软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2023,34(11):5162−5178 [doi: 10.13328/j.cnki.jos.006729]

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

基于异构社交上下文的多视图微博主题检测*

贺瑞芳[1,2], 王浩成[1,2], 刘宏宇[1,2], 王 博[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 300350)

通信作者: 王博, E-mail: Bo_wang@tju.edu.cn

摘 要: 社交媒体主题检测旨在从大规模短帖子中挖掘潜在的主题信息. 由于帖子形式简短、表达非正规化, 且社交媒体中用户交互复杂多样, 使得该任务具有一定的挑战性. 前人工作仅考虑了帖子的文本内容, 或者同时对同构情境下的社交上下文进行建模, 忽略了社交网络的异构性. 然而, 不同的用户交互方式, 如转发, 评论等, 可能意味着不同的行为模式和兴趣偏好, 其反映了对主题的不同的关注与理解; 此外, 不同用户对同一主题的发展和演化具有不同影响, 社区中处于引领地位的权威用户相对于普通用户对主题推断会产生更重要的作用. 因此, 提出一种新的多视图主题模型 (multi-view topic model, MVTM), 通过编码微博会话网络中的异构社交上下文来推断更加完整、连贯的主题. 首先根据用户之间的交互关系构建一个属性多元异构会话网络, 并将其分解为具有不同交互语义的多个视图; 接着, 考虑不同交互方式与不同用户的重要性, 借助邻居级注意力和交互级注意力机制, 得到特定视图的嵌入表示; 最后, 设计一个多视图驱动的神经变分推理方法, 以捕捉不同视图之间的深层关联, 并自适应地平衡它们的一致性和独立性, 从而产生更连贯的主题. 在 3 个月新浪微博数据集上的实验结果证明所提方法的有效性.

关键词: 社交媒体主题检测; 异构社交上下文; 多视图; 注意力机制; 神经变分推理

中图法分类号: TP18

中文引用格式: 贺瑞芳, 王浩成, 刘宏宇, 王博. 基于异构社交上下文的多视图微博主题检测. 软件学报, 2023, 34(11): 5162–5178. http://www.jos.org.cn/1000-9825/6729.htm

英文引用格式: He RF, Wang HC, Liu HY, Wang B. Multi-view Microblog Topic Detection Based on Heterogeneous Social Context. Ruan Jian Xue Bao/Journal of Software, 2023, 34(11): 5162–5178 (in Chinese). http://www.jos.org.cn/1000-9825/6729.htm

Multi-view Microblog Topic Detection Based on Heterogeneous Social Context

HE Rui-Fang[1,2], WANG Hao-Cheng[1,2], LIU Hong-Yu[1,2], WANG Bo[1,2]

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

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

Abstract: Social media topic detection aims to mine latent topic information from large-scale short posts. It is a challenging task as posts are short in form and informal in expression and user interactions in social media are complex and diverse. Previous studies only consider the textual content of posts or simultaneously model social contexts in homogeneous situations, ignoring the heterogeneity of social networks. However, different types of user interactions, such as forwarding and commenting, could suggest different behavior patterns and interest preferences and reflect different attention to the topic and understanding of the topic. In addition, different users have different influences on the development and evolution of the same topic. Specifically, compared with ordinary users, the leading authoritative users in a community play a more important role in topic inference. For the above reasons, this study proposes a novel multi-view topic model (MVTM) to infer more complete and coherent topics by encoding heterogeneous social contexts in the microblog conversation network. For this purpose, an attributed multiplex heterogeneous conversation network is built according to the interaction relationships among users

* 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: 2021-09-26; 修改时间: 2022-04-13; 采用时间: 2022-06-10; jos 在线出版时间: 2023-05-18
CNKI 网络首发时间: 2023-05-19

and decomposed into multiple views with different interaction semantics. Then, the embedded representation of specific views is obtained by leveraging neighbor-level and interaction-level attention mechanisms, with due consideration given to different types of interactions and the importance of different users. Finally, a multi-view neural variational inference method is designed to capture the deep correlations among different views and adaptively balance their consistency and independence, thereby obtaining more coherent topics. Experiments are conducted on a Sina Weibo dataset covering three months, and the results reveal the effectiveness of the proposed method.

Key words: social topic detection; heterogeneous social contexts; multiple views; attention mechanism; neural variational inference

随着如 Twitter (https://twitter.com) 和新浪微博 (https://weibo.com) 等社交媒体的繁荣发展, 互联网上每天都会产生数以百万计的帖子. 面向社交媒体的主题挖掘旨在从这些海量、简短、非正规化表达的帖子中挖掘潜在的主题信息, 帮助分析人员快速掌握潜在的语义结构. 此外, 其对短文本分类[1]、关键词生成[2]、篇章关系识别[3]等下游应用也具有重要意义.

主题模型是一种有效的挖掘文本中潜在主题信息的方法. 传统的主题模型, 如 LDA (latent Dirichlet allocation) 模型[4], 利用两个狄利克雷-多项式共轭分布, 建模文档-主题分布与主题-词分布. 本质上, LDA 通过隐式地捕获文档级的词共现模式来揭示主题, 因而在新闻和科技论文等统计信息丰富的长文档上达到了很好的性能. 然而, 社交媒体平台上的短文本, 如新浪微博中的帖子, 不仅长度远远小于标准长文档, 而且表达不正式, 口语化严重. 这些特点使得长文档主题模型在社交媒体场景中缺乏充足的帖子级词共现模式, 表现不能令人满意[5,6].

现有的面向社交媒体的短文本主题挖掘方法从多个方面缓解短文本的数据稀疏、用词随意等问题, 其可以粗略地分为以下几类.

(1) 捕获跨文档的词共现模式. 一些方法[5−9]基于启发式策略将短帖子聚合成一个伪文档, 缓解了短帖子的数据稀疏性. 比如 Mehrotra 等人[7]综合比较: 基于作者的聚合策略、基于爆发指数的聚合策略、基于时间的聚合策略以及基于 hashtag 的聚合策略, 经过对比发现基于 hashtag 的聚合策略的效果更好. 还有方法直接建模词对 (biterm) 的生成[10,11], 比如 Yan 等人[10]为解决微博、推特等短文本中词共现模式稀少的问题, 提出词对主题模型 (biterm topic model, BTM). 该模型首先抽取一个短文本窗口内出现的所有词的共现词对, 然后直接建模共现词对的生成过程, 并假定每个词对中的两个词有相同的主题. 然而, 该类方法遇到较少的语料数据时, 词对共现频率也随之降低, 最终导致性能大幅下降. 并且此类方法通过频率统计推断文档-主题分布以及主题-词分布, 忽略了词与词之间的语义信息.

