数据集的文档数量和词汇量最大。图 9(a) 至 9(d) 将应用于 DBLP数据集的 ANTM-Bert 与其他模型进行了比较。在图 9(a) 中，ANTM-BERT



(a) 每个时期的聚类数量



(b) 每个时期的主题质量

图 8：DBLP 样本网格搜索研究分布的箱线图。

表 3：网格　　关于各参数如何影响实验整体性能的搜索研究

UMAP #D	UMAP #N	PC 大小	T_0	T_1	T_2	T_3	Q_0	Q_1	Q_2	Q_3	分数
2,3,4,5	10	10	266	246	259	285	0.41	0.41	0.43	0.42	0.42
2,3,4,5	15	10	231	234	229	234	0.42	0.43	0.43	0.42	0.37
2,3,4,5	50	10	176	162	188	200	0.44	0.45	0.45	0.46	0.31
2,3,4,5	15	15	169	151	147	161	0.48	0.50	0.50	0.50	0.29
2,3,4,5	10	20	131	124	124	165	0.52	0.52	0.54	0.50	0.26

的主题连贯性为 66% 至 70%，高于 DTM 和 BERTopic（两者均低于 63%）。另一方面，在图 9(b) 中，ANTM-BERT 的主题多样性比 BERTopic 低不到 5%，比 DTM 高约 10% 至 15%。总体而言，根据 DBLP数据集，ANTM-BERT 的质量（由 TC 和 TD 的乘积得出）仍高于 DTM 和 BERTopic。这项评估说明了两种 ADTMs 相较于 DTM 的能力；DTM 是一种非基于嵌入的概率动态主题模型。此外，这也验证了我们的观点：使用对齐聚类方法进行局部聚类能够保留更多信息，并能更具描述性地解释时间主题。

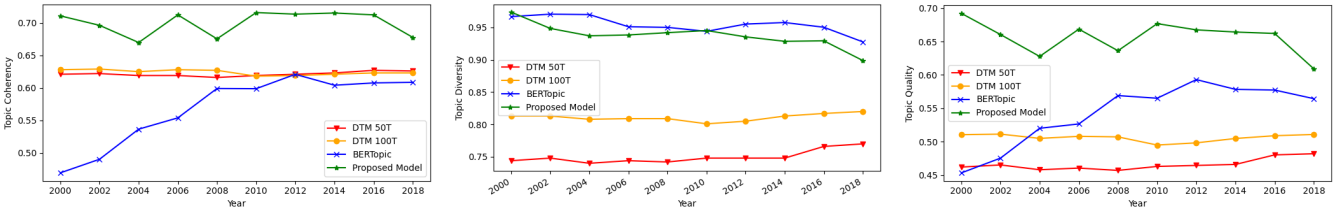
如图 9(d) 至 9(f) 所示，研究在 arXiv 数据集上将 ANTM-Data2Vec 和 ANTM-Bert 与基于嵌入的模型进行了比较。在图 9(e) 中，ANTM 的两种配置生成的主题多样性比 DETM 生成的主题高 25%，但比 BERTopic 获得的主题低约 10%。另一方面，在图 9(d) 中，ANTM-Data2Vec 的主题一致性比 ANTM-BERT 和 BERTopic 的主题高约 20%，但比 DETM 生成的主题低约 15%。由于主题模型的质量是根据主题一致性和主题多样性的综合结果来定义的，即只有在具有一致性的情况下，多样性才有价值，反之亦然，因此，与 BERTopic 和 DETM 相比，ANTM-Data2Vec 的质量分数最高（图 9(f)）。该评估还表明，在 arXiv 上，ANTM-Data2Vec 比 ANTM-BERT 更有效。此外，这进一步支持了以下观点：与采用全局聚类的 ADTMs 相比，对齐聚类方法能够保留更多信息，并且在解释时间主题方面具有更强的描述性。

