分类号 TP39 学校代码 10590
UDC 004 密 级 公开

硕士学位论文

面向短文本主题模型的
数据增强与模型增强方法

学位申请人姓名 陈佳耀
学位申请人学号 2100271007
专 业 名 称 计算机科学与技术
学 科 门 类 工学
学院 (部, 研究院) 计算机与软件学院
导 师 姓 名 吴定明副教授

二〇二四年五月

深圳大学

学位论文原创性声明

本人郑重声明：所呈交的学位论文面向短文本主题模型的数据增强与模型增强方法是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律后果由本人承担。

论文作者签名：　　日期：　　年　　月　　日

深圳大学

学位论文使用授权说明

本学位论文作者完全了解深圳大学关于收集、保存、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属深圳大学。学校有权保留学位论文并向国家主管部门或其他机构送交论文的电子版和纸质版，允许论文被查阅和借阅。本人授权深圳大学可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（涉密学位论文在解密后适用本授权书）

论文作者签名：　　导师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

随着社交媒体和电商平台的不断发展，短文本已经成为当今互联网时代重要的信息载体。如何提取和分析这些文本信息一直是业界的关键研究领域之一。主题建模是一个自动将大量文档压缩成内容摘要的过程，这种摘要以多组相关词汇的形式展示潜在的主题。然而，由于短文本的长度有限，使得传统的主题模型在处理这类文本时，往往会挖掘出嘈杂甚至是没有意义的主题。许多研究为短文本主题建模提出了各种策略和方法，但仍无法为其提供充分且可靠的语义信息。针对现有研究的不足之处，本研究提出一种基于大规模语言模型的短文本数据增强方法，通过生成伪长本文来解决短文本的稀疏性问题，并基于这些文本特性构建主题模型。此外，本研究还基于变分自编码器来构建稀疏主题模型，以增强短文本主题模型的表达能力。本文的主要工作包括：

（1）针对主题模型中无法为短文本提供充分且可靠的语义信息这一问题，本论文提出了一种基于提示的短文本扩充方法（IE），利用可以进行文本自动生成的大规模语言模型，通过指令提示将每个短文本扩充为伪长文本。利用这一方法，可以将 LLMs 的知识迁移到短文本主题建模中，而无需人为收集额外的辅助信息。其次，本论文提出一种基于成对文本的主题模型（TPTM）。其假设短文本与伪长文本是一组成对数据，且短文本中的主题来自其对应的伪长文本中的主题。这种假设利用了丰富的单词共现信息和短文本的独特信息，以改进主题建模过程。

（2）针对主题模型的表达能力不足的问题，本论文基于变分自编码器构建了稀疏增强的非均场主题模型（SpareNTM）。该模型基于短文本的特点，在主题模型的生成过程中用引入伯努利辅助变量来进一步建模文档表示的稀疏性。因此，每个文本都将通过相应的伯努利主题选择器，只关注于一小部分主题。此外，SpareNTM 最大的创新点在于，充分利用了变分自编码器的能力实现非均值场近似来估计真实后验，从而保留了隐变量之间的关系。

关键词：短文本，主题模型，数据增强，变分自编码器，数据挖掘

ABSTRACT

With the continuous development of social media and e-commerce platforms, short texts have become an important carrier of information in today's Internet era. How to extract and analyze these text information has always been one of the key research areas in the industry. Topic modeling is a process that automatically compresses a large number of documents into content summaries, which are presented in the form of groups of related words to reveal underlying themes. However, due to the limited length of short texts, traditional topic models often mine out noisy or even meaningless themes when dealing with such texts. Many studies have proposed various strategies and methods for short text topic modeling, but still fail to provide sufficient and reliable semantic information. To address the shortcomings of existing research, this study proposes a short text data augmentation method based on large-scale language models, which solves the sparsity problem of short texts by generating pseudo-long texts and builds topic models based on these text characteristics. In addition, this study also constructs a sparsity-enhanced topic model based on variational autoencoders to enhance the expressive power of short text topic models. The main work of this paper includes:

(1) To address the issue that topic models cannot provide sufficient and reliable semantic information for short texts, this paper proposes an Instruction-based Expansion (IE) method for short texts, which utilizes large-scale language models capable of automatic text generation to expand each short text into a pseudo-long text through instructional prompts. This method allows the knowledge of LLMs to be transferred to short text topic modeling without the need for manually collecting additional auxiliary information. Secondly, this paper proposes a Topic Model based on Paired Texts (TPTM), which assumes that short texts and their corresponding pseudo-long texts are a pair of data, and the topics in short texts are derived from the topics in their corresponding pseudo-long texts. This assumption takes advantage of the rich word co-occurrence information and the unique information of short texts to improve the topic modeling process.

(2) To address the issue of insufficient expressive power of topic models, this paper constructs a Sparsity Reinforced and Non-Mean-Field Topic Model (SpareNTM) based on variational autoencoders. This model, based on the characteristics of short texts, further models the sparsity of document representations by introducing Bernoulli auxiliary variables in the generation process of the topic model. Thus, each text focuses only on a small subset of topics through the corresponding Bernoulli topic selector. Furthermore, the most innovative aspect of SpareNTM is that it fully utilizes the capability of variational autoencoders to achieve non-mean-field approximation to estimate the true posterior, thereby preserving the relationships between latent variables.