(2) 基于表示学习. 词嵌入 (word embedding)[12]成为近年来大量模型成功的基石. 每个词被表示成一个低维稠密向量, 语义相近的词的嵌入表示被投影到向量空间中的位置也更加邻近[13]. Sridhar 等人[14]以及 Hu 等人[15]利用词嵌入表示含有丰富语义信息的特点建模主题, 在一定程度上缓解了短文本缺乏语义信息的问题. 为了深层次地理解短文本语义, 这些方法将短文本看作由词嵌入组成的集合, 并假设主题-词分布为多维高斯分布, 利用分层贝叶斯模型推断主题. Wang 等人[16]利用词共现图与语义关系图学习词向量特征, 得到了全局的语义表示. 在其他任务, 如短文本匹配中, Lyu 等人[17]引入 HowNet 网络来增强文本表示, 缓解短文本中严重的歧义问题.

以上方法均忽略了帖子在社交网络中的上下文语境. 实际上, 社交媒体中的文本内容与其所在的网络结构密切相关[18], 仅分析文本内容不够充分.

(3) 整合社交上下文. 简短且不规范的微博帖子中词共现模式表现并不明显. 一些研究开始考虑引入社交上下文来缓解稀疏问题. Li 等人[19]以帖子间的回复和转发关系为基础构建会话树 (conversation tree), 借助树的静态结构信息 (如微博中词语数量、词语词性、URL 数量等静态特征) 将其中的帖子划分为领导者 (leader) 和追随者 (follower), 并对两类文本之间的主题依赖关系进行建模. He 等人[20]在同构社交情景中考虑用户动态交互关系. Liu 等人[21]考虑社交网络中用户间灵活阶的相似度, 同时无缝地融合用户的内容特征与结构特征. 然而, 上述方法都忽视了社交网络的异构性, 将用户间的多种交互方式, 如转发、评论同等对待, 这可能会影响对社交上下文的建模, 进而影响主题推断的质量.

实际上, 在社交网络中, 用户之间有多种类型的交互方式, 如转发、评论、关注等. 图 1 展示了一个典型的微博中用户交互的例子. 基于不同的兴趣和不同的理解, U_2 评论了 U_0 关于“梅西”的帖子, 转发了 U_1 关于“马航坠机”

的帖子. 其中, 评论相比于转发可以包含更多的用户个人观点、支持或反对的理由、甚至情绪信息. 在主题描述上, 这些不包含在原始帖子中的内容进一步丰富了主题信息, 有助于缓解数据稀疏; 而转发贴通常会出现在转发用户的个人主页里, 从而引起更多用户对主题的关注、转发和评论, 本质上扩大了主题的传播范围. 因此, 不同的交互类型对主题描述和主题传播的影响是不同的, 进而对主题推断会产生不同的影响.

[O] U_0: #梅西#的控球技术令人难以置信!
[O] U_1: 悲剧!#马航失事飞机#上有许多医学专家!
[C] U_2 (V): 是的，他的能力令人难以置信。@U_0: #梅西#的控球技术令人难以置信!
[R] U_2 (V): 转发微博。@U_1: 悲剧!#马航失事飞机#上有许多医学专家!
[R] U_3: 转发微博。@U_1: 悲剧!#马航失事飞机#上有许多医学专家!
[R] U_4: 转发微博。U_2 (V): 转发微博。@U_1: 悲剧!#马航失事飞机#上有许多医学专家!
[R] U_5: 转发微博。U_2 (V): 转发微博。@U_1:悲剧!#马航失事飞机#上有许多医学专家!

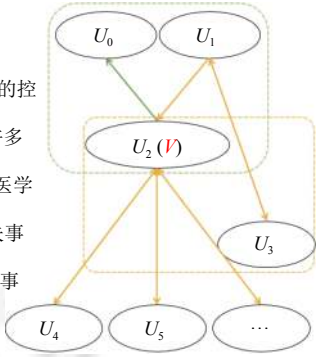


图 1　微博交互的例子

此外, 社交媒体中主题社区内处于引领地位的用户相对于普通用户对主题的发展与检测产生更重要的影响. 考虑到不同交互类型的差异以及不同用户的影响力, 前人方法忽略了上述异构社交媒体情境的上下文信息. 为了充分挖掘异构社交上下文, 本文研究了多样的交互方式以及不同的邻居重要性是否会影响社交媒体中的主题检测, 并提出了一种新的基于编码异构社交上下文的多视图主题模型 (multi-view topic model, MVTM). 该模型通过邻居级注意力和交互级的注意力, 将用户的不同邻居和多种的交互方式对主题的影响嵌入到多个特定视图的编码表示中. 接着, 设计多视图驱动的神经变分推理 (neural variational inference, NVI) 来捕捉不同视图语义之间的复杂关联, 生成更连贯的主题.

本文第 1 节介绍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 节给出属性多元异构会话网络的定义. 第 3 节详细论述本文提出的基于异构社交媒体用户动态行为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 节对本工作进行总结和展望.

1　相关工作

近年来主题检测成为一个重要的研究方向, 涌现了很多优秀的研究[22−25]. 这些方法利用最新的深度学习技术, 比如强化学习、对抗训练等方法解决主题检测面临的各种挑战. 虽然这些方法在标准的长文档上取得了很好的效果, 然而它们没有考虑社交媒体中帖子简短且表达不正式的特性. 聚焦到面向社交媒体的主题检测的研究, 现有工作可以大致地分为两类: (1) 仅考虑文本内容; (2) 整合社交上下文.

