软件学报 ISSN 1000-9825, CODEN RUXUEW
2024,35(11):5133−5148 [doi: 10.13328/j.cnki.jos.007035] [CSTR: 32375.14.jos.007035]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于消息传递和图先验分布的微博主题模型*

王浩成[1,2], 贺瑞芳[1,2], 吴辰昊[1,2], 刘焕宇[1,2]

[1](天津大学 智能与计算学部, 天津 300350)

[2](天津市认知计算与应用重点实验室 (天津大学), 天津 300350)

通信作者: 贺瑞芳, E-mail: rfhe@tju.edu.cn

摘　要: 检测社交媒体文本中的潜在主题是一项有意义的任务. 由于帖子具有表达简短、非正规的特点, 其将带来严重的数据稀疏问题. 不仅如此, 基于变分自编码器 (variational auto-encoder, VAE) 的模型在主题推断过程中还忽视了用户间的社交关系, 考虑 VAE 假设输入的数据点间是相互独立的. 这导致了推断的潜在主题变量间缺少了相关性信息, 进而导致主题不够连贯. 社交网络结构信息不仅聚合上下文信息的线索, 还暗示了用户间的主题相关性. 因此, 提出基于消息传递和图先验分布的微博主题模型, 其借助图卷积网络 (graph convolution network, GCN) 编码更加丰富的上下文信息, 并且在变分自编码器推断主题的过程中, 通过图先验分布整合用户交互关系以促进对多数据点复杂关系的理解, 从而更好地挖掘社交媒体主题信息. 在 3 个真实微博数据集上的实验证明了所提方法的有效性.

关键词: 社交媒体主题检测; 用户相关性; 图先验分布

中图法分类号: TP391

中文引用格式: 王浩成, 贺瑞芳, 吴辰昊, 刘焕宇. 基于消息传递和图先验分布的微博主题模型. 软件学报, 2024, 35(11): 5133–5148. http://www.jos.org.cn/1000-9825/7035.htm

英文引用格式: Wang HC, He RF, Wu CH, Liu HY. Microblog Topic Model Based on Message Passing and Graph Prior Distribution. Ruan Jian Xue Bao/Journal of Software, 2024, 35(11): 5133–5148 (in Chinese). http://www.jos.org.cn/1000-9825/7035.htm

Microblog Topic Model Based on Message Passing and Graph Prior Distribution

WANG Hao-Cheng[1,2], HE Rui-Fang[1,2], WU Chen-Hao[1,2], LIU Huan-Yu[1,2]

[1](College of Intelligence and Computing, Tianjin University, Tianjin 300350, China)

[2](Tianjin Key Laboratory of Cognitive Computing and Applications (Tianjin University), Tianjin 300350, China)

Abstract: Detecting latent topics in social media texts is a meaningful task, and the short and informal posts will cause serious data sparsity. Additionally, models based on variational auto-encoders (VAEs) ignore the social relationships among users during topic inference and VAE assumes that each input data point is independent. This results in the lack of correlation information between the inferred latent topic variables and incoherent topics. Social network structure information can not only provide clues for aggregating contextual messages but also indicate topic correlation among users. Therefore, this study proposes to utilize the microblog topic model based on message passing and graph prior distribution. This model can encode richer context information by graph convolution network (GCN) and integrate the interactive relationship of users by graph prior distribution during VAE topic inference to better understand the complex correlation among multiple data points and mine social media topic information. The experiments on three actual datasets validate the effectiveness of the proposed model.

Key words: topic detection of social media; user correlation; graph prior distribution

社交媒体指基于用户交互的内容发布与传播平台. 随着移动互联网的繁荣, 新浪微博、推特 (Twitter) 等社交媒体平台蓬勃发展, 并迅速吸引了普通大众的视线, 成为人们日常生活中不可缺少的一部分. 主题模型可以用来自

* 基金项目: 国家自然科学基金 (61976154); 国家重点研发计划 (2019YFC1521200)
收稿时间: 2023-01-02; 修改时间: 2023-04-06; 采用时间: 2023-08-24; jos 在线出版时间: 2023-12-06
CNKI 网络首发时间: 2023-12-08

动提取文本中的潜在主题信息, 也被用在社交媒体上, 它可以帮助人们快速掌握帖子中的语义信息, 了解其中的主要内容, 以减少人工消耗. 同时, 该模型对很多其他相关任务都有帮助. 比如, Zeng 等人在短文本分类中提出一种主题记忆机制, 来编码代表类别标签的潜在主题[1]. Xu 等人在隐式篇章关系识别中使用主题张量网络来推断主题级的向量表示[2]. Zou 等人在对话摘要中应用主题模检测不同对话角色中的主题信息, 以捕获对话中不同角色的特征[3].

传统的主题模型, 如 latent Dirichlet allocation (LDA)[4], 广泛地应用在长文档中, 并通过隐式地捕获词共现模式来揭示主题信息. 然而, 在社交媒体场景中, 帖子简短且表达不正式, 导致了严重的数据稀疏性, 使得推断主题不够连贯, 从而影响了模型性能. 现有社交媒体主题检测的相关研究可以大致分为 3 类: (1) 通过启发式聚合策略捕获跨文档的词共现模式[5−7], 或者直接建模语料中词对的生成过程[8,9]; (2) 利用词嵌入技术[10]将词之间的语义信息整合到主题模型中[11,12]. 然而, 上述两种方法都只关注了帖子文本的内容信息; (3) 最新研究[13,14]通过整合社交媒体中关注网络或转发网络来考虑社交上下文信息. 然而, 常用的近似后验分布的方法, 如变分推断[15], MCMC[16]有较高的计算复杂度, 并且模型假设的微小变动, 都会导致推理过程的重新推导.

考虑变分自编码器通过反向传播从而自动执行变分推理的强大能力, 整合社交上下文的模型进一步使用 VAE 作为推理框架. IATM[17]和 PCFTM[18]是基于 VAE 的经典模型, 其通过建模网络结构增强短文本的表示. IATM 考虑好友用户间的动态交互, 并将内容表示和结构表示简单拼接起来, 然而其没有捕获帖子内容和社交网络结构间的内在联系, 是一种次优的组合方式. PCFTM 通过自融合网络无缝地融合文本和结构表示, 学习灵活阶的用户嵌入表示. 然而它将图结构转换为线性结构, 丢失了社交网络中部分复杂关系. DGTM[19]建模不同的传播过程, 同样基于 VAE 进行主题推理, 然而, 这些工作假设帖子彼此独立, 并用标准高斯分布近似帖子的潜在主题[20]. 然而, 在真实的社交媒体情景中, 帖子会受到用户交互的影响, 其主题具有潜在的相关性, 而标准高斯分布数据的独立同分布假设 (i.i.d.) 不能建模这一特性. 因此, 在变分自编码器使用一个整合社交网络中用户交互的先验分布将更为合理.

GCN 处理复杂图数据的能力非常强大, 其将卷积操作拓展到到图领域, 并自然地将结构信息和内容信息整合起来. GCN 在许多领域都取得了成功, 如谣言检测[21]、文本摘要[22]等, 其消息传递架构可以刻画社交网络中的信息传播, 这促使我们使用图卷积网络聚合具有类似主题的相关帖子, 以缓解数据稀疏. 对于 VAE 在主题推理过程中的独立同分布假设, 受文档哈希[23]中整合相关文档信息的启发, 本文进一步将用户相关性集成到主题推断中. 基于社交网络中的交互关系, 我们构造了一个图先验分布, 它诱导变分自编码器推断的潜在主题变量考虑用户相关性. 总体而言, 提出了一种基于消息传递和图先验的主题模型. 该模型在社交网络编码和主题推理过程中都考虑了用户交互. 特别地, 其构建了一个图先验分布, 使变分自编码器推断的每个用户的潜在主题变量蕴含用户在社交网络中的关联关系. 在 3 个真实数据集上的大量实验表明本文提出的模型的有效性.

本文第 1 节归纳社交媒体主题检测的研究现状, 并总结前人方法所存在的问题. 第 2 节论证传统变分自编码器在社交媒体主题检测中的局限性. 第 3 节详细阐述本文提出的基于消息传递和图先验分布的微博主题挖掘方法. 第 4 节介绍实验数据准备、模型评估方法以及实验结果的讨论与分析. 第 5 节为样例分析, 展示模型推断的主题词. 第 6 节进行总结和展望.

1 相关工作

面向社交媒体短文本的主题模型由于帖子存在严重的数据稀疏问题, 影响推断主题的效果. 目前国内外已有相关研究可以分为两大类, 一类是仅关注社交媒体内容的方法, 一类是整合内容与社交上下文的方法.

1.1 仅考虑社交媒体内容

这类方法单纯依靠社交媒体中的内容信息推断主题信息.