5.2 主题质量分析

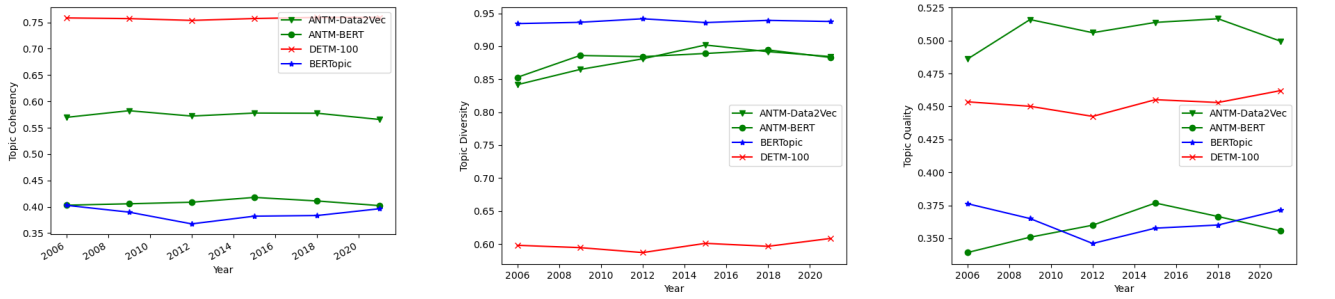
主题级分析的目标是评估一个演化主题是否仍然能够实现多样化，尽管其中包含语义上相近的主题。对于每个模型，我们分别计算每个演化主题的主题连贯性和主题多样性。如图10所示，我们对这些数值进行了排序，并绘制了它们在 DBLP 和 arXiv 数据集上各模型中的分布。这些图包含多项观察结果。

首先，如第2节所述，使用文档和词嵌入会增加演化主题的多样性（图10(a)至10(c)以及图10(e)至10(g)）。基于嵌入的模型具有较高多样性，这可以解释为：演化主题可能会将语义相似但句法不同的词（例如公交车、汽车、自行车）对应的文档重新分组。由于主题多样性基于句法（c-TF-IDF）主题表示，因此这会导致演化主题内部的局部主题表示更加异质，也可能解释了为什么 BERT 和 Data2Vec 得到的演化主题比 DTM（图10(c)和10(d)）以及 DETM（图10(h)）得到的演化主题具有更高的多样性。

另一方面，演变主题中主题的平均主题连贯性则更难解释。首先，我们可以观察到，与其他方法相比，DETM（图 10(h)）实现了非常高且几乎恒定的主题连贯性。这似乎是贝叶斯模型生成过程中使用词嵌入的结果

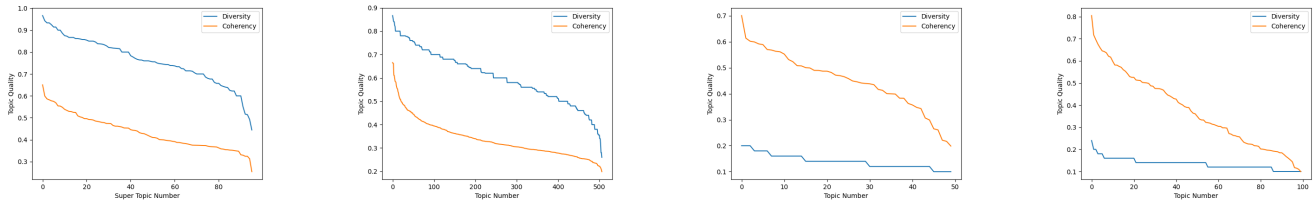


(a) DBLP数据集上的主题连贯性指标 (b) DBLP数据集上的主题多样性指标 (c) DBLP数据集上的主题质量指标

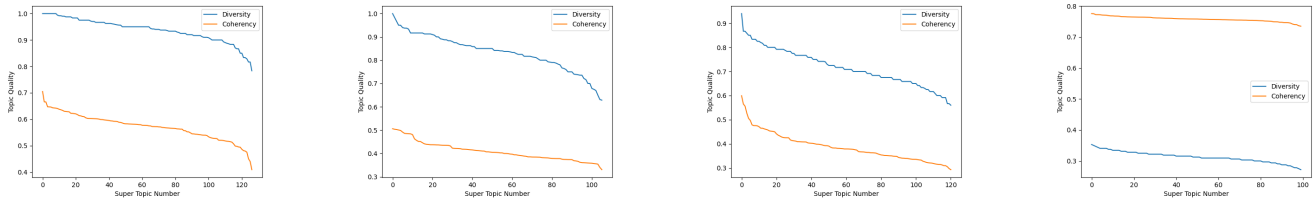


(d) arXiv 数据集上的主题连贯性指标 (e) arXiv 数据集上的主题多样性指标 (f) arXiv 数据集上的主题质量指标

图 9：动态主题模型之间的分时段比较



(a) DBLP 上的 ANTM-BERT (b) DBLP 上的 BERTopic (c) DBLP 上使用 50T 的 DTM (d) DBLP 上使用 100T 的 DTM



(e) arXiv 上的 ANTM-Data2Vec (f) arXiv 上的 ANTM-BERT (g) arXiv 上的 BERTopic (h) arXiv 上使用 100T 的 DETM