1.1　仅考虑文本内容

此类方法只建模帖子中的文本内容, 生成文档-主题分布以及主题-词语分布. 具体可分为: (1) 基于集聚策略. 由于短文本的数据稀疏性导致传统的主题模型无法直接应用于短文本的主题挖掘, 文献 [5−8] 基于启发式的聚集策略将短文本聚集成长文档, 聚集策略包括基于文本作者关系[5,6], 或基于主题标签 (hashtag) 关系[7,8]. 聚合之后利用层次贝叶斯结构的模型挖掘主题信息. 为克服模型对启发式规则的依赖, 一些文献提出伪文档的概念. 如 Quan 等人[9]提出基于自聚合的模型 (self-aggregation based topic model, SATM). 它采用基于短文本主题亲和度的聚合策略, 该方法将短文本按照主题相近程度聚合成伪文档, 之后采用 3 层贝叶斯结构的主题模型挖掘其主题信息. (2) 基于词对的方法. 另外一些研究[10,11]直接对一个短文本窗口内出现的所有词的共现词对 (biterm) 建模. 如 Yan 等人[10]提出词对主题模型 (biterm topic model, BTM), 该模型直接建模词对的生成过程, 同时利用这个语料的词对共现模式, 推断整个语料库中全局的主题分布. 利用语料库中相对更多的词对共现模式缓解数据稀疏. 然而当微博

数量减少时, 基于聚合和基于词对的方法所依赖的跨文档共现模式也会大大减少, 此外, 这些方法完全忽略了词语本身携带的语义信息. (3) 基于嵌入表示. 词嵌入[12]是指学习特定语境中单词特征表示的技术, 其将语料中的词语映射为低维稠密的向量. 通过该技术训练得到的分布式词向量具有词语的语义信息. 即语义相近的词嵌入表示投影到表示空间上也位置也更加接近[13]. 由于词嵌入 (word embedding) 携带有该词语的语义信息, 文献 [14,15] 利用词嵌入表示来丰富短文本语义. Sridhar 等人[14]以及 Hu 等人[15]为了更好地理解文本语义, 将文档看作由词嵌入组成的集合, 无需聚合短文本内容, 利用层次贝叶斯模型推断主题. Wang 等人[16]借助词共现图与语义关系图来学习全局语义表示. Lyu 等人[17]引入 HowNet 网络来增强文本表示.

然而, 社交媒体平台允许用户通过转发、评论微博帖子进行交流. 因此, 社交媒体情境下的内容和结构是相互关联的[18], 而仅关注文本内容的方法忽略了社交上下文对主题检测的影响.

1.2 整合社交上下文

同时考虑短文本的内容和社交网络结构, 该类方法进一步可分为基于静态网络结构和动态社交关系. (1) 基于静态网络结构. Li 等人[19]提出了 LeadLDA 模型, 将微博帖子及其相应的转发和评论微博按照转发和评论关系组织成会话树 (conversation tree), 并借助会话树的静态信息 (如短文本中词语数量、词语词性、URL 数量等静态特征) 将该树上的每一条微博帖子区分为领导者 (leader) 和跟随者 (follower). 领导者是在主题描述上含有简洁新内容的微博, 而跟随者是对领导者微博的简单重复. 该模型根据会话树上领导者和跟随者的主题依赖关系生成潜在主题. Lim 等人[26]通过利用社交媒体上微博作者、主题标引以及用户粉丝网络等额外信息辅助缓解短文本缺乏词共现的问题, 提出 TNTM (Twitter-network topic model) 模型. 该模型以贝叶斯非参数方式对文本内容以及用户粉丝网络同时建模进行主题挖掘. 然而存在为了提升自己的影响力而购买僵尸粉的现象. 僵尸粉仅关注其买主而且长时间没有动态更新, 即不会发出任何帖子. 故僵尸粉的存在会干扰用户粉丝网络, 从而导致该网络结构变得不可靠. 这使得利用该网络的主题模型性能欠佳. (2) 基于动态社交关系. IATM[20]整合了动态交互行为, 结合了文本内容以及会话网络结构信息, 并将两者通过网络表示学习和注意力机制编码为交互感知的边嵌入表示. 随后将该边嵌入表示输入到变分自编码器中, 从而生成连贯性更佳的主题. PCFTM[21]在此基础上建模灵活阶的用户相似度, 效果进一步提升.

以上方法均忽略了社交网络的异构性. 本文重点探索建模不同类型的交互关系以及不同的邻居用户是否有助于面向社交媒体的短文本主题挖掘.

1.3 异构网络嵌入以及神经变分推理

网络表示学习[27,28]旨在为网络中的节点学习低维、稠密的向量表示. 最近关于异构网络表示学习的研究[29]启发我们对社交网络中不同用户和交互关系的重要性进行建模, 整合异构社交上下文. 同时, 深度神经网络在函数近似上表现十分出色, 并且其应用于无监督模型中可以学习得到与数据较为一致的复杂分布. 神经变分推理是一种深度神经网络, 可为主题模型提供有效的变分推断算法. 其利用神经网络分别建模文档-主题分布和主题-词语分布. Miao 等人[30]以及 Srivastava 等人[31]将神经变分推理运用于主题检测中, 使用编码器计算得到话题后验分布的均值和方差, 利用解码器重构文本输入. He 等人[20,21]将神经变分推理应用于社交媒体数据, 同时建模用户的动态交互行为. 然而这些工作均没有考虑社交网络的异构性. 我们尝试建模异构社交媒体情境下的用户交互行为, 检测微博帖子中的潜在主题. 综上所述, 我们试图深入挖掘异构社交上下文, 借鉴异构网络表示学习以及神经变分推理的研究成果来推动社交媒体主题检测的发展.

2 属性多元异构会话网络

社交媒体中存在的大规模嘈杂短文本使得数据稀疏性以及主题不连贯性问题十分尖锐. 如图 1 所示, 社交媒体中用户的不同交互行为给我们带来了机遇. 本文为挖掘异构社交上下文而构建了如图 2 所示的属性多元异构会话网络, 下面给出了该网络的具体定义.

在社交媒体平台上, 用户之间通过转发、评论以及提及等方式相互交流. 基于不同类型的动态用户行为模式, 我们构建属性多元异构会话网络 (attributed multiplex heterogeneous conversation network, AMHCN). 将其形式化为

网络 $G=(V, E, T)$, $E=U_{r\in R} E_r$, 其中 V 表示节点集, 包含网络 G 中所有的用户, E 表示边集, 包含网络 G 中所有类型的用户-用户交互边. 交互边的类型有多种, 对应多元异构, E_r 是 E 的子集, 由交互类型为 r 的边组成, $r\in R$ 指代转发或评论关系, 且$|R|>1$, E_r 是所有交互类型为 r 的边构成的集合, T 是节点属性信息的集合, 包含网络 G 中所有用户节点的帖子文本信息, 对应属性多元异构中属性. 如图 2 所示, 网络中的节点集 V 共包含 5 个用户, $|R|=2$ 表示共有两种类型的交互关系, 分别为转发和评论. 每个用户所发表的帖子内容构成了属性信息的集合 T. 为了初步缓解数据稀疏问题, 我们采用了一种基于用户的聚合策略. 具体来说, 将同一用户发表的所有帖子消息, 包括原始微博、评论和转发微博, 聚合在一起成为用户的帖子文本信息. 以网络 G 中的用户节点 $u\in V$ 为例, 其聚合后的文本信息表示为 $M_u=(w_1, w_2, \ldots, w_n)$, 其中 n 是文本信息 M_u 中词语的个数.