(1) 基于聚合的方法. 为了解决短文本的数据稀疏性, Zhao 等人根据作者属性经验性地将短本文聚合起来[5]. 相似地, Mehrotra 等人综合实验了基于作者、爆发指数、时间以及标签的聚合, 并表明基于标签的聚合方法可以

为主题模型的性能带来更大的提升[6]. Alvarez-Melis 等人提出一种按照会话关系聚合帖子的策略[24], 帖子与其所有评论帖子被看作一段会话, 且被聚合成一篇独立的文档, 其中的用户被认为是这篇文档的共同作者. Quan 等人提出基于自聚合策略的主题模型[7], 其本质是建立在主题亲和度 (topical affinity) 上的聚合. 该模型假设短文本是长文档随机分解的结果, 即每个短文本都可以从一个看不到的伪长文档中采样得到. 除了启发式聚合策略和自聚合策略, BTM[8]等方法基于窗口聚合的策略, 直接建模窗口内词对 (biterm) 的生成过程. 文本中词对的数量比单个词的数量多, 因此可以一定程度上缓解数据稀疏问题. 然而, 上述基于聚合策略的方法依然依赖词共现模式, 仅关注了词与词之间的统计关系, 忽视了它们之间的语义关联.

(2) 基于表示的方法. 词嵌入 (word embedding) 技术的出现为学习词之间的语义关联提供了支持[25]. 简单来说, 具有相似上下文的词可以学到相似的向量表示, 在特征空间中距离更近. 词是自然语言中的基本单位, 在词的表示中嵌入语义关联使得很多任务受益, 也推动了面向社交媒体的概率主题模型的发展. LCTM[11]不再建模词的生成过程, 而是建模词向量的生成. 其引入一个新的变量——潜在概念, 同时将主题定义为潜在概念上的分布. Li 等人[26]指出, 当人们去理解一段短文时, 不仅基于短文中的词本身, 还基于它的背景知识, 比如词与词之间的语义关联. 两个词在语义上关系紧密, 但共同出现次数少, 基于词共现模式的方法不能捕获这种语义关联. 于是作者提出在模型中整合从大量辅助文本中学到的词语关联信息, 改进了短文本的主题模型. Shi 等人[12]认为尽管词嵌入对短文本主题建模有帮助, 但其训练文本 (维基百科、谷歌新闻等) 和短文本有很大不同, 于是提出 SeaNMF 模型, 将短文本作为一个窗口, 直接使用 skip-gram 方法学习短文本中的语义关联并嵌入到主题模型中. 基于词嵌入的方法学习词与词之间的语义关联, 但在社交媒体情境下, 帖子文本与社交网络结构紧密关联, 仅对文本内容进行建模是不充分的.

1.2 整合内容和社交上下文

不同于一般短文本, 社交媒体中存在丰富的结构信息, 它们不仅可以丰富帖子文本的表示, 还可以为发现主题信息提供有效的线索. 帖子与帖子之间、用户与用户之间根据不同的交互方式形成了多种多样的交互网络. 整合其中的结构信息已经成为目前社交媒体主题模型研究的热点.

(1) 基于静态网络结构. SRTM[13]联合建模帖子文本和社交网络中的关注关系, 并且判断一段关注关系是否由主题因素造成, 从而排除非主题因素导致的关注关系. Li 等人[14]和 Chen 等人[27]基于网络中的评论关系改进主题模型. 以 Li 等人提出的 LeadLDA 模型为例, 根据转发评论构建帖子的会话树. 出于对当前帖子的主题更感兴趣, 用户才会愿意花时间转发与评论, 因此会话树中的帖子主题更加相关、一致. 会话树中的帖子可以分为对主题贡献较大的领导者 (leader) 以及对主题贡献较小的追随者 (follower). 在这一观察的基础上, LeadLDA 使用条件随机场检测领导者与追随者, 并将这一信息整合到模型中, 得到了性能上的提升.

(2) 基于用户动态行为. IATM[17]和 PCFTM[18]利用网络表示学习技术挖掘用户间的交互行为, 并基于学到的嵌入表示采用 VAE 推断主题. 然而, 学到的表示没能充分利用社交网络中的复杂结构特征. DGTM[19]综合考虑宽度、深度两种传播模式, 采用 GCN 整合更加丰富的社交上下文信息. 然而在主题推断过程中, 标准变分自编码器的数据独立假设忽视了用户间的相关性. 另外, Zheng 等人[28]、Zhou 等人[29]和 Zhang 等人[30]整合文档间的结构关系, 然而这 3 种方法都适用于长文本数据, 不能解决短文本中的数据稀疏问题. 本文使用图卷积网络的消息传递机制编码社交上下文, 并通过构建图先验分布建模数据之间的复杂关联信息, 并诱导产生更加连贯的主题信息.

1.3 图卷积网络

社交网络可以看做是一个结构复杂的图, 用户之间存在错综复杂的交互关联, 因此对社交网络的分析与编码十分重要. 卷积操作在图像领域取得了很大的成功, 图卷积网络直接处理图结构数据, 将卷积操作泛化到图结构中, 并且天然地将网络结构和节点属性融合在一起, 学习统一的节点表示, 已成功应用在很多任务中. 比如, DeepInf[31]使用 GCN 预测社会影响力, MOGANED[32]通过 GCN 编码候选触发词来辅助事件检测任务, BiGCN[21]利用 GCN 在社交网络中检测谣言, 都取得了很好的效果. 如前所述, 帖子简短且表达不正式, 在主题检测时遭遇了严重的数据稀疏问题, 导致推断出的主题词不够连贯. GCN 的“消息传递”机制可以很好地刻画信息在社交网络中的传播过

程. DGTM[19]利用图卷积网络建模宽度、深度两种传播模式, 捕获更加丰富的社交上下文. 本文在 GCN 的基础上增加残差网络, 借鉴人们在阅读帖子时参考相关帖子的过程, 采用 GCN 将社交媒体中的帖子信息根据网络中的链路结构传播、聚合, 从而缓解单一帖子所面临的数据稀疏问题, 以学习上下文增强的节点表示.

2 变分自编码器的局限性

现有的方法, 如 IATM[17]、PCFTM[18]以及 DGTM[19], 将变分自编码器作为主题推断的基本框架, 并假设数据之间相互独立. 接下来, 本文将先简要介绍在多数据点输入的情况下, 之前方法中变分自编码器的推断过程. 在了解这一过程的基础上, 分析当处理复杂社交网络数据时, 由独立同分布假设导致的忽略用户相关性的问题. 下面就相关概念和基本知识予以介绍.

如图 1 所示, 假设输入的数据点集合为 $X=\{x_1,\ldots,x_n\}$, 其中 x_i 表示第 i 个数据 (用户帖子) 的向量表示, 此处可以假设为词袋向量, n 表示数据点总数. 标准的变分自编码器对每个用户数据点 x_i 独立地执行如下过程.

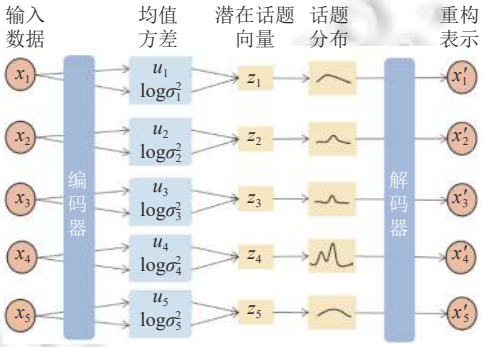


图 1 VAE 在多数据点输入下的主题推断

(1) 编码器计算高斯主题分布 $q(z_i|x_i)$ 的均值 μ_i 和方差 σ_i^2, 这一部分通常选择使用多层感知机 (multi-layer perception, *MLP*):

$$\mu_i = MLP_1(x_i) \tag{1}$$

$$\log\sigma_i^2 = MLP_2(x_i) \tag{2}$$

(2) 数据点 i 的潜在主题向量 z_i 通过重参数技巧从后验分布 $q(z_i|x_i)$ 中采样得到, 在潜在主题向量上使用 *Softmax* 函数获得每个用户对应的主题分布 θ_i, 上述过程如公式 (3), 公式 (4) 所示:

$$z_i = \mu_1 + \varepsilon_i \times \sigma_i \tag{3}$$

$$\theta_i = Softmax(z_i) \tag{4}$$

其中, $\sigma_i \in N(0, I)$.