Key word: short text, topic model, data augmentation, variational autoencoder, data mining

目　　录

第一章　绪论

1.1　研究背景

随着信息技术和互联网媒体的崛起，如博客、维基百科、社交媒体平台等，文本数据已经成为当代社会信息传播的重要载体。其中，短文本作为信息传播的一种高效形式，其数量在互联网时代经历了爆炸性的增长。短文本通常指的是字数较少、内容简洁的文本数据。它们的主要特点是信息量密集，但表达形式极为简洁。比如在社交平台中，不论是用户发表的微博和小红书，还是标题、弹幕以及评论等，绝大多数都以短文本的形式存在。由于短文本在个人日常交流、商业广告、新闻报道等领域扮演着重要的角色，对短文本进行分析研究不仅对于理解和挖掘网络社会的信息动态具有重要意义，也对于商业智能和公共管理等领域的决策支持具有实际价值。

目前，主题模型仍是一种高效挖掘短文本数据集的常用算法。主题挖掘能够自动发现文本中隐藏的主题，是为了文本信息提取这一任务所设计的一类无监督机器学习技术，使得人们能够更快速且全面地理解文本所包含的内容。该方法通过对文档语料库进行统计分析，能够将包含大量文档的语料库压缩成一个简短的摘要，以揭示语料库中的潜在主题。这个简短的摘要采用主题的形式，即一组相关的词语，因此被称为主题模型。同时，每个文档可以被表示为在这些潜在主题上的一个分布。这种“文档-主题-词”结构提供了语义可解释性，使我们能够更好地理解语料库所包含的主题信息。

基于主题模型的有效性和可解释性，短文本主题挖掘在现实生活中得到了广泛的应用。例如，在商业领域，通过分析消费者的短评和反馈，能够及时捕捉市场趋势和消费者需求，为产品开发和市场营销提供指导。在社交媒体分析中，通过对大量用户发表的短文本进行主题挖掘，可以迅速了解公众情绪、舆论走向或热点话题。自动检测社交媒体上的事件讨论，这在许多不同的领域中都有用途。例如，在国家安全行动中用于预测骚乱并检测虚假信息**DisruptiveEventDetection**，而人为地监控社交媒体中的反社会行为和反事实行为是十分受限制的。利用主题模型，可以深入挖掘和利用海量短文本背后有价值的信息。

1.1.1　短文本特征概述

随着社交媒体和电商平台的不断发展，短文本已经成为当今互联网时代常见的一类文本。短文本的长度，少则几个词，多则十几个词，与常规长度的文本相比存在相当大的差别。传统的书籍和科学文章存在着具有辨识度的单词共现模式，而短文本可能拥有非常不同的模式[survey_2023]。短文本数据，尤其是互联网中的短文本，具有如下的特征：（1）稀疏性：社交媒体短文本不同于编辑精良的文章，往往只包含少数匆忙输入的单词。因此短文本的长度十分有限。此外，社交媒体数据的词汇不断演变，新词和标签不断涌现。帖子中包含多种语言并不罕见，缩写更是常规。在社交媒体帖子中，相较于长文本，高频的单词共现模式几乎不存在。（2）语境缺失：由于篇幅限制，短文本往往缺乏足够的上下文信息来清晰地表达其含义。这使得理解和分析短文本的语义内容更加困难。（3）动态性和时效性：社交媒体等平台上的短文本往往与当前事件、趋势和公众关注点密切相关，展现出强烈的动态性和时效性。（4）潜在主题的复杂性：单个短文本所包含的信息有限，但文本与文本之间的联系可以揭示复杂且多样的主题信息。

1.1.2　短文本主题模型的挑战

由于主题模型往往依赖于文档层面的单词共现信息来推断潜在主题[STTM]，但由于短文本噪声多、数据量大、文本长度短等特征，要找到短文本数据集中的词共现性信息是相当有挑战性的。即使是最先进的（针对常规文本的）主题模型，在应用于短文本时，往往也会挖掘出嘈杂甚至是无法解释的主题。因此，短文本主题模型的核心问题在于短文本可使用的单词共现信息相对较少。如果能够使模型利用更为充分和可靠的语义信息，就有望有效提升主题模型在短文本集上的效果。因此已有的研究工作大致可分成两个类别[ZJHNY]，一类方法基于短文本的特性，引入先验知识以增加词语共现信息，从而缓解稀疏性问题。另一类方法借助外部知识引入除词语共现信息之外的语义信息，与词语共现信息在建模过程中形成互补。

第一类方法引入适用于短文本特性的假设来缓解稀疏性问题。这类方法通常利用对短文本集的先验知识，设计特定的模型以更好地适应短文本的特征。例如，考虑到短文本的长度有限，一个短文本的内容可能仅涉及有限的主题。利用这种先验知识设计模型，可以使主题模型更贴近短文本集的特性。然而，这些基于先验知识的主题模型主要依赖传统的统计推断方法，如变分推断和 Gibbs 采