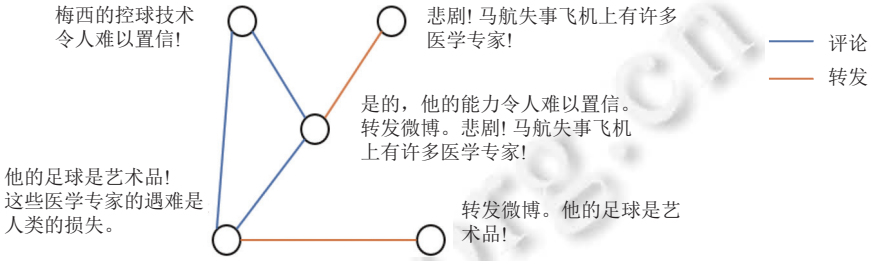


图 2 属性多元异构会话网络

3 MVTM 研究框架

MVTM 模型整体框架如图 3 所示, 主要包含两个模块: 特定视图的嵌入和基于多视图神经变分推理的主题检测. 给定属性多元异构会话网络 AMHCN, 我们尝试编码异构的社交上下文. 首先按照交互类型的不同, 将网络分解为多个视图, 如转发视图 (reposting view) 和评论视图 (commenting view), 同时利用邻居级注意力机制和交互级注意力机制建模不同邻居和不同交互类型对主题检测的影响, 学习特定视图的嵌入; 然后将多个视图的嵌入同时输入至神经变分推理来推断主题. 下面将对这两部分依次作详细介绍.

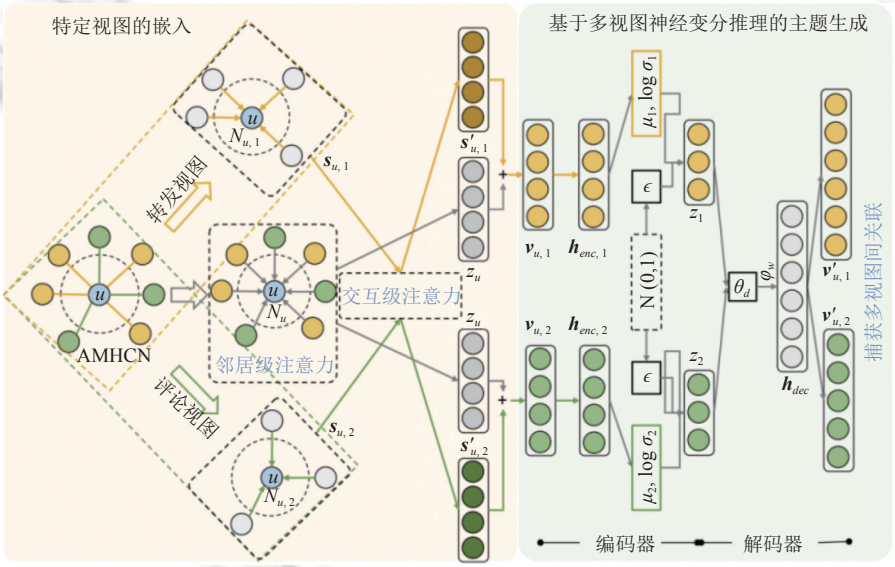


图 3 MVTM 的整体框架

3.1 特定视图的嵌入

社交媒体中通常会出现关于某个热门主题 (topic) 发生一系列社交事件. 基于对主题产生的不同兴趣偏好以

及多样的思考理解, 用户产生如转发、评论等不同行为. 这些不同类型的交互关系蕴含了不同的主题语义, 因而属性多元异构会话网络 AMHCN 呈现出携带有不同交互语义的多个视图. 为捕获不同视图级主题语义间的复杂关联来更好地检测主题, 我们按照交互类型的不同将网络分解为多个视图, 每个视图记为 $G_r=(V_r, E_r, T_r)$, r 表示交互类型. 在本节中, 我们为每个用户学习关于视图 G_r 的嵌入, 其由两个部分组成: 用户嵌入和交互类型嵌入. 需要注意的是, 用户嵌入在不同的视图间是共享的.

● 基于邻居级注意力机制的用户嵌入. 如前文所述, 处于引领地位的权威用户倾向于对主题推断产生更多的贡献. 为挖掘不同用户对主题的不同重要性并降低噪声信息对主题推断的影响, 我们设计邻居级注意力机制为网络中的用户学习统一的表示, 我们称之为用户嵌入.

以任一用户 $u \in V$ 为例, 首先需要将其文本信息 $\boldsymbol{M}_u$ 处理为一个特征向量, 为减少手动干预, 我们采用了统一的嵌入层:

$$\boldsymbol{h}_u = Emb(\boldsymbol{M}_u) \tag{1}$$

$Emb(\cdot)$ 函数可以是任何特征学习的神经网络, 本文采用了词嵌入[32]和全连接层进行语义特征学习. 其还可以替换为预训练模型的词向量, 但本文的重点是探索异构社交上下文对话题检测的影响, 因此不再聚焦于初始的文本表示. $\boldsymbol{h}_u$ 为用户 u 初始的表示向量. 相似地, 用户 $u \in V$ 的所有邻居, 包括在每种交互类型下的邻居, 都得到了唯一的语义特征表示向量. 经验地, 考虑到直接相连的用户通常在主题描述中发挥更重要的作用. 因此, 本文仅建模了用户 $u \in V$ 的一阶邻居 N_u 对主题推断的影响. 给定一阶邻居用户对 (u, v), 设计邻居级注意力机制学习邻居 v 对用户 u 的重要性系数 e_{uv}. 其本质是一个自注意力机制, 如下公式所示:

$$e_{uv} = att(\boldsymbol{h}_u, \boldsymbol{h}_v) \tag{2}$$

其中, att 指代执行邻居级注意力机制的神经网络, $\boldsymbol{h}_u$ 和 $\boldsymbol{h}_v$ 分别是用户 u 和用户 v 的语义特征表示向量. 为进行公平比较, att 的参数对所有的用户对共享. 为使得重要性系数 e_{uv} 易于比较, 我们通过 *Softmax* 函数对其作归一化, 进一步得到邻居 v 对用户 u 的权重系数 β_{uv}. 计算重要性系数与归一化的过程可以形式化为下面的公式.