(3) 解码器通过神经网络层独立地重构每个数据点的向量表示, 其中参数矩阵 ϕ_w 即为所求的主题-词分布, 主题分布与主题-词分布矩阵相乘重构得到原始的数据表示:

$$x_i' = ReLU(\theta_i \times \phi_w) \tag{5}$$

总体来说, 变分自编码器对每个用户数据独立地执行上述过程. 证据下界 (evidence lower bound, *ELBO*) 如公式 (6) 所示, 是每个数据点的 *ELBO* 计算结果的累加. 其中包含两部分, 一是重构误差, 即 x_i' 与 x_i 之间的误差, 二是主题后验分布与先验分布的 *KL* 散度.

$$ELBO = \sum_{i=1}^{n} [E_{q(z_i|x_i)}[\log p(x_i|z_i)] - KL(q(z_i|x_i)\|p(z_i))] \tag{6}$$

最大化 *ELBO* 使得 *KL* 散度最小化, 进而使得后验分布 $q(z_i|x_i)$ 逐渐逼近先验分布 $p(z_i)$, 先验分布通常选为标准高斯分布. 标准高斯分布的独立同分布假设认为潜在变量之间是相互独立的. 这意味着在使用变分自编码器进

行主题推断时没有考虑用户间的相关性. 在社交媒体中, 用户之间存在丰富的交互行为, 如转发、评论等, 它们暗含了用户间的相关性信息. 可以想到, 具有转发、评论关系的两个用户关注的主题是更加相关的, 它们的潜在主题变量服从这种相关结构更为合理, 这更有利于挖掘相邻好友间的相关主题, 从而推断更加连贯的主题信息. 本文拟针对这一问题, 构建图先验分布替代标准高斯分布, 进一步在变分自编码器推断主题过程融入用户相关性.

3 基于消息传递和图先验分布的微博主题模型

针对现有模型的局限性, 本文设计了一个基于消息传递和图先验分布的社交媒体主题模型 (message passing and graph prior for topic model, MGTM), 在网络表示和主题推断两阶段整合用户交互. 该模型依然以用户级的社交网络作为输入, 首先在网络表示学习阶段, 使用残差图卷积网络学习社交网络中用户节点的向量表示, 借助消息传递机制缓解数据稀疏, 并将用户间的交互特征融入向量表示中. 然后在主题推理阶段, 构建图先验分布替代标准高斯分布, 使变分自编码器在主题推断时能够拥有用户相关性的先验知识, 进而使推断的每个用户潜在主题变量能够服从社交网络中的交互结构. 接下来, 将分阶段阐述 MGTM 模型, 其结构框架如图 2 所示.

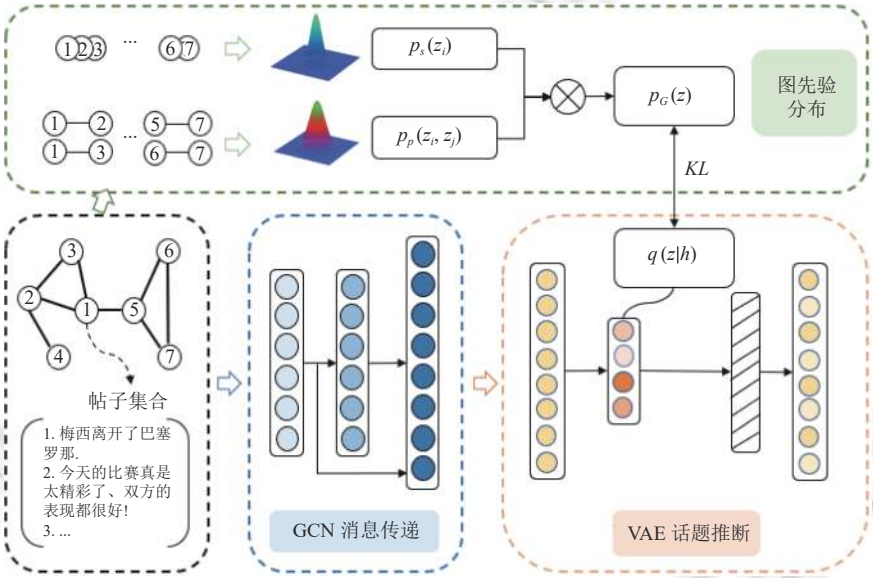


图 2 基于消息传递与图先验分布的微博主题模型框架图

3.1 构建用户级社交网络

根据用户之间的转发、评论行为, 构建用户级的社交网络 $G=(V,E,T)$. $V=\{v_i|1\leqslant i\leqslant n\}$ 表示节点集合, 其中 v_i 代表社交网络中的用户 i, n 代表用户总数. 如果两个用户存在转发、评论行为, 则称它们为一对好友. $E_i=\{e_{ij}|1\leqslant i,j\leqslant n\}$ 表示边的集合, 如果 v_i 所代表的用户 i 与 v_j 所代表的用户 j 有过交互, 则 $e_{ij}=1$, 两个节点之间建立一条边; 如果 v_i 所代表的用户 i 与 v_j 所代表的用户 j 从未交互过, 则 $e_{ij}=0$, 两个节点之间不存在直接连通的边. 将每个用户发表的帖子作为该用户节点的属性信息, $T=\{t_1,t_2,\ldots,t_n\}$ 表示用户帖子集合, 其中 t_i 表示用户 i 发表的帖子文本内容. 根据用户交互行为, 可以得到社交网络的邻接矩阵 A. 接着, 将所有用户的帖子向量化得到社交网络的属性矩阵 X.

3.2 基于消息传递的用户嵌入表示

使用残差图卷积网络作为编码社交网络中用户节点的基础框架. 图卷积网络可以将不同的关系矩阵作为输入, 捕获不同类型的交互关系. 本文以邻接矩阵 A 和属性矩阵 X 作为输入, 经过两层卷积网络, 学习好友用户间的相关性, 计算出每个用户节点的向量表示, 如公式 (7)–公式 (9) 所示:

$$\hat{A}=\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}} \tag{7}$$

$$H^1=ReLU\left(\hat{A}XW^1\right) \tag{8}$$

$$H^2=ReLU\left(\hat{A}\left(H^1+X\right)W^2\right) \tag{9}$$

其中, $\tilde{A}$、$\hat{A}$、$\tilde{D}$ 分别表示增加自连接的邻接矩阵、归一化的邻接矩阵以及度矩阵. H^1 表示经过一层图卷积网络后的隐藏节点嵌入. W^1 和 W^2 表示模型参数. 两层图卷积网络依然使用 $ReLU$ 作为激活函数. 最终, H^2 表示所有用户节点的嵌入表示构成的矩阵. 在第 2 层图卷积网络中将 H^1 与 X 组合起来作为属性矩阵, 构成残差网络. 一方面它强化了每个用户的原始帖子信息在更新嵌入表示时的作用, 防止出现用户表示同质化的现象; 另一方面它还可以缓解梯度消失问题.

主题检测是一个无监督的任务, 因此图卷积网络不能像其他任务那样以有监督的方式训练. 本文为图卷积网络设计了一个自监督的损失函数, 如下所示:

$$loss_G=-\sum\nolimits_{v_i\in V}\sum\nolimits_{v_j\in N_i}\log P(v_j|v_i) \tag{10}$$

$$P(v_j|v_i)=\frac{\exp\left(h_j^{\mathrm{T}}\cdot h_i\right)}{\sum_{v_k\in N_i}\exp\left(h_k^{\mathrm{T}}\cdot h_i\right)} \tag{11}$$

给定用户节点 v_i, 我们的目标是最大化 v_i 与其在 N_i 中的相邻节点的相似度. 其中, N_i 是节点 v_i 的一阶邻居的集合. h_i 是节点 v_i 的嵌入表示, 对应 H^2 中的第 i 行.

在图卷积网络“消息传递”机制的作用下, 每个用户节点会聚合其邻居节点的属性信息, 这些聚合来的属性信息丰富了当前用户的帖子内容, 补充了单一用户的帖子由于篇幅简短和表达随意而丢失的信息, 缓解了社交媒体中的数据稀疏. 同时, 社交网络中的交互关系也保留在了嵌入表示中, 好友节点在表示上拥有了更高的相似度.

3.3 基于图先验分布的主题推断

在已有基于变分自编码器进行主题推理的方法[17–19]中, 假设数据点相互独立, 在主题推理阶段忽视了用户间的相关性. 变分自编码器在新闻文稿等标准文档中得到了广泛应用, 但对于具有复杂关系的多数据点 (用户) 输入的情况, 如前所述, 将数据点间的相关性整合到潜在主题向量中更为合理.

变分自编码器中的先验分布采用标准高斯分布, 导致了潜在主题变量相互独立[20]. 混合高斯分布与流模型都未出现解决用户相关性的研究. 受之前工作[22]的启发, 本文提出一种融入用户交互的神经主题变分推断方法. 该方法首先构造一个图先验分布来替代标准高斯分布. 图先验分布在构造过程中, 考虑用户间的交互行为, 从而使得变分自编码器在推断过程中掌握更多用户相关性的先验知识, 进而鼓励每个用户的潜在主题变量考虑用户间的关系结构, 学习到包含相关性的主题.