样。随着主题模型结构变得复杂，这些推断方法的复杂度显著增加。同时，在处理大规模文本集时，难以有效扩展或利用 GPU 等并行计算设备[NTMsurvey]。

第二类方法引入外部语义信息，让模型能够同时使用词语共现信息和其他语义信息。以微博为例，元信息如标签、作者、地点和时间戳等可以用于将短文聚合成长文，再对将传统的主题模型应用于这些短文上。然而，由于元数据的有限外部来源，这些方法未能达到预期的结果。同时聚合策略会减少文档数量，进而导致后续模型在建模时会面临文档数量有限这一问题。另一种直观的想法是借助外部语料将短文本扩充为长文本。但是，短文本集和扩充后得到的文本集需要在语义上尽量保持一致，否则这一做法会引入噪声，让主题模型的结果更差[ZJHNY]。

综上，尽管许多研究提出了各种短文本主题模型以应对短文本的稀疏性，但是这些方法仍存在着适用性的问题以及计算时间复杂度较大的问题。下面本文将更详细的介绍已有的研究工作。

1.2 国内外研究现状

国内外研究者从多个方面展开工作来解决短文本主题挖掘存在的问题，这些工作主要分为两大类：第一类方法是引入适用于短文本特性的假设，从而增加词共现信息对短文本进行建模。第二类方法引入外部语义信息，让模型能够同时使用词语共现信息和其他语义信息。

1.2.1 引入适用于短文本特性的假设

这类方法可以分成四种策略[STTM,XSJMXQQ]，分别是基于狄利克雷多项混合（Dirichlet Multinomial Mixture，DMM）的模型，基于稀疏先验假设的模型，基于全局词共现信息的模型，以及基于文本自聚合的模型。

（1）基于 DMM 的模型

DMM[DMM]模型最初 Nigam 等人由提出，已被广泛应用于推断短文本中的潜在主题。该模型基于一种简单的假设策略，即一个短文本仅包含一个潜在主题。相较于 LDA 模型假设一个文本包含多个主题，这更适用于短文本。DMM 模型设计之初采用的是基于 Expectation－Maximization（EM）的推断算法。而 GSDMM[GSDMM]提出了一个基于吉布斯采样（Gibbs Sampling）的 DMM 模型。许多方法都集中在利用词的嵌入表示来检索语义信息来进一步降低了稀疏度。

PDMM[PDMM]模型认为主题数应该服从于泊松分布，设计了一个基于泊松分布和词嵌入的 DMM 模型。GPU-DMM[GPUDMM]模型和 GPU-PDMM[PDMM]模型基于广义波利亚罐（Generalized Pólya Urn, GPU）模型和词嵌入，利用语义相似的词去提升 DMM 模型的采样过程。GPM[GPM]采用伽马分布和泊松分布改进 DMM，但在处理复杂短文本时性能有限。LF-DMM[LFDMM]提出了基于潜在特征向量的 DMM 模型，通过改进特征词表示并引入词-主题映射来提高模型性能。Lap-DMM[LAPDMM]提出了一种带有变分流形正则化的 DMM 主题模型，以提高主题分类准确率。然而，它基于文档之间的相似程度，具有一定的复杂性。MultiKE-DMM[MKEDMM]模型利用知识图谱和词嵌入增强了 DMM 模型的采样过程。APU-DMM[APUDMM]则通过自适应调整 GPU-DMM 的提升权重获得了更好的性能。

（2）基于稀疏先验假设的模型

稀疏主题模型对于主题数的假设区别于 DMM 和 LDA 模型。通常来说，每一个文本的内容应该关注在一小部分主题上，而不是仅仅一个主题或者所有的主题这样的极端情况。Dual-Sparse[DualSparse]模型通过“Spike and Slab”[SpikeSlab]先验限制短文本对应的主题数量以及每个主题包含的词汇数。稀疏主题编码[STC]（STC）利用拉普拉斯先验直接控制文档主题分布的稀疏性。随着深度学习的发展，研究者们也将目光转向利用神经网络实现主题模型。NSTC[NSTC]联合利用词嵌入和神经网络实现 STC 模型。Sparsemax-NVDM[SparseMax,NVDM]利用变分自编码器[VAE]（Variational AutoEncoder，VAE）和 Sparsemax 激活函数构建主题模型，并且限制了一个短文本只能对应少数主题。CRNTM[CRNTM]模型在神经网络主题模型的解码阶段用高斯分布引入了辅助数据集的词嵌入信息，并且利用贝塔分布限制了短文本对应的主题数量。DVAE.Sp[DVAE]模型，直接利用 Sigmoid 激活函数函数选择与当前文本相关的主题。NQTM[NQTM]模型则基于 VQ-VAE[VQVAE]（Vector Quantised Variational AutoEncoder）构建了稀疏的文档主题分布。TSCTM[TSCTM]在 NQTM 的基础上加入了对比学习的思想以增强模型的表现。

（3）基于全局词共现信息的模型

为了解决词共现信息是不足的问题，一些模型尝试利用原始数据集中丰富的全局词共现信息来推断隐藏的主题。全局词共现可在一定程度上缓解短文本稀疏性问题。这些模型需要配置一个滑动窗口来提取词共现。这种类型的模型可以根据全局词共现的利用策略分为两类。第一类可以通过利用全局词共现信