$$\beta_{uv} = \frac{\exp\left(\sigma\left(a^{\mathrm{T}}\left[\boldsymbol{h}_u \| \boldsymbol{h}_v\right]\right)\right)}{\sum_{j \in N_u} \exp\left(\sigma\left(a^{\mathrm{T}}\left[\boldsymbol{h}_u \| \boldsymbol{h}_j\right]\right)\right)} \tag{3}$$

其中, σ 指代激活函数, a 邻居级注意力向量, T 指代矩阵或向量的转置操作, ||指代拼接操作, N_u 指代用户 $u \in V$ 的一阶邻居. 通过对用户 u 的所有一阶邻居的语义特征表示向量作加权求和, 我们进一步得到用户嵌入 z_u, 其挖掘了不同一阶邻居的重要性并降低了噪声信息的权重, 且其对于所有视图是共享的, 即用户 u 在所有视图中共享一个用户嵌入 z_u.

$$z_u = \sigma\left(\sum_{j \in N_u} \beta_{uj} \cdot \boldsymbol{h}_j\right) \tag{4}$$

其中, σ 指代激活函数, $\boldsymbol{h}_j$ 是用户 u 的一阶邻居 $j \in N_u$ 的语义特征表示.

● 基于交互级注意力机制的交互类型嵌入. 在视图 $G_r=(V_r, E_r, T_r)$ 中, E_r 由所有交互类型为 r 的用户-用户边组成. 由于不同类型的交互关系表明用户对社交媒体事件产生的不同程度的兴趣和多样的理解, 因此本文采用交互级注意力机制捕获不同类型交互关系间的相互影响, 并获得特定交互类型嵌入. 对于一个视图 G_r, 首先聚合每个用户在该视图下所有一阶邻居的交互类型嵌入, 得到用户 u 关于交互类型 r 的初始表示 $\boldsymbol{s}_{u,r}$:

$$\boldsymbol{s}_{u,r} = \sigma\left(\boldsymbol{W}_s \cdot \mathrm{mean}\left(\left\{\boldsymbol{s}_{j,r}, \forall j \in N_{u,r}\right\}\right)\right) \tag{5}$$

其中, σ 指代激活函数, $\boldsymbol{W}_s$ 指可训练的参数矩阵, mean 指代均值函数, $\boldsymbol{s}_{j,r}$ 指代用户 u 在交互类型 r 下的邻居 j 的交互类型嵌入, $N_{u,r}$ 表示用户 u 在交互类型 r 下的所有一阶邻居, 即在该视图 G_r 中的所有一阶邻居. 为建模包括交互类型 r 在内的其他交互类型对当前交互类型的影响 $\boldsymbol{t}_{u,r}$, 我们首先将用户 u 的所有交互类型嵌入拼接起来:

$$\boldsymbol{C}_u = concat(\boldsymbol{s}_{u,1}, \boldsymbol{s}_{u,2}, \ldots, \boldsymbol{s}_{u,m}) \tag{6}$$

其中, m 指代交互类型的数量, $concat$ 指代拼接函数, $\boldsymbol{s}_{u,r}$ 指代用户 u 关于交互类型 i 的嵌入. 然后采用交互级注意力机制计算各个交互类型嵌入分量对当前交互类型 r 的重要性 $\boldsymbol{t}_{u,r}$, 其本质是一个自注意力机制:

$$\boldsymbol{t}_{u,r}=Softmax\left(\boldsymbol{w}_r^{\mathrm{T}}\tanh\left(\boldsymbol{W}_r\boldsymbol{C}_u\right)\right)^{\mathrm{T}} \tag{7}$$

其中, $\boldsymbol{w}_r$ 和 $\boldsymbol{W}_r$ 是关于交互类型 r 的可训练的参数, 上标 T 指代向量或矩阵的转置操作, $\boldsymbol{C}_u$ 是用户 u 的所有交互类型嵌入的拼接. 最后, 用户 u 关于交互类型 r 的嵌入计算如下:

$$\boldsymbol{s}'_{u,r}=\boldsymbol{M}_r^{\mathrm{T}}\boldsymbol{C}_u\boldsymbol{t}_{u,r} \tag{8}$$

其中, $\boldsymbol{M}_r$ 是可训练的参数矩阵. $\boldsymbol{s}'_{u,r}$ 为计算得到的用户 u 关于交互类型 r 的嵌入表示.

● 视图嵌入. 通过邻居级注意力机制和交互级注意力机制分别捕获不同邻居的重要性和不同类型交互关系间的相互影响, 学习得到用户嵌入和交互关系嵌入, 我们可以进一步得到用户 u 关于视图 G_r 的嵌入:

$$\boldsymbol{v}_{u,r}=\boldsymbol{z}_u+\alpha\boldsymbol{s}'_{u,r} \tag{9}$$

其中, α 是一个超参数, 指代交互类型嵌入 $\boldsymbol{s}'_{u,r}$ 对于整个视图嵌入 $\boldsymbol{v}_{u,r}$ 的重要性, $\boldsymbol{z}_u$ 是在不同视图间共享的用户嵌入. 通过将不同的交互类型嵌入与用户嵌入相加, 可以得到不同视图的嵌入表示. 为建模不同邻居和不同类型交互关系对主题检测的影响, 我们为特定视图嵌入部分定义了如下目标函数:

$$L_v=\sum_{j\in\boldsymbol{C}_r}-\log p(j\mid u) \tag{10}$$

其中, $\boldsymbol{C}_r\in N_{u,r}$ 指代用户 u 在视图 G_r 中的上下文用户节点, $p(j|u)$ 指代给定用户 u 后生成用户 j 的概率. 我们利用 *Softmax* 函数归一化, 如下所示:

$$\log P(j\mid u)=\frac{\exp\left(\boldsymbol{c}_{j,r}^{\mathrm{T}}\cdot\boldsymbol{v}_{u,r}\right)}{\sum_{k\in\boldsymbol{C}_r}\exp\left(\boldsymbol{c}_{k,r}^{\mathrm{T}}\cdot\boldsymbol{v}_{u,r}\right)} \tag{11}$$

其中, $\boldsymbol{c}_{k,r}^{\mathrm{T}}$ 是任意上下文用户节点 $k\in\boldsymbol{C}_r$ 关于视图 G_r 的嵌入. 为减少生成概率 $p(j|u)$ 的计算代价, 我们进一步使用负采样技术进行优化, 得到下列的最终目标函数:

$$L_v=\sum\nolimits_{j\in\boldsymbol{C}_r}\left(-\log\sigma\left(\boldsymbol{c}_{j,r}^{\mathrm{T}}\cdot\boldsymbol{v}_{u,r}\right)-\sum\nolimits_{l=1}^{L}E_{k\sim V}\left[\log\sigma\left(-\boldsymbol{c}_{k,r}^{\mathrm{T}}\cdot\boldsymbol{v}_{u,r}\right)\right]\right) \tag{12}$$

其中, σ 是 Sigmoid 函数, $\boldsymbol{v}_{u,r}$ 是用户 u 关于视图 G_r 的嵌入, L 指代负采样的数量, k 指代采样自用户集合 V 的噪声用户节点.