图先验分布: 用户间的交互关系通过社交网络 $G=(V,E)$ 体现, v_i 与 v_j 间存在交互关系, 则在图 G 中 $e_{ij}=1$, 即 $(v_i,v_j)\in E$. 本文构造一个潜在主题向量 $(z_1,\ldots,z_n)$ 上的图先验分布 $p_G(z)$, 它基于社交网络中的链路结构构建而得, 因此与标准高斯分布相比, 它学习了用户间的交互关系, 其具体形式如下所示:

$$p_G(z)=\sum\nolimits_{i=1}^{n}p_s(z_i)\prod\nolimits_{(v_i,v_j)\in E}\frac{p_p(z_i,z_j)}{p_s(z_i)\,p_s(z_j)} \tag{12}$$

其中, z_i 和 z_j 是用户 v_i 和 v_j 的潜在主题向量. $p(z)$ 包含两部分: (1) 对于社交网络中的每个用户 v_i, $p(z_i)$ 负责捕获其潜在主题向量上的分布特征, 这里可以使用简单的标准高斯分布. (2) 对于社交网络中每条边 $(v_i,v_j)\in E$, v_i 和 v_j 是有过转发、评论行为的好友用户, $p_p(z_i,z_j)$ 可以捕获两个好友用户的潜在主题向量间的相关性特征. 由于其涉及两个用户间的交互行为, 标准的高斯分布并不能建模这种关系, 因此对于 $p_p(z_i,z_j)$, 本文采用如下形式:

$$p_p(z_i,z_j)=N\left(\mu=0,\Sigma=\begin{pmatrix}I & \alpha\cdot I\\ \alpha\cdot I & I\end{pmatrix}\right) \tag{13}$$

其中, α 是高斯分布的超参数, 代表先验分布中的相关性强度. 图先验分布 $p_G(z)$ 中注入了用户交互关系 $(v_i,v_j)\in E$, 且 VAE 的损失函数中 KL 散度项会将后验主题分布推向先验分布, 因此变分自编码器学到的主题后验分布也将

包含用户间的交互关系.

证据下界: 公式 (6) 展示了标准变分自编码器的证据下界, 将其中的标准高斯分布替换为图先验分布并对公式进行整理, 得到如下形式的证据下界:

$$ELBO=\sum_{i=1}^{N}(E_{q(z_i|h_i)}[\log p(h_i|z_i)]-KL(q(z_i|h_i)\|p(z_i)))$$
$$-\sum_{(v_i,v_j)\in E}(KL(q(z_i,z_j|h_i,h_j)\|p_p(z_i,z_j))-KL(q(z_i|h_i)\|p_s(z_i))-KL(q(z_j|h_j)\|p_s(z_j))) \tag{14}$$

其中, 对于每个用户数据点, $\log p(h_i|z_i)$ 表示由潜在主题向量重构用户表示, $KL(q(z_i|h_i)\|p(z_i))$ 表示主题后验分布与先验分布之间的 KL 散度. 对于网络中每对交互关系 (每条边), $KL(q(z_i,z_j|h_i,h_j)\|p_p(z_i,z_j))$ 、 $KL(q(z_i|h_i)\|p_s(z_i))$ 和 $KL(q(z_j|h_j)\|p_s(z_j))$ 都表示主题后验分布与先验分布之间的 KL 散度. $ELBO$ 中后验变分分布 $q(z_i,z_j|h_i,h_j)$ 采用与先验分布中 $p_p(z_i,z_j)$ 相对应的形式, 如下所示:

$$q(z_i,z_j|h_i,h_j)=N\left(\begin{bmatrix}\mu_i\\ \mu_j\end{bmatrix},\begin{bmatrix}\sigma_i^2 & \gamma_{ij}*\sigma_i*\sigma_j\\ \gamma_{ij}*\sigma_i*\sigma_j & \sigma_j^2\end{bmatrix}\right) \tag{15}$$

其中, μ_i , μ_j 和 σ_i^2 , σ_j^2 表示均值和方差, γ_{ij} 是对角线矩阵, 控制 z_i 和 z_j 间的相关性强度, 是模型训练的参数之一, *表示哈达玛积. 将各个高斯分布的参数代入新的证据下界中, 得到基于图先验分布的变分自编码器的损失函数, 如公式 (16) 所示:

$$loss_v=-\sum_{i=1}^{N}\left(E_{q(z_i|h_i)}[\log p(\mathrm{h}_i|z_i)]-\frac{1}{2}\left(\mu_i^2+\sigma_i^2-1-2\log\sigma_i\right)\right)+\sum_{(v_i,v_j)\in E}\left(0.5\cdot\left\{\log\left(1-\alpha^2\right)\right.\right.$$
$$\left.\left.-\left(\mu_i^2+\mu_j^2+\sigma_i^2+\sigma_j^2+\log\left(1-\gamma_{ij}^2\right)\right)+\frac{\mu_i^2+\mu_j^2+\sigma_i^2+\sigma_j^2-2\alpha\gamma_{ij}\sigma_i\sigma_j-2\alpha\mu_i\mu_j}{2}\right\}\right) \tag{16}$$

从损失函数可以看出, 图先验变分自编码器包括以下 3 部分.

(1) 编码器 (推断网络): 与标准的变分自编码器一样, 以一个用户数据 $[h_i]$ 作为输入, 计算主题后验分布的均值 μ_i 和方差 σ_i^2 , 并利用重参数技巧采样得到潜在主题向量 z_i , 进而得到主题分布, 如公式 (1)–公式 (4) 所示.

(2) 相关性计算网络: 将一对好友节点 $(v_i,v_j)\in E$ 的向量表示 $[h_i,h_j]$ 作为输入, 计算后验变分分布中的相关性参数 γ_{ij} . 计算结果需要满足对称性, 即 $\gamma_{ij}=\gamma_{ji}$, 本文分别以拼接后的向量表示 $[h_i|h_j]$ 与 $[h_j|h_i]$ 作为输入, 通过全连接层捕获两个用户的相关性强度, 并将结果相加取平均作为最终的相关性系数, 以满足对称性.

(3) 解码器 (生成网络): 与现有工作中的变分自编码器一样, 以潜在变量 z_i 作为输入, 重构用户节点的向量表示 h_i , 如公式 (5) 所示.

以上为对基于消息传递的网络表示学习和基于图先验的变分自编码器的介绍. MGTM 从两阶段建模用户的交互行为, 特别是引入图先验分布, 将社交网络中用户的相关性信息作为先验知识, 改变了现有工作中标准高斯分布认为用户相互独立的假设, 在主题推断过程中整合用户相关性, 使得用户的潜在主题向量服从用户在社交网络中的关系结构, 从而推断出更加连贯的主题.

4 实验分析

4.1 数据集

本文所使用的的数据集需包含社交关系和帖子文本内容, 目前英文社交媒体, 如 tweet, 我们还未发现合适的数据集. 本文在 3 个真实的新浪微博数据集上进行实验. Li 等人[14]按照如下过程构建了一个新浪微博语料库: 首先收集微博中以“#”开头的 hashtag 标签, 然后利用新浪微博官方提供的搜索接口 (hashtag-search API) 搜索与给定 hashtag 相对应的帖子, 最终构建了一个由 2014 年 5 月 1 日–7 月 31 日的微博数据构成的语料库. 语料中包含了帖子文本、用户信息 (用户名使用随机生成的编号代表, 以保护用户隐私)、用户的交互信息等. 为了得到包含不同主题分布的数据集, Li 等人[14]将整个语料库按照月份分成 3 个数据集, 每个数据集包含一个月的微博数据,

由此构成了 3 个验证数据集. 这一数据集在 2016 年公开, 并且被后续工作广泛使用[17,18].

跟随之前的工作[17]在这 3 个数据集上进行如下处理: (1) 对每个帖子进行分词, 删除少于 3 个词的帖子记录; (2) 过滤没有转发、评论行为的用户; (3) 将每个用户发表的原微博和转发微博聚合起来构成用户的文本属性; (4) 根据用户间的转发、评论行为, 构建用户级的社交网络. 表 1 中列出了本文最终使用的 3 个验证数据集中的统计信息, 包括用户数、交互数 (转发、评论) 以及词表大小.

表 1　3 个验证数据集的统计信息

数据集	用户数	交互数	词表大小	帖子平均长度
5月	8907	10435	5914	9.5
6月	19293	35962	9368	10.8
7月	16990	20971	9663	9.6

4.2 评价指标

内部评价: 在早期的研究中, 经常使用困惑度 (perplexity) 评估推断主题的质量, 然而 Chang 等人[33]证明困惑度评价高不一定对应人类认知中语义连贯的主题. 为了有意义且客观地评估模型, 本文跟随 Mimno 等人[34]使用更接近人类判断且目前更为流行的连贯性 (topic coherence) 作为评价指标. 它计算每个主题下词语的连贯性分数, 具体公式如下所示:

$$C=\frac{1}{K}\cdot\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{i-1}\log\frac{D\left(w_i^k,w_j^k\right)+1}{D\left(w_j^k\right)} \tag{17}$$

其中, w_i^k 表示主题 k 中根据话题-词分布排序得到的出现概率第 i 大的词, $D\left(w_i^k,w_j^k\right)$ 表示数据集中同时出现词 w_i^k 和词 w_j^k 的帖子的数量, $D\left(w_j^k\right)$ 表示数据集中包含单词 w_j^k 的帖子数量. 公式对 K 个主题的连贯性得分取均值作为最终评价得分. 对于每个主题, 根据主题-词分布找出概率最大的前 N 个词, 计算连贯性分数. 若两个词共同出现的概率高, 则认为这两个词在语义上更加连贯. 本文在实验中分别在主题数 K 为 50、100 以及词数 N 为 10、15、20 这 6 种组合设定下计算提出模型与对比方法的连贯性分数, 以评价它们推断主题的质量.