息推断潜在主题。例如，BTM[BTM]模型假设构成一个双词的两个词具有相同的主题，这个主题是从整个数据集中的各种主题中得出的。而这两类中的第二类，比如 WNTM[WNTM]则基于全局词共现创建一个词共现网络，然后从构建的网络中找出隐藏的主题，其中每个单词都代表构建网络的一个节点。

然而，BTM 可能会丢失一些在语料库中无法观察到的显著且连贯的词共现信息。它还容易受到噪声干扰，提取出许多不相关的双词。LS-BTM[LS-BTM]利用潜在语义细节进行主题提取，并改善 BTM 的性能。然而，该模型使用了更多与主题不相关的双词。R-BTM[R-BTM]通过使用词嵌入的相关词相似性列表来克服 BTM 模型连贯性消失的问题。NBTMWE[NBTMWE]结合了来自外部语料库的噪声 BTM 和词嵌入技术，以改善主题的连贯性。UGTM[UGTM]基于上下文数据的语义关系开发了用户图主题模型。这种方法在动态主题提取方面非常高效。GLTM[GLTM]基于全局和局部词嵌入进行主题建模，该模型使用连续 Skip-Gram 模型与负采样的适当编码来训练全局嵌入，以获取局部词嵌入。CSTM 是一种共同语义主题模型，通过使用单词来过滤短文本主题发现中的噪音。但是，这个模型在设置优先级和确定主题标签数量方面存在局限性。

（4）基于文本自聚合的模型

基于自聚合的模型增加了一个新的隐变量，长文本，然后构造了一个短文本－长文本－主题－词语的联合概率分布。只要让长文本的数量少于短文本的总数，就能够让长文本成为短文本的一个聚类。其好处是在主题推理过程中同时进行主题建模和文本自我聚类。SATM[SATM]模型是最早提出的自聚合方法，它将每个短文视为隐藏的长篇伪文档的样本，并将其合并，使用吉布斯采样进行主题提取。优点是不依赖元数据或辅助信息，但却很容易过度拟合，而且计算成本也很高。为了提高 SATM 模型的性能，PTM[PTM]模型和 SPTM[SPTM]模型提出了伪文档的概念，将短文隐性地结合起来，以解决数据稀少的问题。PYSTM[PYSTM]模型通过狄利克雷过程采样伪长文本的数量，实现伪长文本数量的参数化。SenU-PTM[SenU-PTM]根据词与词嵌入的语义相似性生成短语，并用新的词汇表对原始语料进行标注，接着根据原始文本和语义关系生成具有共现关系的意义单元。

1.2.2　引入外部语义信息

一些研究把目光转向了除词共现信息以外的语义信息。一种直观的方法是利用外部知识将短文本进行扩充变成长文本，从而增加单词共现信息。目前

关于短文本主题建模的文档扩展方法，大多数侧重于扩展 Twitter 数据。ET-LDA[ExTwitter]提出了基于推文作者或语料库词汇中的聚合方案。考虑到元信息，DLDA[DualLDA]使用由推文中 URL 链接的网页作为元长文档，以识别推文中更好的主题。然而，在某些领域中可能无法获得有利的元数据。Pooling[PoolTwitter]评估了四种推文汇总方案，以改善 LDA 的结果。DREx[DREx]提出了共频扩展（CoFE）和基于分布表示的扩展（DREx），将短文本扩展为一个可观的伪文档。AOTM[AOTM]提出了一种新的主题模型，用于从短的用户评论文本和普通文本中提取主题。通过考虑作者身份，AOTM 为每个短文本的作者提供了一个概率分布，该分布覆盖了一系列仅由短文本示例化的主题。一些短文本主题模型就试图利用上下文语义信息，从而弥补稀疏的共现信息。SeaNMF[SeaNMF]模型通过非负矩阵分解的方式，把短文本集的词嵌入引入模型中。RLSeaNMF[RLSeaNMF]模型构造了一个结合强化学习和 SeaNMF 的模型。COTM[COTM]模型面向博客及其评论数据，将主题分成标准主题和非标准主题，把常规文本集的语义信息迁移到短文本集中去。ASTM[ASTM]模型和 AATM[AATM]模型则引入辅助文本集的词语共现信息，并将模型与注意力机制结合起来。

1.3 本文解决的主要问题

近年来，许多研究针对主题模型提出了各种策略和方法来解决短文本数据的稀疏性问题。但是这些方法要么无法为主题模型提供充分且可靠的语义信息，要么存在适用性问题。在本论文中，我们主要解决了以下问题：

（1）为了给主题模型提供充分的语义信息，一种直接的方法是从短文本自身出发，对其进行扩充以实现数据增强的目的，从而增加文档层面的词共现信息。但现有的文本扩充方法往往无法保留文本原来的语义信息，即扩充之后的“伪”长文与原来的短文之间存在语义不一致的问题。我们的方法则可以保证这种语义一致性。此外，不同于抛弃原始短文本，只利用扩充文本的主题模型，我们的模型将短文本与扩充文本联合建模，既利用伪长文本的词共现信息，又注重原始短文本中的独特信息，实现了高质量的主题挖掘。