3.2 基于多视图神经变分推理的主题生成

近些年, 具有出色函数近似能力和数据拟合能力的深度神经网络在无监督模型中得到广泛应用, 神经变分推理[33]首次将神经网络应用于主题推理, 通过用变分分布逼近真实的后验分布提供了有效的主题推断算法, 使得生成的主题连贯性更佳. 与此同时, 属性多元异构会话网络中转发、评论等多个视图的嵌入从不同方面丰富了主题语义. 为了进一步深入挖掘不同视图级主题语义间的复杂关联, 受神经主题模型[30,31]所启发, 本文提出一个新的多视图神经变分推断方法. 具体来说, 将第 3.1 节得到的多个视图嵌入同时送到多输入的变分自编码器中, 捕获多重交互语义间的复杂关联, 得到文档-主题分布和主题-词分布.

● 文档-主题分布. 此处, 文档指代聚合的用户文本信息, 用 d 表示. 此时, 文档主题分布可表示为 $\theta_d=(p(t_1|d),p(t_2|d),\ldots,p(t_K|d))$, 其中 K 指代主题的数量, t_i 指代第 i 个主题, $p(t_i|d)$ 指代文档 d 包含第 i 个主题的概率.

给定多个视图嵌入, 首先将其编码到非线性的隐式空间, 以第 i 个视图为例:

$$\boldsymbol{h}_{enc,i}=ReLU\left(\boldsymbol{W}^h\boldsymbol{v}_{u,i}+\boldsymbol{b}^h\right) \tag{13}$$

其中, *ReLU* 是非线性的激活函数, $\boldsymbol{W}^h$ 和 $\boldsymbol{b}^h$ 是编码器神经网络的参数, $\boldsymbol{v}_{u,i}$ 是第 i 个视图的嵌入. 我们采用与神经主题模型一致的做法, 使用隐空间的表示 $\boldsymbol{h}_{enc,i}$ 来学习特定视图包含的主题信息, 其服从高斯分布, 利用两个独立的神经网络计算其均值与方差:

$$\boldsymbol{u}_i=\boldsymbol{W}^u\boldsymbol{h}_{enc,i}+\boldsymbol{b}^u \tag{14}$$

$$\log\boldsymbol{\sigma}_i^2=\boldsymbol{W}^\sigma\boldsymbol{h}_{enc,i}+\boldsymbol{b}^\sigma \tag{15}$$

其中, $\boldsymbol{u}_i$ 和 $\boldsymbol{\sigma}_i^2$ 分别为与第 i 个视图对应的后验高斯分布的均值和方差. 通过使用重参数化技巧得到相应视图的潜

在语义向量$z_i = u_i + \epsilon \times \sigma_i$, 其中为捕获不同视图级语义信息间的复杂关联, 首先将所有视图的潜在语义向量z_i, $i \in \{1, \ldots, m\}$共同编码为稠密空间中的表示$z \in R^K$, 该表示保留了来自不同视图的本质信息:

$$z = ReLU(\boldsymbol{W}^z concat[z_1, \ldots, z_m] + \boldsymbol{b}^z) \tag{16}$$

其中, *concat* 指代拼接操作, $\boldsymbol{W}^z$ 和 $\boldsymbol{b}^z$ 是可训练的参数. 文档主题分布 θ_d 可通过 *Softmax* 函数对 z 作归一化得到.

● 主题-词分布. 我们采用与文献 [20] 类似的做法, 将主题词分布 $\phi_w = (p(w|t_1), p(w|t_2), \ldots, p(w|t_K))$ 作为解码器神经网络的参数, 解码器如下公式所示:

$$\boldsymbol{h}_{dec} = Softmax\left(\phi_w \times \theta_d^{\mathrm{T}}\right) \tag{17}$$

之后通过全连接神经网络层得到重构的多个视图嵌入:

$$\boldsymbol{v}'_{u,i} = ReLU\left(\boldsymbol{W}^{d,i}\boldsymbol{h}_{dec} + \boldsymbol{b}^{d,i}\right) \tag{18}$$

如上所述, 不同的视图嵌入从多个方面反映了主题的语义聚合, 其深度关联可通过模型中的两个方面捕获: (1) 交互级注意力机制捕获了不同类型交互关系间的相互影响, 学习了融合其他交互类型对当前交互类型影响的特定交互类型嵌入表示. (2) 多视图神经变分推理以多个视图表示作为输入, 将来自不同视图的多方面主题语义通过关联的非线性神经网络整合至隐式空间的稠密表示 z 中, 其保留了不同视图携带的不同主题语义的本质信息; 此外, 多视图神经变分推理通过重构多个视图嵌入, 平衡了不同视图级主题语义间的一致性和独立性, 进一步促进了不同主题语义的本质信息的保留.

在解码过程中, 多视图神经变分推理适应性地平衡了不同视图级主题语义间的一致性和独立性, 共同促进主题生成, 该部分的目标函数定义如下:

$$L_g = \sum_{i=1}^{m}(-E_{z_i} \sim p(z_i \mid \boldsymbol{v}_{u,i})[\boldsymbol{v}_{u,i} \mid z_i] + KL(q(z_i) \| p(z_i \mid \boldsymbol{v}_{u,i}))) \tag{19}$$

其中, 第 1 项表示重构期望, 第 2 项表示 *KL* 散度, 用于衡量先验分布和后验分布的相近程度, $q(z_i)$ 是先验高斯分布 $N(0; \mathbf{I})$. 模型需最大化重构期望和最小化 *KL* 散度, 保证生成主题的一致性.

3.3 模型训练

结合上述特定视图嵌入和多视图神经变分推理这两个模块, 通过最小化如下公式所示的整体目标函数来学习模型参数.

$$L = L_v + \lambda L_g \tag{20}$$

其中, λ 用于控制两个模块间的平衡. 在模型训练时, 采用了 Adam[34]优化算法; 在生成文档主题分布和主题词分布时, 采用了 dropout[35]技术.