图 10：动态主题模型的逐主题质量指标分布

。在基于文档嵌入的方法中，ANTM-Data2Vec 的平均主题连贯性（以及主题多样性）（图 10(e)）优于其他方法。

6 结论

现有动态主题模型通过为动态主题配置全局结构（例如每个时期的文档簇数量相同），忽略了演化主题的某些时间变化。此外，其中一些模型无法扩展到词典规模庞大的大型档案，也无法处理短文本适用于社会网络分析等应用。这些局限性直接影响对主题演化的观察，并降低了序列化表示的演化主题的可解释性和多样性。本文提出了一系列名为对齐神经主题模型（ANTM）的动态神经主题模型算法，该系列结合了新颖的数据挖掘算法，为发现演化主题提供了一个模块化框架。在针对四个不同数据集进行的一系列实验中，与当前最先进的动态主题模型相比，ANTM 在可解释性和主题词多样性方面表现更优。作为未来工作，我们计划

定义一个用于评估主题对齐质量及其检测主题涌现或消退能力的指标。我们还将研究基于嵌入的方法，以便从文档簇中推断出更全面的主题标签。

致谢

我们衷心感谢索邦人工智能中心（SCAI）通过博士研究员奖学金为本研究提供部分经费。我们还要感谢学生 Lucie Chen、Mohamed Allaa Eddine Boutaleb 和 Mouloud Samy Nehlil 提供的协助。本文谨献给伊朗勇敢的女性。

