分类号 TP39 学校代码 10590

UDC 004 密 级 公开

硕士学位论文

xxxxxxxxxxx

yyyyyyyyyyyyyyyy

学位申请人姓名	xxx
学位申请人学号	xxxxxxxxx
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	xxxxxxxxxxx

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 xxxxxxxxxxxyyyyyyyyyyyyyyyy 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

中文摘要

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

ABSTRACT

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。

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 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 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

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

其中 $\vec{l}=\{\vec{l}_d\}_{d=1}^{D}=\{\vec{z}^+,\vec{z}\}=\{\vec{z}^+_d,\vec{z}_d\}_{d=1}^{D}$。$n^{(k)}_{P_d}$ 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：

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

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:

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

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)