4 实验结果与分析

4.1 实验数据

实验基于 Li 等人[19]构建的新浪微博原始语料. 该语料搜集了新浪微博平台上从 2014 年 5 月 1 日到 2014 年 7 月 31 日期间 50 个高频标签的微博, 并按照月份划分为 3 个数据集. 目前这份语料是公开的, 并广泛应用于社交媒体主题建模的任务[17,18]. 我们对这 3 个数据集按照如下步骤进行预处理: (1) 考虑本文需要对不同的交互方式建模, 因而过滤掉没有交互记录的用户; (2) 将同一个用户的所有原始微博、转发微博、评论微博聚合起来, 形成用户的文本信息, 以初步缓解数据稀疏. 按照以上步骤, 我们得到了用于评估模型性能的 3 个月份数据集. 统计信息如表 1 所示.

表 1 数据集的统计信息

月份	用户数	转发数	评论数	微博帖子数
5月	44395	27666	36626	70893
6月	89979	59855	90597	163420
7月	119169	90597	87557	188657

4.2 评价指标

由于社交文本具有严重的数据稀疏性, 面向社交媒体领域的主题检测模型在本质上是很难评估的. 在以往研究中, 困惑度是一种常用的评估推断主题质量的指标, 用于衡量在包含不可见词的外部数据集上主题模型的预测能力. 然而, 有研究[36]证明了具有高困惑度的模型不一定生成人类感知中语义连贯的主题, 即困惑度高不一定表明生成的主题在语义上是连贯的. 为进行客观且有意义的比较, 我们延续文献 [37] 的做法, 采用目前广泛使用的且更适合社交媒体文本的主题连贯性指标, 为所有方法计算主题的连贯性分数来评估主题模型的性能. 具体来说, 主题连贯性分数的计算如下:

$$C=\frac{1}{K}\sum_{k=1}^{K}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\log\frac{D\left(w_i^k,w_j^k\right)+1}{D(w_j^k)} \tag{21}$$

其中, K 指代主题的数量, N 表示用于计算连贯性分数的词语个数, 即按照主题词概率从高到低进行排序的前 N 个词语用于计算连贯性分数. w_k^i 是按照主题词分布排序的主题 k 中第 i 个词, $D(w_i^k,w_j^k)$ 表示词 w_i^k 和词 w_j^k 共同出现的文档个数, $D(w_j^k)$ 指代包含词 w_j^k 的文档个数. 此处, 文档指聚合用户的原始消息、转发消息和评论消息得到的用户文本信息.

4.3 对比方法选择与超参数设置

为了验证 MVTM 的性能, 我们选择与几个具有代表性的基线模型进行对比. 如第 1 节所述, 本文将面向社交媒体的主题挖掘方法分为仅考虑文本内容的算法和整合文本内容以及网络结构的算法两大类. 因此, 我们选取的基线模型也来自这两类. 此外, 我们还研究了长文档主题挖掘的最新研究[22−25], 其中 BAT[24]属于神经主题模型, 以词袋向量作为输入, 基于双向对抗网络进行主题推断, 我们选择其进行对比, 验证以词袋向量为基础的对抗训练在社交媒体主题检测中的效果; AdjEnc[25]同样将数据中的网络结构信息整合到模型中, 我们选择其作为对比方法以验证整合网络结构对长文档数据的作用. 基线模型的具体说明如下.

● 只考虑文本内容

(1) BAT[24] 没有选择变分自编码器的推断框架, 而是探索了双向对抗训练在神经主题模型中的应用. 它没有考虑微博帖子简短的特性, 在应用于社交媒体数据时面临严重的稀疏问题.

(2) LCTM[15]通过潜在概念的共现模式来揭示主题, 这种共现模式考虑了文本的语义信息. 已有研究证明 LCTM 在社交媒体上的性能优于原始 LDA.

● 整合社交上下文

(3) LeadLDA[19] 将微博消息按照转发、评论关系组织成对话树, 并将微博消息按照包含关键主题词的概率程度分为领导者消息和跟随者消息, 并根据相互之间的主题依赖来增强推断主题的连贯性.

(4) AdjEnc[25]将网络结构引入到结构化长文档 (如学术论文、网页) 的主题推理中.

(5) ForumLDA[38]联合建模了原帖子、相关和不相关回复帖子的生成过程来推断主题.

(6) IATM[20]采用网络表示学习方法[39,40]对动态用户行为进行建模, 并采用神经变分推理生成主题.

(7) PCFTM[21]整合高阶邻域内的用户交互关系, 并通过平行社交上下文捕获网络结构与帖子文本之间的非线性关系. 由于 MVTM 只建模了用户的一阶邻域, 为了公平地比较并证明建模异构社交上下文的有效性, 我们选择 PCFTM 的变体 PCFTM(-seq) 进行比较, 其仅考虑了用户间的一阶相似度.

同时, 为了验证在第 3 节提出的邻居级注意力和交互级注意力是否有利于主题挖掘, 我们设计了本文提出模型 MVTM 的两个变种模型: MVTM(-nei) 和 MVTM(-mul).

(8) MVTM(-nei) 忽略了用户邻居的不同重要性. 此时邻居级注意力不被使用.

(9) MVTM(-mul) 忽略了 AMHCN 中不同交互方式对主题的影响. 此时不使用交互级的注意力, 也没有多个视图嵌入.

基线模型中概率生成模型的超参数已调节至在本文所采用的数据集上表现最佳时的设置. 具体地, 文档-主题分布的狄利克雷先验分布的超参数 α=50/K, β=50/K (K 代表主题数). LCTM 中概念-词分布的多维高斯先验分布的超参数 μ 设定为所有词嵌入的平均值; LeadLDA 中领导者消息 (leader) 与追随者消息 (follower) 之间主题依赖的

狄利克雷先验超参数 γ=50/K, 主题-词分布的对称贝塔 (beta) 先验分布的超参数 δ=0.5; ForumLDA 中相关主题 (serious topic) 的贝塔先验分布的超参数 γ_0=1, γ_1=5, 不相关主题数 S=1. 为了保证 LCTM、LeadLDA 以及 ForumLDA 收敛, 我们运行吉布斯采样 (Gibbs sampling) 1 000 次. BAT、AdjEnc 以及 IATM 都采用了论文中所报告的参数并训练至收敛.

本文提出的 MVTM, 使用 Adam 进行优化, 学习率设定为 0.001. 嵌入维数设置为 200, 窗口大小设置为 4, 负样本数设置为 5. 损失函数中的系数 λ 设置为 0.8. 其他参数从正态分布 $N(1.0, 0.3^2)$ 中随机初始化得到. 主题数 K 设置为 50 和 100, 主题词的数量 N 设置为 10、15、20. 相关参数调节实验, 我们将在第 4.6 节进行分析.