外部评价: 除了连贯性, 本文还在下游任务——链路预测上进一步验证模型性能. 评价指标选择常用的 *AUC* (area under the receiver operating characteristic curve), 基于测试集中边的相似值和不存在的边的相似值的比较进行评测.

4.3 基线模型

为验证本文提出的 MGTM 的性能, 我们选择以下社交媒体主题模型进行实验对比: LCTM[11] 和 NQTM[35]属于仅关注帖子文本内容的方法; LeadLDA[14]、AdjEnc[36]、PCFTM[18]、IATM[17]、DGTM[19]同时建模帖子的内容信息和社交网络的结构信息. 值得注意的是, 在社交媒体主题检测领域, LCTM 和 LeadLDA 是最先进的概率主题模型, 其已经被证明优于 LDA、BTM 等模型. 下面是对各个对比模型以及退化模型的简要介绍.

(1) LCTM 通过潜在概念的共现模式揭示主题. 其将潜在概念作为变量引入模型, 捕获单词间的概念相似度. 在 LCTM 中, 每个主题被建模为潜在概念上的分布, 而每个潜在概念被建模为词嵌入空间上的局部高斯分布. 引入词嵌入, 捕获词与词之间的语义关联.

(2) NQTM 通过使用分布量子化机制和负采样解码器, 得到区分度更大的峰值主题分布, 从而提升自编码器框架在短文本上推断主题的质量.

(3) LeadLDA 利用微博中的会话结构提出了一种新的概率主题模型. 该方法根据转发和回复关系将微博中的帖子组织成会话树. 基于会话树结构, LeadLDA 将帖子区分为领导者帖子 (leader) 和追随者帖子 (follower), 并在模型中建模它们包含关键主题词的不同概率.

(4) AdjEnc 的输入不仅包括文档的内容, 还包括文档的邻接向量. 同样地, 模型除了重构文档本身的内容外, AdjEnc 还重构输入文档的邻接向量. 这促进了网络中的文档相互协作学习, 从而使得相邻文档在潜在空间中具有

更加相似的表示, 在表示中融入了结构特征.

(5) PCFTM 通过随机游走获取社交网络中的游走序列, 序列中包含了用户间的交互关系, 并且利用 LSTM[37] 无缝地融合游走序列中的内容表示和结构表示. 基于学到的表示, PCFTM 利用变分自编码器推断主题信息.

(6) IATM 建模用户与其好友间的交互行为, 学习感知交互的边嵌入表示. 该方法将一对好友节点的表示拼接起来, 得到融合对话上下文的边表示, 并以此作为主题推断的基础.

(7) DGTM 在建模社交网络时考虑主题的不同传播模式, 得到更加丰富的上下文表示, 并以此为基础送入标准变分自编码器进行主题推断.

4.4 实验设置

对于本文提出的 MGTM, 同样先预处理每个用户节点的帖子 t_i. 聚合每个用户的所有帖子, 进行截断与填充, 每个用户的帖子文本将标准化为 50 个词. 然后将 t_i 通过词嵌入层, 并对每条帖子取平均得到用户帖子的向量表示 x_i, 进而得到社交网络的属性矩阵 X. 词嵌入层通过随机初始化得到, 并会随着模型的训练而更新优化. 需要指出的是, Meng 等人和 Zhang 等人分别在文献 [38,39] 中提到用 Bert 等预训练模型学习词表示在文本聚类任务上表现不佳, 原因在于预训练词表示基于词序, 而文本聚类任务基于词共现, 即两任务的目标存在差异[40], 而主题推理阶段对应的本质任务也可被视为文本聚类, 并且本文重在两阶段整合社交媒体结构特征, 因此没有使用 BERT 作为词嵌入表示的初始化. 实验中采用网格搜索方法寻找超参数最优值. 在图卷积网络中, 隐藏层的维度分别设置为 200 和 400. 在图先验变分自编码器中, 第 1 层编码器的维度设置为 200, 第 2 层编码器的维度根据主题数的不同设置为 50 或 100. 训练过程中学习率在设置为 0.001, 公式 (13) 中的超参数 α 设置为 0.9. 所有模块采用 Adam 进行优化. 对于所有对比方法中的超参数, 都使用了它们在原始论文中报告的参数设置并进行微调. 对于概率主题模型 (LCTM、LeadLDA), 运行迭代吉布斯采样并保证收敛.

主题的数量 K 设置为 50 和 100, 这意味着模型将为每个数据集推断 50 或 100 个主题. 固定主题数 K, 按主题-词分布进行排序选择每个主题下出现概率最大的 10、15 和 20 个词, 并根据公式 (17) 评估模型的性能.

4.5 连贯性结果分析

4.5.1 本文方法与基线模型的比较

表 2–表 4 展示了 MGTM 和其他对比方法在 3 个评估数据集上的主题连贯性分数. 分数越高, 代表主题越连贯. K50 代表 50 个主题, K100 代表 100 个主题. N: 根据主题-词分布 ϕ_w 选择的 top 词数. 加粗表示最佳性能. 根据表 2–表 4 中结果, 可以得到进行以下观察.

表 2 MGTM 和对比方法在 5 月数据集上的主题连贯性结果

类型	模型	K50			K100		
		N=10	N=15	N=20	N=10	N=15	N=20
仅考虑文本内容的方法	LCTM	−70.91	−165.37	−296.36	−58.65	−140.10	−261.40
	NQTM	−93.04	−214.33	−384.82	−90.98	−207.64	−376.32
整合社交网络结构的方法	LeadLDA	−53.91	−138.53	−258.38	−58.15	−141.34	−261.65
	AdjEnc	−67.57	−159.66	−290.10	−70.33	−164.62	−300.47
	PCFTM	−74.03	−167.52	−303.48	−77.65	−178.18	−317.88
	IATM	−43.34	−112.64	−228.27	−47.32	−121.46	−219.96
	DGTM	−36.90	−88.59	**−172.62**	−43.59	−106.31	−205.02
我们的方法	MGTM	**−31.59**	**−86.45**	−179.76	**−40.54**	**−100.31**	**−199.88**

● 整合社交网络中的交互关系有助于推断出更加连贯的主题. 直觉上, 由交互关系构成的网络结构不仅可以丰富节点的表示, 还可以提供主题传播的有效线索. 从表 2–表 4 中可以看出, 整合网络结构的方法整体上表现优于仅考虑帖子文本内容的方法. 仅考虑文本内容的方法中, LCTM 依赖于从 Wikipedia 中训练得到的词嵌入向量, 而 Wikipedia 和帖子属于不同类型的文本, 文本格式与表达习惯的不同给模型带来了噪声与偏差. NQTM 既没有

采用词嵌入技术, 也没有整合相关的上下文信息, 数据稀疏导致其难以产生连贯的主题. 这两种仅考虑帖子内容的方法在主题连贯性上的表现较差. 在整合结构的方法中, IATM 取得了更好的效果, 这说明建模社交网络中好友间的交互行为可以有效地提升主题检测的性能. PCFTM 的表现没有其他建模网络结构的方法稳定, 这可能是由于其对网络结构的建模完全依赖于随机游走, 对网络的拓扑结构以及规模比较敏感. DGTM 在建模社交网络过程中融入了更加丰富的社交上下文信息, 从话题连贯性结果可以看出, 其取得了进一步的提升.