（2）在一个大的主题集合中，一个文本往往只涉及其中的小部分主题。稀疏主题模型针对这一事实，引入了额外的正则化约束项或稀疏先验分布，以增强短文本主题模型的表达能力。然而，这些主题模型在引入辅助变量时，往往以牺牲

部分准确性为代价，假设隐变量的后验分布之间是无关的，以简化目标函数的推导过程，即平均场假设[VI]。但在传统的非深度学习框架下，这些模型的推导和求解仍旧复杂。我们的模型则利用变分自编码器构建了稀疏增强的非均场主题模型，并且利用了非均值场假设对后验分布进行建模解决了上述的问题，从模型增强的角度实现了高质量的主题挖掘。

1.4　论文主要研究内容

1.4.1　短文本主题模型的数据增强方法

数据增强的目的是解决短文本的长度有限和词共现信息匮乏的问题。论文将从两个方面展开研究：首先是短文本扩充技术，旨在生成高质量的扩充文本，称为伪长文本。其次是研究短文本与伪长文本之间的联合建模方法，以提升短文本主题建模的有效性。具体的研究内容如下：

（1）**短文本扩充技术**：文本扩充是一种常用的数据增强方法，通过依据某个数据模型，从原始文本计算得到符合该模型的目标系统所需的文本数据。随着在主题建模任务中，文本扩充技术已被证明能显著提高模型性能。然而，在实际应用中，由于外部知识的有限性，可能难以得到高质量的文本数据以提升模型性能。为解决这一问题，本论文将研究基于大语言模型的短文本扩充技术，利用大语言模型强大的外部知识进行文本生成，从而获得与短文本语义尽可能一致的伪长文本数据。

（2）**短文本与伪长文本的联合建模**：以往的研究通常直接利用针对常规文本的主题模型对扩充后的文本进行主题建模。然而，文本生成过程中不可避免的会引入额外的主题信息，从而影响主题挖掘的效果。为了应对这个问题，本论文将研究一种短文本与伪长文本联合的方法，既利用伪长文本的词共现信息，又注重原始短文本中的稀疏信息，从而更好挖掘出短文本中的主题信息。

1.4.2　短文本主题模型的模型增强方法

模型稀疏性增强的目标是使模型学习到更具语义明确和清晰的潜在表达，以在一定程度上缓解短文本词共现信息的稀疏问题。本研究主要集中在以下两个主要方面进行研究：稀疏主题模型的设计以及如何利用神经网络实现稀疏主题模型的变分推断算法。具体的研究内容如下：

（1）**稀疏主题模型设计**：为了增强文本表示的稀疏性，一种直接的方法是生

成峰值分布，使每个短文本只关注少数几个主题。为实施表达的稀疏性约束，往往通过引入了额外的正则化约束项或额外的先验变量来实现。然而，在实际应用中，由于神经网络的主要优化算法是梯度反传算法。而主题模型涉及到分布的采样，并不是所有的分布采样过程都能够得到有效的梯度计算。为了解决这个问题，将研究适合的稀疏先验，与经典的主题模型结合以使后续的深度学习方法能够进行推断。

（2）**基于非均值场的变分推断算法：**神经主题模型通常使用变分自编码器得到模型的参数估计然而，当引入额外的变量时，这些模型普遍基于均值场理论，假设隐变量之间具有强独立性，以降低其复杂的理论推导过程。但在实际的应用过程中，隐变量之间存在着一定的关系。为了简化推导过程而忽略隐变量之间的关系，将难以得到高质量的主题。针对以上问题，本论文将研究非均值场假设的神经变分推断算法，通过使用神经网络来建模稀疏主题模型的生成过程，从而简化复杂的推断过程。

1.5　论文组织结构

本文分为五章，具体组织安排如下：

第 1 章介绍了短文本主题模型的研究背景及意义、短文本主题模型的研究现状，分析了现有短文本主题模型存在的问题和挑战，介绍说明了本文的研究目标、研究内容与核心贡献。

第 2 章介绍了相关技术与理论，包括主题模型的概述以及经典的适用于常规文本的 LDA 模型和适用于短文本的 DMM 模型，最后介绍了吉布斯采样算法以及基于变分自编码器的主题模型。

第 3 章介绍了一种短文本数据增强的方法，基于引导的短文本扩充方法，该方法将短文本扩充为（伪）长文本。同时提出了提出一种将伪长文本与短文本视为成对数据的主题模型 TPTM。

第 4 章介绍了一种基于稀疏性增强的短文本主题模型增强方法，基于非均值场推理的神经稀疏主题建模方法 SpareNTM。

第 5 章介绍了本论文的研究内容总结，并且对本文研究内容做出展望。

附录 A　IETM 的吉布斯采样公式推导

本附录将为 IETM 的吉布斯采样提供推导细节。IETM 模型的联合分布函数如公式 (1.1) 所示：

$$p(\mathcal{D},\vec{l}|\vec{\alpha},\vec{\eta})=p(\mathcal{D}|\vec{l},\vec{\eta})p(\vec{l}|\vec{\alpha})=p(\mathcal{D}|\vec{l},\vec{\eta})\cdot\prod_{d=1}^{D}p(\vec{l}_d|\vec{\alpha}) \tag{1.1}$$