参考文献

[1] D. M. Blei，《概率主题模型》，《ACM通讯》，第55卷，第4期，第77–84页，2012年。[2] R. Alghamdi 和 K. Alfalqi，《文本挖掘中的主题建模综述》，《国际先进计算机科学与应用期刊》（IJACSA），第6卷，第1期，2015年。[3] R. Churchill 和 L. Singh，《主题建模的演进》，《ACM计算调查》，第54卷，第10s期，第1–35页，2022年。[4] D. M. Blei 和 J. D. Lafferty，《动态主题模型》，载于《第23届国际机器学习会议论文集》，第113–120页，2006年。[5] A. Abdelrazek、Y. Eid、E. Gawish、W. Medhat 和 A. Hassan，《主题建模算法与应用：综述》，《信息系统》，第112卷，102131，2023年，第。[6] X. Wang 和 A. McCallum，《随时间变化的主题：一种关于主题趋势的非马尔可夫连续时间模型》，载于《第12届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，第424–433页，2006年。[7] J. Hu、X. Sun、D. Lo 和 B. Li，《使用动态主题模型刻画开发主题的演变》，载于《2015年IEEE第22届软件分析、演进与再工程国际会议（SANER）论文集》，第3–12页，IEEE，2015年。[8] K. Li、H. Naacke 和 B. Amann，《用于探索科学演进的分析型图数据模型与查询语言》，《大数据研究》，第26卷，100247，2021年。[9] H. Sha、M. A. Hasan、G. Mohler 和 P. J. Brantingham，《美国州长和内阁高官关于COVID-19的Twitter叙事动态主题建模》，arXiv预印本 arXiv:2004.11692，2020年。[10] L. Breiman，《统计建模：两种文化（附评论及作者回应）》，《统计科学》，第16卷，第3期，第199–231页，2001年。[11] D. M. Blei、A. Y. Ng 和 M. I. Jordan，《潜在狄利克雷分配》，《机器学习研究杂志》，第3卷，1月号，第993–1022页，2003年。[12] M. Steyvers 和 T. Griffiths，《概率主题模型》，载于《潜在语义分析手册》，第439–460页，心理学出版社，2007年。[13] L. Thompson 和 D. Mimno，《利用上下文化词语表示簇进行主题建模》，arXiv预印本 arXiv:2010.12626，2020年。[14] M. Grootendorst，《BERTopic：采用基于类别的 TF-IDF 流程进行神经主题建模》，arXiv预印本 arXiv:2203.05794，2022年。[15] S. A. Bahrainian、M. Jaggi 和 C. Eickhoff，《自监督神经主题建模》，载于《计算语言学协会论文集：EMNLP 2021》，（蓬塔卡纳，多米尼加共和国），第3341–3350页，计算语言学协会，2021年11月。[16] D. Angelov，《Top2Vec：主题的分布式表示》，arXiv预印本 arXiv:2008.09470，2020年。[17] Y. Teh、M. Jordan、M. Beal 和 D. Blei，《在相关群组间共享簇：层次狄利克雷过程》，《神经信息处理系统进展》，第17卷，2004年。[18] C. Wang、D. Blei 和 D. Heckerman，《连续时间动态主题模型》，arXiv预印本 arXiv:1206.3298，2012年。[19] A. B. Dieng、F. J. Ruiz 和 D. M. Blei，《动态嵌入主题模型》，arXiv预印本 arXiv:1907.05545，2019年。[20] Q. Gao、X. Huang、K. Dong、Z. Liang 和 J. Wu，《语义增强的主题演化分析：动态主题模型与 word2vec 的结合》，《科学计量学》，第127卷，第3期，第1543–1563页，2022年。[21] A. Eklund、M. Forsman 和 F. Drewes，《通过聚类预训练语言模型的嵌入进行动态主题建模：一项研究计划》，载于《计算语言学协会亚太分会第二届会议暨第12届自然语言处理国际联合会议：学生研究研讨会论文集》，第84–91页，2022年。[22] D. Greene 和 J. P. Cross，《使用动态主题建模方法探索欧洲议会的政治议程》，《政治分析》，第25卷，第1期，第77–94页，2017年。[23] F. Yao 和 Y. Wang，《基于动态主题模型追踪城市地理主题》，《计算机、环境与城市系统》，第79卷，101419，2020年。[24] A. Bhadury、J. Chen、J. Zhu 和 S. Liu，“动态主题模型的扩展”，载于《第25届国际万维网大会论文集》，第381–390页，2016年。[25] D. C. Zhang 和 H. Lauw，“面向时间文档网络的动态主题模型”，载于《第39届国际机器学习大会论文集》（K. Chaudhuri、S. Jegelka、L. Song、C. Szepesvari、G. Niu 和 S. Sabato 编），《机器学习研究论文集》第162卷，第26281–26292页，PMLR，2022年7月17日至23日。[26] R. Albalawi、T. H. Yeap 和 M. Benyoucef，“短文本数据的主题建模方法：比较分析”，《人工智能前沿》第3卷，第42页，2020年。[27] D. C. Zhang 和 H. Lauw，“神经主题模型中短文本语义的元补全”，《神经信息处理系统进展》第35卷，第29498–29511页，2022年。[28] J. Feng、Z. Zhang、C. Ding、Y. Rao、H. Xie 和 F. L. Wang，“短文本上的上下文增强神经主题建模”，《信息科学》第607卷，2022年，第91页–79页。[29] M. Chen、J. Tworek、H. Jun、Q. Yuan、H. P. d. O. Pinto、J. Kaplan、H. Edwards、Y. Burda、N. Joseph、G. Brockman 等，“评估基于代码训练的大语言模型”，arXiv预印本 arXiv:2107.03374，2021年。[30] A. Baevski、W.-N. Hsu、Q. Xu、A. Babu、J. Gu 和 M. Auli，“Data2Vec：语音、视觉和语言自监督学习的通用框架”，载于《国际机器学习大会论文集》，第1298–1312页，PMLR，2022年。[31] T. Iwata、T. Yamada、Y. Sakurai 和 N. Ueda，“在线多尺度动态主题模型”，载于《第16届ACM SIGKDD国际知识发现与数据挖掘大会论文集》，第663–672页，2010年。[32] S. A. Bahrainian、I. Mele 和 F. Crestani，“离散动态主题建模”，载于《应用计算研讨会论文集》，第858–865页，2017年。[33] L. Ren、D. B. Dunson 和 L. Carin，“动态层次狄利克雷过程”，载于《第25届国际机器学习大会论文集》，第824–831页，2008年。[34] X. Wei、J. Sun 和 X. Wang，“多时间序列的动态混合模型”，载于 IJCAI，第7卷，第2909–2914页，2007年。[35] T. Iwata、T. Yamada、Y. Sakurai 和 N. Ueda，“在线多尺度动态主题模型”，载于《第16届ACM SIGKDD国际知识发现与数据挖掘大会论文集》，KDD ’10，（美国纽约州纽约），第663–672页，计算机协会，2010年。[36] A. Bhadury、J. Chen、J. Zhu 和 S. Liu，“动态主题模型的扩展”，载于《第25届国际万维网大会论文集》，WWW ’16，（瑞士日内瓦共和国及州），第381–390页，国际万维网会议指导委员会，2016年。[37] E. Zosa 和 M. Granroth-Wilding，“多语言动态主题模型”，《RANLP 2019——自然语言处理：深度学习的世界》，2019年。[38] A. B. Dieng、F. J. Ruiz 和 D. M. Blei，“嵌入空间中的主题建模”，《计算语言学协会汇刊》第8卷，2020年，第453页–439页。[39] F. Tomasi、M. Lalmas 和 Z. Dai，“大型词汇表动态主题建模的高效推断”，载于《第三十八届人工智能不确定性会议论文集》（J. Cussens 和 K. Zhang 编），《机器学习研究论文集》第180卷，第1950–1959页，PMLR，2022年8月1日至5日。[40] P. Jähnichen、F. Wenzel、M. Kloft 和 S. Mandt，“可扩展的广义动态主题模型”，载于《第二十一届国际人工智能与统计会议论文集》（A. Storkey 和 F. Perez-Cruz 编），《机器学习研究论文集》第84卷，第1427–1435页，PMLR，2018年4月9日至11日。[41] “如何使用 AlignedUMAP。”[42] T. S. Kuhn，《科学革命的结构》。芝加哥大学出版社，2012年。[43] A. Ghosal、A. Nandy、A. K. Das、S. Goswami 和 M. Panday，“不同聚类技术及其应用简述”，《建模与图形学新兴技术：IEM Graph 2018论文集》，第69–83页，2020年。[44] M. T. Islam 和 J. W. Fleischer，“流形对齐邻域嵌入”，arXiv预印本 arXiv:2205.11257，2022年。[45] L. McInnes、J. Healy 和 J. Melville，“UMAP：用于降维的统一流形近似与投影”，arXiv预印本 arXiv:1802.03426，2018年。[46] R. J. Campello、D. Moulavi 和 J. Sander，“基于层次密度估计的基于密度的聚类”，载于《知识发现与数据挖掘进展：第17届亚太地区会议，PAKDD 2013，澳大利亚黄金海岸，2013年4月14日至17日，论文集，第二部分17》，第160–172页，施普林格，2013年。[47] M. Ester、H.-P. Kriegel、J. Sander、X. Xu 等，“一种用于在含噪声的大型空间数据库中发现簇的基于密度的算法”，载于 KDD，第96卷，第226–231页，1996年。[48] A. M. Jarman，“层次聚类分析：单链接、完全链接、平均链接和质心链接方法的比较”，佐治亚南方大学，2020年。[49] M. Ley，“DBLP计算机科学文献目录：演变、研究问题与展望”，载于《字符串处理与信息检索：第9届国际