表 3 MGTM 和对比方法在 6 月数据集上的主题连贯性结果

类型	模型	*K*50			*K*100		
		N=10	*N*=15	*N*=20	*N*=10	*N*=15	*N*=20
仅考虑文本内容的方法	LCTM	−91.72	−208.75	−367.76	−81.88	−181.57	−323.16
	NQTM	−102.93	−239.49	−431.17	−102.23	−239.41	−432.95
整合社交网络结构的方法	LeadLDA	−63.54	−150.18	−278.19	−72.07	−169.80	−309.40
	AdjEnc	−67.57	−159.66	−290.10	−70.33	−165.87	−303.37
	PCFTM	−77.95	−181.93	−330.45	−79.77	−182.12	−325.38
	IATM	−46.69	−113.09	−213.61	−59.11	−133.96	−225.48
	DGTM	−39.24	−90.86	−176.63	−56.59	−132.21	−245.12
我们的方法	MGTM	**−36.25**	**−80.75**	**−170.66**	**−52.67**	**−93.49**	**−175.75**

表 4 MGTM 和对比方法在 7 月数据集上的主题连贯性结果

类型	模型	*K*50			*K*100		
		N=10	*N*=15	*N*=20	*N*=10	*N*=15	*N*=20
仅考虑文本内容的方法	LCTM	−72.78	−160.08	−275.58	−63.56	−137.36	−238.31
	NQTM	−78.35	−185.85	−339.82	−73.31	−178.51	−331.89
整合社交网络结构的方法	LeadLDA	−70.40	−157.83	−268.23	−59.75	−130.83	−226.62
	AdjEnc	−51.72	−123.78	−225.29	−55.73	−140.63	−250.75
	PCFTM	−60.96	−146.28	−266.24	−63.00	−147.89	−268.46
	IATM	−50.75	−119.48	−212.26	−46.80	−120.27	−204.35
	DGTM	−47.70	−108.28	−198.97	−42.37	−99.60	−210.82
我们的方法	MGTM	**−39.62**	**−96.38**	**−184.09**	**−41.98**	**−96.38**	**−179.26**

● GCN 可以聚合更丰富的上下文信息. 为了验证 GCN 的效果, 我们将 MGTM 中的图先验 VAE 退化为具有标准高斯先验的传统 VAE, 记为 MGTM (标准高斯), 其主题连贯性得分如表 5 所示 (后文为了描述方便, 将表 5 中 PCFTM (标准高斯) 简记为 PCFTM, MGTM (图先验) 简记为 MGTM), 进而将其与 PCFTM 和 IATM 比较 (见表 2–表 4), PCFTM 和 IATM 都使用网络嵌入技术来建模社会上下文. 然而, PCFTM 将图结构简化为线性结构, 失去了图中部分复杂关系, IATM 则直接拼接内容特征和结构特征, 这不是最佳的融合方法. MGTM (标准高斯) 使用 GCN 来聚集相邻用户的相关信息. 它直接对图结构执行卷积操作, 并自然地将内容和结构特征融合在一起. 从结果可以看出, MGTM (标准高斯) 比 PCFTM 表现更好, 这表明 GCN 的消息传递有效地缓解了社交媒体帖子的数据稀疏性, 从而获得更连贯的主题. 另外, DGTM 使用双流图卷积网络聚合更加丰富的上下文信息, 在大部分情况下, 其性能优于 PCFTM, IATM 和 MGTM (标准高斯), 也表明了 GCN 在聚合社交媒体信息上的有效性.

● 当主题数 K 固定时, 主题连贯性分数会随着关键词数量 N 的增加而降低. 这可能是因为随着 N 的增加, 生成的主题关键词中容易出现越来越多的无关信息, 计算连贯性得分的公式 (18) 也能反映这一现象.

4.5.2 图先验分布的效果

图先验分布通过在 VAE 主题过程中集成用户交互关系来提高主题连贯性. 通过表 5 观察, 将 MGTM 与 MGTM (标准高斯) 进行比较, 可以看出 MGTM 的连贯性得分更高. 这表明引入的图先验分布有利于社交媒体主题检测. 图先验分布整合了好友之间的相关性, 使得主题推断过程考虑了社交网络的结构特征, 从而产生更连贯的

主题. 我们还将 PCFTM 中 VAE 的先验分布替换为图先验分布进行比较. 结果表明, 采用图先验分布的算法性能得到了提升. 图先验需要每个用户节点的表示, 而 IATM 仅对每条交互边学习唯一的嵌入表示, 因此本文无法验证使用图先验分布的 IATM 的性能. 此外, 由于图先验分布是根据用户间的一阶交互关系构建, 对于社交网络中的二阶, 甚至高阶关系并没有建模. 然而 DGTM 在建模社交上下文时融入了高阶的社交关系, 其节点表示不适合用于构建图先验分布, 因此本文只在 MGTM 和 PCFTM 上进行了图先验分布的验证实验.

表 5　在 MGTM 与 PCFTM 上验证图先验分布的作用

数据集	模型	*K*50			*K*100		
		N=10	*N*=15	*N*=20	*N*=10	*N*=15	*N*=20
5月	MGTM (标准高斯)	−39.56	−95.63	−196.16	−45.16	−116.67	−215.66
	MGTM (图先验)	−31.59	−86.45	−179.76	−40.54	−100.31	−199.88
	PCFTM (标准高斯)	−74.03	−167.52	−303.48	−77.65	−178.18	−317.88
	PCFTM (图先验)	−66.34	−161.74	−300.82	−72.87	−167.60	−302.93
6月	MGTM (标准高斯)	−58.81	−129.53	−226.67	−60.21	−110.14	−191.68
	MGTM (图先验)	−36.25	−80.75	−170.66	−52.67	−93.49	−175.75
	PCFTM (标准高斯)	−77.95	−181.93	−330.45	−79.77	−182.12	−325.38
	PCFTM (图先验)	−68.92	−165.03	−303.04	−72.54	−175.18	−322.99
7月	MGTM (标准高斯)	−48.47	−116.19	−231.45	−46.66	−114.39	−215.58
	MGTM (图先验)	−39.62	−96.38	−184.09	−41.98	−96.38	−179.26
	PCFTM (标准高斯)	−60.96	−146.28	−266.24	−63.00	−147.89	−268.46
	PCFTM (图先验)	−54.73	−131.72	−246.26	−57.79	−138.72	−258.09

4.6 链路预测

除了主题连贯性, 本文还在下游任务——链路预测上验证 MGTM 中用户节点表示的效果, 这是网络表示学习中最常用的下游任务[41]. 基线模型中的主题向量可以在用户级社交网络中预测节点之间的链接关系. 本节依然在 3 个数据集上进行实验, 主题向量的维度设置为 50. 在对比方法中, IATM 学习社交网络中每条交互边的表示, PCFTM 中的用户主题向量随着随机游走的进行而动态变化. 这两个模型都不适合学习用户级社交网络中节点的唯一主题向量, 因此它们没有作为本文的对比方法.

按照如下步骤划分训练集和测试集: 对于每个节点 $v_i \in V$, 如果 v_i 有 3 条以上的边, 则随机将其中一条边选入测试集 E_{test}, 再将其余的所有边放入训练集 E_{train}. MGTM 和对比方法在图 $G_{\text{train}} = \{V, E_{\text{train}}, T\}$ 上训练, 并基于学到的潜在主题向量 z 预测 E_{test} 中的边.

预测与结果评估过程如下: 首先计算图 G 中节点间的距离矩阵 $D \in R^{n \times n}$. 节点 v_i 和节点 v_j 之间的距离, d_{ij} 定义为两个节点之间的欧氏距离, 距离越小意味着节点之间存在链路的可能性越大. 对于评价指标, 选择在链路预测任务中经常使用的 *AUC*, 它的计算方式如下:

$$AUC = \frac{n' + 0.5 \times n''}{n} \tag{18}$$

在评测过程中, 将测试集 E_{test} 中一条边上两个节点的相似度与图 G 中任意两个不存在边的节点 (不存在的边) 的相似度进行比较. 如果测试集中边的相似度大于不存在边的相似度, 则分子加 1, n' 表示这种情况的边的数量. 如果相等, 表示预测结果近似于随机选择, 则分子加 0.5, n'' 表示对应情况下边的数量. 如果小于, 则分子加 0. 分母 n 是测试集中的边与不存在的边的比较次数, 实验中随机选择 20 条不存在的边进行比较. 实验结果如表 6.

从表 6 中可以看出, LCTM 和 NQTM 的表现较差, AdjEnc 和 LeadLDA 的性能较好. NQTM 和 LCTM 仅根据帖子文本内容的相似度预测链路结构. LeadLDA 依托会话树结构, 检测其中的领导者与追随者, 直接将领导者与追随者之间的转发结构嵌入到模型中. AdjEnc 和 DGTM 将网络结构整合到节点表示中, 嵌入了节点间的链接信

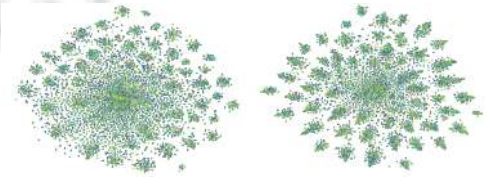
息. 相比仅依据文本内容的方法, 它们都能更加准确地预测用户节点间的潜在关系. 本文提出的模型 MGTM 通过在图先验分布中注入社交结构, 能够更好地捕捉到用户之间的交互关系, 使得潜在的好友间的主题向量具有更高的相似度.