其中 $\vec{l}=\{\vec{l}_d\}_{d=1}^{D}=\{\vec{z}^+,\vec{z}\}=\{\vec{z}_d^+,\vec{z}_d\}_{d=1}^{D}$。首先，我们可以推导出

$$p(\vec{l}_d|\vec{\alpha})=p(\vec{z}_d|\vec{z}_d^+)\int p(\vec{z}_i^+|\vec{\theta}_d)p(\vec{\theta}_d|\vec{\alpha})\mathrm{d}\vec{\theta}_d=\frac{\Delta(\vec{n}_{P_d}+\vec{\alpha})}{\Delta(\vec{\alpha})}\cdot\prod_{k=1}^{K}\left(\frac{n_{P_d}^{(k)}}{N_d}\right)^{n_{S_d}^{(k)}} \tag{1.2}$$

其中，$\vec{n}_{P_d}=\{n_{P_d}^{(k)}\}_{k=1}^{K}$ 和 $\vec{n}_{S_d}=\{n_{S_d}^{(k)}\}_{k=1}^{K}$。$n_{P_d}^{(k)}$ 和 $n_{S_d}^{(k)}$ 分别是第 d 个伪文档和原始文档中属于第 k 个主题的词的数量。在这里，我们采用了 Δ 函数，如下所示：

$$\Delta(\vec{\alpha})=\frac{\prod_{k=1}^{K}\Gamma(\alpha)}{\Gamma(\sum_{k=1}^{K}\alpha)} \tag{1.3}$$

$$\Delta(\vec{n}_{P_d}+\vec{\alpha})=\frac{\prod_{k=1}^{K}\Gamma(n_{P_d}^{(k)}+\alpha)}{\Gamma(\sum_{k=1}^{K}n_{P_d}^{(k)}+\alpha)}=\frac{\prod_{k=1}^{K}\Gamma(n_{P_d}^{(k)}+\alpha)}{\Gamma(N_d+K\alpha)} \tag{1.4}$$

类似的，我们可以改写 $p(\mathcal{D}|\vec{l},\vec{\eta})=p(\mathcal{S}|\vec{z},\vec{\eta})p(\mathcal{P}|\vec{z}^+,\vec{\eta})$ 为

$$p(\mathcal{D}|\vec{l},\vec{\eta})=\prod_{i=1}^{W}\beta_{z_i}^{(w_i)}\cdot\prod_{j=1}^{W^+}\beta_{z_j^+}^{(w_j)}=\prod_{k=1}^{K}\left(\prod_{\{i:z_i=k\}}\beta_k^{(w_i)}\cdot\prod_{\{j:z_j^+=k\}}\beta_k^{(w_j)}\right)=\prod_{k=1}^{K}\prod_{v=1}^{V}(\beta_k^{(v)})^{n_k^{(v)}} \tag{1.5}$$

其中，W 和 W^+ 分别是 $\mathcal{S}$ 和 $\mathcal{P}$ 中的词数，$n_k^{(v)}$ 是分配给 $\mathcal{D}$ 中第 k 个主题的词 v 的出现次数。然后，通过对 $\vec{\beta}$ 积分，我们可以得到

$$p(\mathcal{D}|\vec{l},\vec{\eta})=\prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\vec{\eta})}{\Delta(\vec{\eta})} \tag{1.6}$$

$$\Delta(\vec{\eta})=\frac{\prod_{v=1}^{K}\Gamma(\eta)}{\Gamma(\sum_{v=1}^{V}\eta)} \tag{1.7}$$

$$\Delta(\vec{n}_k+\vec{\eta})=\frac{\prod_{v=1}^{V}\Gamma(n_k^{(v)}+\eta)}{\Gamma(\sum_{v=1}^{V}n_k^{(v)}+\eta)}=\frac{\prod_{v=1}^{V}\Gamma(n_k^{(v)}+\eta)}{\Gamma(n_k+V\eta)} \tag{1.8}$$

其中 $\vec{n}_k=\{n_k^{(v)}\}_{v=1}^{V}$，且 $n_k=\sum_{v=1}^{V}n_k^{(v)}$。现在联合概率分布 Eq.(1.1) 变成:

$$p(\mathcal{D},\vec{l}|\vec{\alpha},\vec{\eta})=\prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\vec{\eta})}{\Delta(\vec{\eta})}\cdot\prod_{d=1}^{D}\left[\frac{\Delta(\vec{n}_{P_d}+\vec{\alpha})}{\Delta(\vec{\alpha})}\cdot\prod_{k=1}^{K}\left(\frac{n_{P_d}^{(k)}}{N_d}\right)^{n_{S_d}^{(k)}}\right] \tag{1.9}$$