研讨会，*SPIRE 2002*，葡萄牙里斯本，*2002*年*9*月*11*日至*13*日，论文集第*9*卷，第1–10页，施普林格，2002年。[50] C. B. Clement、M. Bierbaum、K. P. O'Keeffe 和 A. A. Alemi，“将 arXiv 用作数据集”，*arXiv* 预印本 *arXiv:1905.00075*，2019年。[51] S. J. H. Shahzad、M. Anas 和 E. Bouri，“加密货币的价格爆炸性与埃隆·马斯克的推文”，《金融研究快报》，第47卷，第102695页，2022年。[52] Y. Pinter、C. L. Jacobs 和 M. Bittker，“Nytwit：《纽约时报》新词数据集”，*arXiv* 预印本 *arXiv:2003.03444*，2020年。[53] Z. Yan 和 X. Tang，“使用动态主题模型探索天涯社区公众意见的演变”，《系统科学与信息学报》，第8卷，第4期，第309–324页，2020年。[54] D. Newman、J. H. Lau、K. Grieser 和 T. Baldwin，“主题连贯性的自动评估”，载于人类语言技术：计算语言学协会北美分会*2010*年年会，第100–108页，2010年。[55] M. Röder、A. Both 和 A. Hinneburg，“探索主题连贯性度量的空间”，载于第八届 *ACM* 网络搜索与数据挖掘国际会议论文集，第399–408页，2015年。[56] H. Rahimi、J. L. Hoover、D. Mimno、H. Naacke、C. Constantin 和 B. Amann，“上下文化主题连贯性指标”，2023年。[57] T. Hashimoto、D. L. Shepard、T. Kuboyama、K. Shin、R. Kobayashi 和 T. Uno，“使用图聚类分析主题多样性的时间模式”，《超级计算期刊》，第77卷，第4375–4388页，2021年。[58] G. Bouma，“搭配抽取中的归一化（逐点）互信息”，*GSCL* 论文集，第30卷，第31–40页，2009年。[59] H. Face，“Hugging Face。”https://huggingface.co，2021年。访问日期：2023年2月1日。[60] N. Reimers 和 I. Gurevych，“Sentence-BERT：使用孪生 BERT 网络的句子嵌入”，*arXiv* 预印本 *arXiv:1908.10084*，2019年。[61] A. Anderson、D. Jurafsky 和 D. McFarland，“迈向 ACL 的计算历史：1980—2008年”，载于*ACL-2012*重发现*50*年成果特别研讨会论文集，第13–21页，2012年。

收稿日期：2023年2月2日；修订日期：2023年6月2日