表 6　3 个数据集上链路预测实验的结果

模型	5月数据集	6月数据集	7月数据集
LCTM	0.562	0.576	0.603
NQTM	0.447	0.490	0.490
LeadLDA	0.628	0.642	0.634
AdjEnc	0.746	0.734	0.733
DGTM	0.753	0.772	0.748
MGTM	**0.760**	**0.776**	**0.757**

4.7　用户主题向量可视化

为了进一步验证图先验分布的作用, 本文提取用户的潜在主题向量, 借助可视化工具直观地感受其在空间中的相对位置. 我们首先分别提取退化模型 MGTM (标准高斯) 与 MGTM 计算的潜在主题向量, 进而利用 t-SNE[42] 将每个主题向量降到二维并对其可视化, 如图 3 所示. 其中, 每个点代表一个用户. 每个用户的潜在主题向量通过重参数技巧从主题后验分布中采样得到, 向量维度设置为 50. 图 3(a) 是 MGTM (标准高斯) 推断主题的结果, 图 3(b) 是 MGTM 推断主题的结果. 可以看出, 图 3(b) 中的潜在主题向量聚合得更好, 聚集簇更具可分性. 相反地, 图 3(a) 中更多的点散落在聚集簇之外. 这表明 MGTM 学习的主题向量更好地保持了好友间的相关性, 使得好友之间的主题向量在低维空间中更加相似和接近. 可视化分析进一步证明了图先验分布可以整合用户间的交互特征, 从而使得模型在主题推理阶段能考虑用户间的相关性信息.



(a) MGTM (标准高斯)　　(b) MGTM

图 3　模型推断的潜在主题向量的可视化

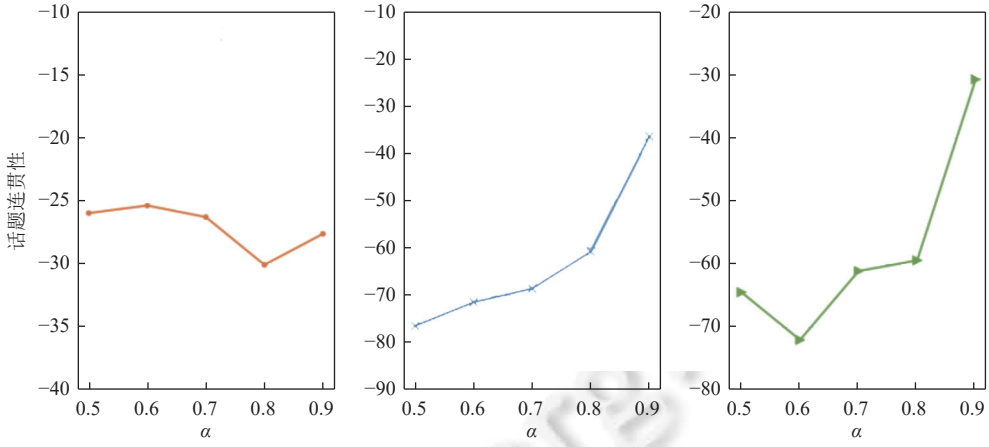
4.8　超参数分析

在本文提出的 MGTM 中有一个重要的超参数: 公式 (13) 中的 α, 它控制着图先验分布中整体的相关性强度. 为了研究不同数据集中这一参数对主题连贯性的影响, 我们将其从 0.5 变化到 0.9 以观察其影响. 从图 4 可以看出, 不同的数据集中最佳的 α 是不同的. 在 5 月数据集中, $\alpha = 0.6$ 时效果最好, 在 6 月、7 月数据集中, $\alpha = 0.9$ 时连贯性得分最高, 且它们的变化趋势也不同. 这表明不同的数据集中存在不同的数据特征, 其中用户间的相关性强度也会随之相应改变.

5　样例分析

为了进一步直观地展示模型提取的主题词, 表 7 和表 8 列出了所有模型推断的关于“韩国明星”和“小米发布会”两个潜在主题的前 10 个主题词, 其中红色斜体字被认为与主题相关度较低. 此外, 表中列出的词按主题词出现的概率进行排序, 位置靠前表示模型认为与主题更相关. 因此, 红色斜体字越少或出现位置越靠后越好. 对比基线模型可以发现: 对于基线模型, 以 LCTM 为例, 在“韩国明星”主题中, 它推断出了“上座率”“海外”“模特”“表现”“主舞”等相关主题词, 然而也错误推断出了“小女孩”“内容”等与主题无关词. 这可能是由于在 LCTM 训练词向量的文本中, 文本形式与表达习惯带来了语义上的偏差, 导致了模型的错误判断. 在 MGTM 提取的主题词中, 非相关词的

个数最少, 但在两个主题中也出现了“科普”“随意”“肢体”等主题无关词. 通过观察数据集, 我们发现这可能是由于相关主题与噪声主题同时受到了很多用户的关注, 导致模型在建模用户交互时引入了噪声信息.



(a) 5 月数据集 (b) 6 月数据集 (c) 7 月数据集

图 4 超参数 α 对 MGTM 主题连贯性的影响

表 7 关于潜在主题“韩国明星”的前 10 个主题词

模型	主题词
LCTM	*小女孩, 内容, 决定*, 选手, 上座率, 海外, *凡哥*, 模特, 表现, 主舞
NQTM	演员, *无暇, 无关*, 首映, *丈母娘, 红色, 走进*, 歌声, *kc党*, 电视剧
LeadLDA	吴亦凡, 公司, EXO, *傻*, 解约, *相信*, 成员, *平凡, 真的*, SM (经济公司)
AdjEnc	*流星*, 金贤重 (明星), *人口, 超级, 要紧*, 表演, 艺名, *痕迹*, 追星, *奶爸*
PCFTM	*带话, 目前*, 德艺, 大地 (歌), 演员, *pls, 晴天*, 袁姗姗, *紫金*, 娱乐圈
IATM	*依靠*, 金俊勉 (偶像), *柔弱, 海底*, 角色, 搭档, *东西*, 伟大, 美丽, 数量
DGTM	心中, *搜索*, 钢琴, 微笑, 金贤重 (明星), 观看, *提示*, 新歌, 呐喊, *美丽*
MGTM	金俊勉, 靠谱, 难忘, 迷倒, 标准, 里面, 坐满, *科普*, 首场 (演唱会), 在

表 8 关于潜在主题“小米发布会”的前 10 个主题词

模型	主题词
LCTM	*虔诚*, 开卖, *世事无常*, *反省*, *倾斜*, 苹果, 金色, 收视率, 创意, *放弃*
NQTM	厉害, 反超, *道路*, 电量, *米饭*, *高尚*, 苹果, *复刻*, *容貌*, *目睹*
LeadLDA	小米, 手机, 新品, 购买, *微博*, *资格*, 专场, 手环, *f码*, *依然*
AdjEnc	包装, 今夏, *帅*, *伙伴*, 周边, *善意*, *白发魔女, 挑到*, good, *争论*
PCFTM	*拿到*, 电池, 卖点, 涨幅, *变成*, *理所当然*, *电视剧*, 压轴, *呼吁*, 科技
IATM	缓存, *黑*, 破万, 雷军, *大声*, *悲剧*, 真棒, *开拍*, 童鞋们, 超值
DGTM	雷总, 结论, 干实事, *语文*, *说说而已*, *了解*, 路由器, 尖端, 亲临, 家居
MGTM	拍摄, 可视化, 安全, 获奖, 智能, *随意*, *肢体*, 最强, 旅游, 稳定

6 总 结

针对社会媒体帖子文本简短, 社交网络结构复杂, 以及现有基于 VAE 的方法, 在主题推理过程中并不能很好地整合用户节点间的复杂关系等问题, 本文提出了一个基于消息传递和图先验分布的社交媒体主题模型 MGTM. 其在两阶段建模用户间的交互关系, 首先在编码社交网络阶段, 图卷积网络通过消息传递机制, 聚合相关的上下文

信息缓解数据稀疏, 并将用户交互结构天然地嵌入到节点的向量表示中. 特别是在基于变分自编码器的主题推理阶段, 与标准 VAE 相比, MGTM 构建图先验分布, 将用户交互整合到潜在主题向量中. 在 3 个真实微博数据集上进行实验, 本文提出的模型获得了更高的连贯性分数, 表明引入图先验分布可以有效提高社交媒体中主题检测的性能, 链路预测和可视化分析也表明 MGTM 中潜在主题向量更好地融入了用户间的交互特征.

References:

[1] Zeng JC, Li J, Song Y, Gao CY, Lyu MR, King I. Topic memory networks for short text classification. In: Proc. of the 2018 Conf. on Empirical Methods in Natural Language Processing. Brussels: Association for Computational Linguistics, 2018. 3120–3131. [doi: 10.18653/v1/D18-1351]

[2] Xu S, Li PF, Kong F, Zhu QM, Zhou GD. Topic tensor network for implicit discourse relation recognition in Chinese. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: IEEE, 2019. 608–618. [doi: 10.18653/v1/P19-1058]

[3] Zou YC, Zhao LJ, Kang YY, Lin J, Peng ML, Jiang ZR, Sun CL, Zhang Q, Huang XJ, Liu XZ. Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. Washington: AAAI Press, 2021. 14665–14673. [doi: 10.1609/aaai.v35i16.17723]

[4] Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. The Journal of Machine Learning Research, 2003, 3: 993–1022.