接下来需要求解两个条件后验概率分布：（1）为伪文档 $\boldsymbol{P}_d$ 中的词 $w_{d,n}^+$ 采样一个主题 $z_{d,n}^+$ 的条件后验概率分布；（2）对于原始文档 $\boldsymbol{S}_d$ 中的词 $w_{d,n}$，将采样一个主题 $z_{d,n}$ 条件后验概率分布。对于（1），我们有

$$
\begin{aligned}
p(z^{+}_{d,n}=k|\vec{\boldsymbol{l}}_{\neg(P_{d,n})},\mathcal{D}) &= \frac{p(\vec{\boldsymbol{l}},\mathcal{D})}{p(\vec{\boldsymbol{l}}_{\neg(P_{d,n})},\mathcal{D})} \propto \frac{p(\vec{\boldsymbol{l}},\mathcal{D})}{p(\vec{\boldsymbol{l}}_{\neg(P_{d,n})},\mathcal{D}_{\neg(P_{d,n})})} \\
&= \frac{\Delta(\vec{n}_k+\vec{\eta})}{\Delta(\vec{n}_{k,\neg(P_{d,n})}+\vec{\eta})}\cdot\frac{\Delta(\vec{n}_{P_d}+\vec{\alpha})}{\Delta(\vec{n}_{P_d,\neg(P_{d,n})}+\vec{\alpha})}\cdot\prod_{j=1}^{K}\left(\frac{N_d-1}{N_d}\cdot\frac{n^{(j)}_{P_d}}{n^{(j)}_{P_d,\neg(P_{d,n})}}\right)^{n^{(j)}_{S_d}} \\
&\propto \frac{n^{(v)}_{k,\neg(P_{d,n})}+\eta}{\sum_{i=1}^{V}(n^{(i)}_{k,\neg(P_{d,n})}+\eta)}\cdot\frac{n^{(k)}_{P_d,\neg(P_{d,n})}+\alpha}{N_d-1+K\alpha}\cdot\left(\frac{N_d-1}{N_d}\cdot\frac{n^{(k)}_{P_d,\neg(P_{d,n})}+1}{n^{(k)}_{P_d,\neg(P_{d,n})}}\right)^{n^{(k)}_{S_d}}
\end{aligned}
\tag{1.10}
$$

其中 $\vec{\boldsymbol{l}}=\{\vec{\boldsymbol{l}}_d\}_{d=1}^{D}=\{\vec{\boldsymbol{z}}^{+},\vec{\boldsymbol{z}}\}=\{\vec{\boldsymbol{z}}^{+}_d,\vec{\boldsymbol{z}}_d\}_{d=1}^{D}$。$n^{(k)}_{P_d}$ 和 $n^{(k)}_{S_d}$ 分别是第 d 个伪文档和原始文档中属于第 k 个主题的词的数量。而 $n^{(v)}_{k}$ 是分配给 $\mathcal{D}$ 中第 k 个主题的词 v 的出现次数。所有带有 $\neg\bullet$ 的计数表示排除来自 $\bullet$ 的计数。类似地，对于（2），原始文档 S_d 中的词 $w_{d,n}$，其采样一个主题 $z_{d,n}$ 的条件后验概率分布为公式为：

$$
\begin{aligned}
p(z_{d,n}=k|\vec{\boldsymbol{l}}_{\neg(S_{d,n})},\mathcal{D}) &= \frac{p(\vec{\boldsymbol{l}},\mathcal{D})}{p(\vec{\boldsymbol{l}}_{\neg(S_{d,n})},\mathcal{D})} \propto \frac{p(\vec{\boldsymbol{l}},\mathcal{D})}{p(\vec{\boldsymbol{l}}_{\neg(S_{d,n})},\mathcal{D}_{\neg(S_{d,n})})} \\
&= \frac{\Delta(\vec{n}_k+\vec{\eta})}{\Delta(\vec{n}_{k,\neg(S_{d,n})}+\vec{\eta})}\cdot\frac{\Delta(\vec{n}_{P_d}+\vec{\alpha})}{\Delta(\vec{n}_{P_d}+\vec{\alpha})}\cdot\prod_{j=1}^{K}\left(\frac{n^{(j)}_{P_d}}{N_d}\right)^{n^{(j)}_{S_d}}\Big/\left(\frac{n^{(j)}_{P_d}}{N_d}\right)^{n^{(j)}_{S_d,\neg(S_{d,n})}} \\
&\propto \frac{n^{(v)}_{k,\neg(S_{d,n})}+\eta}{\sum_{i=1}^{V}(n^{(i)}_{k,\neg(S_{d,n})}+\eta)}\cdot\left(\frac{n^{(k)}_{P_d}}{N_d}\right)
\end{aligned}
\tag{1.11}
$$