[5] Zhao WX, Jiang J, Weng JS, He J, Lim EP, Yan HF, Li XM. Comparing Twitter and traditional media using topic models. In: Proc. of the 33rd European Conf. on Advances in Information Retrieval. Dublin: Springer, 2011. 338–349. [doi: 10.1007/978-3-642-20161-5_34]

[6] Mehrotra R, Sanner S, Buntine W, Xie LX. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: Proc. of the 36th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Dublin: ACM, 2013. 889–892. [doi: 10.1145/2484028.2484166]

[7] Quan XJ, Kit C, Ge Y, Pan SJ. Short and sparse text topic modeling via self-aggregation. In: Proc. of the 24th Int'l Conf. on Artificial Intelligence. Buenos Aires: AAAI Press, 2015 2270–2276.

[8] Yan XH, Guo JF, Lan YY, Cheng XQ. A biterm topic model for short texts. In: Proc. of the 22nd Int'l Conf. on World Wide Web. Rio de Janeiro: ACM, 2013. 1445–1456. [doi: 10.1145/2488388.2488514]

[9] Chen WZ, Wang JP, Zhang Y, Yan HF, Li XM. User based aggregation for biterm topic model. In: Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th Int'l Joint Conf. on Natural Language Processing. Beijing: Association for Computational Linguistics, 2015. 489–494. [doi: 10.3115/v1/P15-2080]

[10] Mikolov T, Le Q. Distributed representations of sentences and documents. In: Proc. of the 26th Int'l Conf. on Neural Information Processing Systems. Lake Tahoe, 2013. 3111–3119.

[11] Hu WH, Tsujii JI. A latent concept topic model for robust topic inference using word embeddings. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 380–386. [doi: 10.18653/v1/P16-2062]

[12] Shi T, Kang K, Choo J, Reddy CK. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In: Proc. of the 2018 World Wide Web Conf. Lyon: Int'l World Wide Web Conf. Steering Committee, 2018. 1105–1114. [doi: 10.1145/3178876.3186009]

[13] Guo WY, Wu S, Wang L, Tan TN. Social-relational topic model for social networks. In: Proc. of the 24th ACM Int'l on Conf. on Information and Knowledge Management. Melbourne: ACM, 2015. 1731–1734. [doi: 10.1145/2806416.2806611]

[14] Li J, Liao M, Gao W, He YL, Wong KF. Topic extraction from microblog posts using conversation structures. In: Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Berlin: ACL, 2016. 2114–2123. [doi: 10.18653/v1/P16-1199]

[15] Kingma DP, Welling M. Auto-encoding variational Bayes. In: Proc. of the 2nd Int'l Conf. on Learning Representations. Banff: ICLR, 2014.

[16] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations. San Diego: ICLR, 2014.

[17] He RF, Zhang XF, Jin D, Wang LB, Dang JW, Li XG. Interaction-aware topic model for microblog conversations through network embedding and user attention. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. 1398–1409.

[18] Liu HY, He RF, Wang HC, Wang B. Fusing parallel social contexts within flexible-order proximity for microblog topic detection. In: Proc. of the 29th ACM Int'l Conf. on Information & Knowledge Management. ACM, 2020. 875–884. [doi: 10.1145/3340531.3412024]

[19] Wang HC, He RF, Liu HY, Wu CH, Wang B. Topic model on microblog with dual-streams graph convolution networks. In: Proc. of the

2022 Int'l Joint Conf. on Neural Networks. Padua: IEEE, 2022. 1–8. [doi: 10.1109/IJCNN55064.2022.9892645]

[20] Zhang ZH, Fang M, Chen L, Rad MRN. Is neural topic modelling better than clustering? An empirical study on clustering with contextual embeddings for topics. In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2022. 3886–3893. [doi: 10.18653/v1/2022.naacl-main.285]

[21] Bian T, Xiao X, Xu TY, Zhao PL, Huang WB, Rong Y, Huang JZ. Rumor detection on social media with bi-directional graph convolutional networks. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Washington: AAAI Press, 2020. 549–556. [doi: 10.1609/aaai.v34i01.5393]

[22] Cui P, Hu L, Liu YC. Enhancing extractive text summarization with topic-aware graph neural networks. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 5360–5371. [doi: 10.18653/v1/2020.coling-main.468]

[23] Ou ZJ, Su QL, Yu JX, Liu B, Wang JW, Zhao RH, Chen CY, Zheng YF. Integrating semantics and neighborhood information with graph-driven generative models for document retrieval. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Association for Computational Linguistics, 2021. 2238–2249. [doi: 10.18653/v1/2021.acl-long.174]

[24] Alvarez-Melis D, Saveski M. Topic modeling in Twitter: Aggregating tweets by conversations. In: Proc. of the 10th Int'l AAAI Conf. on Web and Social Media. Limassol: AAAI Press, 2016. 519–522. [doi: 10.1609/icwsm.v10i1.14817]

[25] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proc. of the 2013 Int'l Conf. on Learning Representations. Scottsdale: ICLR, 2013.

[26] Li CL, Wang HR, Zhang ZQ, Sun AX, Ma ZY. Topic modeling for short texts with auxiliary word embeddings. In: Proc. of the 39th Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. Pisa: ACM, 2016. 165–174. [doi: 10.1145/2911451.2911499]

[27] Chen CT, Ren JT. Forum latent Dirichlet allocation for user interest discovery. Knowledge-based Systems, 2017, 126: 1–7. [doi: 10.1016/j.knosys.2017.04.006]

[28] Zheng B, Wen HY, Liang YB, Duan N, Che WX, Jiang DX, Zhou M, Liu T. Document modeling with graph attention networks for multi-grained machine reading comprehension. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 6708–6718. [doi: 10.18653/v1/2020.acl-main.599]

[29] Zhou DY, Hu XM, Wang R. Neural topic modeling by incorporating document relationship graph. In Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 3790–3796. [doi: 10.18653/v1/2020.emnlp-main.310]

[30] Zhang DC, Lauw HW. Dynamic topic models for temporal document networks. In: Proc. of the 39th Int'l Conf. on Machine Learning. Baltimore: PMLR, 2022. 26281–26292.

[31] Qiu JZ, Tang J, Ma H, Dong YX, Wang KS, Tang J. DeepInf: Social influence prediction with deep learning. In: Proc. of the 24th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. London: ACM, 2018. 2110–2119. [doi: 10.1145/3219819.3220077]

[32] Yan HR, Jin XL, Meng XB, Guo JF, Cheng XQ. Event detection with multi-order graph convolution and aggregated attention. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing. Hong Kong: Association for Computational Linguistics, 2019. 5766–5770. [doi: 10.18653/v1/D19-1582]

[33] Chang J, Boyd-Graber J, Gerrish S, Wang C, Blei DM. Reading tea leaves: How humans interpret topic models. In: Proc. of the 22nd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2009. 288–296.

[34] Mimno DM, Wallach HM, Talley EM, Leenders M, McCallum AK. Optimizing semantic coherence in topic models. In: Proc. of the 2011 Conf. on Empirical Methods in Natural Language Processing. Edinburgh: Association for Computational Linguistics, 2011. 262–272.

[35] Wu XB, Li CP, Zhu Y, Miao YS. Short text topic modeling with topic distribution quantization and negative sampling decoder. In: Proc. of the 2020 Conf. on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2020. 1772–1782. [doi: 10.18653/v1/2020.emnlp-main.138]

[36] Zhang C, Lauw HW. Topic modeling on document networks with adjacent-encoder. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. Washington: AAAI Press, 2020. 6737–6745. [doi: 10.1609/aaai.v34i04.6152]

[37] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997, 9(8): 1735–1780. [doi: 10.1162/neco.1997.9.8.1735]

[38] Meng Y, Huang JX, Wang GY, Zhang C, Zhuang HL, Kaplan L, Han JW. Spherical text embedding. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 737.

[39] Zhang LH, Hu XM, Wang BY, Zhou DY, Zhang QW, Cao YB. Pre-training and fine-tuning neural topic model: A simple yet effective approach to incorporating external knowledge. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics

(Vol. 1: Long Papers). Dublin: Association for Computational Linguistics, 2022. 5980–5989. [doi: 10.18653/v1/2022.acl-long.413]

[40] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: Association for Computational Linguistics, 2019. 4171-4186. [doi: 10.18653/v1/N19-1423]

[41] Tu CC, Yang C, Liu ZY, Sun MS. Network representation learning: An overview. Scientia Sinica: Informationis, 2017, 47(8): 980–996 (in Chinese with English abstract). [doi: 10.1360/N112017-00145]

[42] van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(86): 2579–2605.

附中文参考文献:

[41] 涂存超, 杨成, 刘知远, 孙茂松. 网络表示学习综述. 中国科学: 信息科学, 2017, 47(8): 980–996. [doi: 10.1360/N112017-00145]

王浩成(1997－), 男, 硕士, 主要研究领域为社会媒体话题检测.

吴辰昊(1999－), 女, 硕士, 主要研究领域为社会媒体话题检测.

贺瑞芳(1979－), 女, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为自然语言处理, 社会媒体挖掘, 机器学习.

刘焕宇(1996－), 男, 硕士, 主要研究领域为社会媒体摘要.