附录 B　SpareNTM 的损失函数推导细节

在正文中外我们定义了变分分布 $q(\theta, b|x) = q(b|x; \hat{\lambda})q(\theta|x, b; \hat{\alpha})$ 去估计真实的后验分布 $p(\theta, b|x)$, 其中 $q(b|x; \hat{\lambda}) = \prod_{k=1}^{K} q(b_k|\hat{\lambda}_k)$, $q(b_k|\hat{\lambda}_k)$ 是一个参数为 $\hat{\lambda}_k$ 的伯努利分布。同时，我们定义了 $q(\theta|x, b; \hat{\alpha}) = \mathrm{Dir}(b \cdot \hat{\alpha})$. 因此, SpareNTM 的变分推断将优化以下 ELBO：

$$\begin{aligned}
\mathcal{L}(x) &= E_{q(\theta,b|x)}\left[\log p(x, \theta, b|\alpha, \lambda, \beta) - \log q(\theta, b|x)\right] \\
&= E_{q(\theta,b|x)}\left[\log p(x|\theta) + \log p(\theta|b) + \log p(b) - \log q(b|x) - \log q(\theta|x, b)\right] \\
&= E_{q(\theta,b|x)}\left[\log p(x|\theta)\right] - E_{q(\theta,b|x)}\left[\log \frac{q(\theta|x, b)}{p(\theta|b)}\right] - E_{q(\theta,b|x)}\left[\log \frac{q(b|x)}{p(b)}\right] \\
&= \mathcal{L}_{rec} + \mathcal{L}_{\theta} + \mathcal{L}_{b}
\end{aligned} \tag{1.12}$$

B.1 $\mathcal{L}_{\theta}$ 项的推导

Term $\mathcal{L}_{\theta} = -E_{q(\theta,b|x)}\left[\log \frac{q(\theta|x,b)}{p(\theta|b)}\right]$ can be written to:

$$\begin{aligned}
&E_{q(\theta,b|x)}\left[\log \frac{q(\theta|x, b)}{p(\theta|b)}\right] = \int_{\theta,b} q(b|x)q(\theta|x, b) \log \frac{q(\theta|x, b)}{p(\theta|b)} \,\mathrm{d}\theta, b \\
&= \int_{b} q(b|x) \int_{\theta} q(\theta|x, b) \log \frac{q(\theta|x, b)}{p(\theta|b)} \,\mathrm{d}\theta \,\mathrm{d}b = E_{q(b|x)}[KL(q(\theta|x, b)||p(\theta|b))]
\end{aligned} \tag{1.13}$$

B.2 $\mathcal{L}_{b}$ 项的推导

$\mathcal{L}_{b} = -E_{q(\theta,b|x)}\left[\log \frac{q(b|x)}{p(b)}\right]$ 将被改写为:

$$\begin{aligned}
&E_{q(\theta,b|x)}\left[\log \frac{q(b|x)}{p(b)}\right] = \int_{\theta,b} q(b|x)q(\theta|x, b) \log \frac{q(b|x)}{p(b)} \,\mathrm{d}\theta, b \\
&= \int_{b} q(b|x) \log \frac{q(b|x)}{p(b)} \left(\int_{\theta} q(\theta|x, b) \,\mathrm{d}\theta\right) \mathrm{d}b \\
&= \int_{b} q(b|x) \log \frac{q(b|x)}{p(b)} \,\mathrm{d}b = \int_{b} \prod_{k=1}^{K} q(b_k|x) \cdot \log \prod_{k=1}^{K} \frac{q(b_k|x)}{p(b_k)} \,\mathrm{d}b \\
&= \int_{b_2 \ldots b_K} q(b_2|x) \ldots q(b_K|x) \left[\int_{b_1} q(b_1|x) \log \frac{q(b_1|x)}{p(b_1)} + q(b_1|x) \log \frac{q(b_2|x) \ldots q(b_K|x)}{p(b_2) \ldots q(b_K)} \,\mathrm{d}b_1\right] \mathrm{d}b_2 \ldots b_K \\
&= KL(q(b_1|x)||p(b_1)) + \int_{b_2 \ldots b_K} q(b_2|x) \ldots q(b_K|x) \log \frac{q(b_2|x) \ldots q(b_K|x)}{p(b_2) \ldots q(b_K)} \,\mathrm{d}b_2 \ldots b_K \\
&= \sum_{k=1}^{K} KL(q(b_k|x)||p(b_k))
\end{aligned} \tag{1.14}$$

致　谢

攻读硕士学位期间的研究成果

[1] Chen, J., Wang, R., He, J., Li, M. J. (2023, September). Encouraging Sparsity in Neural Topic Modeling with Non-Mean-Field Inference. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, ECML-PKDD (pp. 142-158). Cham: Springer Nature Switzerland. (CCF B 类会议)

[2] Li, M. J., Chen, J., Li, J., Wang, R., Zhang, Q.. Transferring Knowledge from Large Language Models for Short Text Topic Modeling. International Conference on Data Engineering, ICDE. (CCF A 类会议，在投)

[3] He, J., Chen, J., Li, M. J. (2022, November). Multi-knowledge Embeddings Enhanced Topic Modeling for Short Texts. In International Conference on Neural Information Processing (pp. 521-532). Cham: Springer International Publishing. (CCF C 类会议)

[4] Li, M. J., Wang, R., Li, J., Bao, X., He, J., Chen, J., He, L. (2023, November). Topic Modeling for Short Texts via Adaptive Pólya Urn Dirichlet Multinomial Mixture. In International Conference on Neural Information Processing (pp. 364-376). Singapore: Springer Nature Singapore. (CCF C 类会议)