4.4 实验结果分析

4.4.1 本文方法与基线模型的比较

为验证本文方法的有效性, 我们通过设置不同的主题数{50, 100}与不同数量的主题词{10, 15, 20}, 将本文方法与各种模型挖掘的主题进行对比. 实验结果如表 2–表 5 所示, 我们可以观察到如下情况.

表 2 所有模型在 5 月数据集的实验性能

模型		*K*50			*K*100		
		N=10	*N*=15	*N*=20	*N*=10	*N*=15	*N*=20
仅考虑文本的方法	BAT	−98.49	−233.32	−422.48	−86.14	−201.02	−363.78
	LCTM	−70.91	−165.37	−296.36	−58.65	−140.10	−261.40
整合社交上下文的方法	LeadLDA	−53.91	−138.53	−258.38	−58.15	−141.34	−261.65
	AdjEnc	−67.57	−159.66	−290.10	−72.02	−165.87	−303.37
	ForumLDA	−55.76	−129.57	−231.90	−55.84	−132.23	−236.89
	IATM	−58.75	−112.64	−228.27	−47.32	−121.46	219.96
	PCFTM(-seq)	−31.34	−73.19	−131.65	−32.11	−72.43	−128.84
本文方法	**MVTM**	**−30.40**	**−70.04**	**−120.36**	**−26.94**	**−63.04**	**−113.63**

表 3 所有模型在 6 月数据集的实验性能

模型		*K*50			*K*100		
		N=10	*N*=15	*N*=20	*N*=10	*N*=15	*N*=20
仅考虑文本的方法	BAT	−132.37	−311.32	−558.90	−143.13	−328.34	−581.57
	LCTM	−91.72	−208.75	−367.76	−81.88	−181.57	−323.16
整合社交上下文的方法	LeadLDA	−63.54	−150.18	−278.19	−72.07	−169.80	−309.40
	AdjEnc	−67.57	−159.66	−290.10	−72.02	−165.87	−303.37
	ForumLDA	−78.22	−140.46	−229.62	−82.33	−160.46	−258.72
	IATM	−46.69	−113.09	−213.61	−59.11	−133.96	−225.48
	PCFTM(-seq)	−27.94	**−64.83**	**−118.38**	−30.98	−70.74	−125.83
本文方法	**MVTM**	**−26.72**	−66.17	−122.31	**−29.40**	**−67.87**	**−122.07**

表 4 所有模型在 7 月数据集的实验性能

模型		*K*50			*K*100		
		N=10	*N*=15	*N*=20	*N*=10	*N*=15	*N*=20
仅考虑文本的方法	BAT	−65.49	−167.11	−322.43	−89.36	−194.96	−339.22
	LCTM	−72.78	−160.08	−275.58	−63.56	−137.36	−238.31
整合社交上下文的方法	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
	ForumLDA	−89.16	−215.47	−396.20	89.96	−213.59	−386.65
	IATM	−50.75	−119.48	−212.26	−46.80	−110.27	−204.35
	PCFTM(-seq)	−31.19	−74.28	−135.26	−35.29	−80.32	−144.38
本文方法	**MVTM**	**−27.68**	**−64.68**	**−118.39**	**−31.79**	**−72.89**	**−132.83**

表 5 IATM 模型在不同稀疏程度的数据集上的实验性能

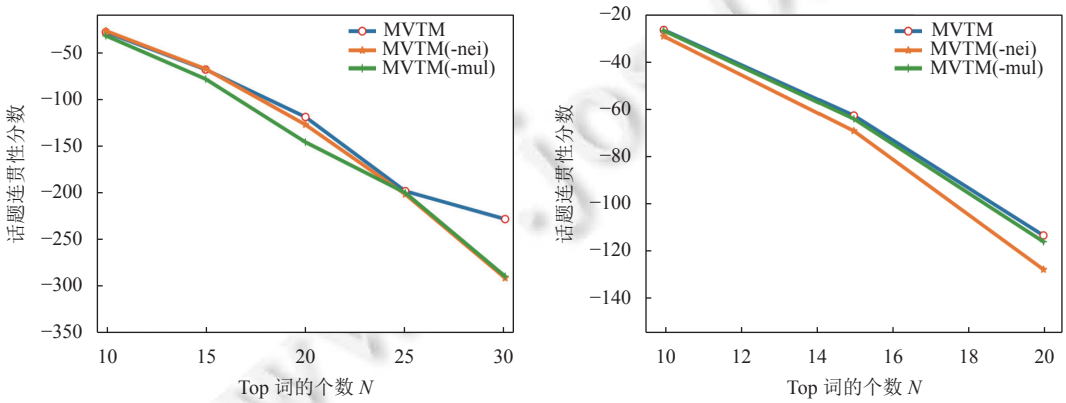
月份	*K*50			*K*100			交互密度
	N=10	*N*=15	*N*=20	*N*=10	*N*=15	*N*=20	
5月	−30.40	−70.04	−120.36	−26.94	−63.04	−113.63	1.45
6月	−26.72	−66.17	−122.31	−29.40	−67.87	−122.07	1.67
7月	−27.68	−64.68	−118.39	−31.79	−72.89	−132.83	1.49

● 模型 BAT 的主题连贯性得分最低. 尽管对抗训练在长文档数据中取得了很好的效果, 但对于短文本的数据输入, 它既没有采用聚合策略, 也没有建模社交上下文, 因此遭遇了严重的数据稀疏问题, 效果较差. 另一个长文档主题模型 AdjEnc 表现比 BAT 好. 这是因为 AdjEnc 将数据中存在的网络结构整合到模型中, 建模了社交网络中交互结构, 对内容信息起到了补充作用, 因此得到了更好的效果. 然而, AdjEnc 没有考虑社交媒体中用户的动态交互, 也只建模了同构的社交网络结构, 因此效果不如社交媒体主题模型. 总体来说, 长文档主题模型没有考虑帖子简短、表达不正式的特点, 效果并不能令人满意.

● 对于社交媒体主题模型, 我们可以看到 LCTM 的连贯性得分较低. 这可能是由于它只关注文本内容, 而忽略了静态网络结构和动态用户行为等丰富的社交上下文. 此外, LCTM 严重依赖于诸如 Wikipedia 之类的高质量外部数据来训练词嵌入表示. 由于 Wikipedia 文章与社交媒体数据在语言形式和词汇选择上有着较大的区别, 因此在此基础上训练得到的词的分布式表示可能会在主题推理中引入噪声和偏差. 进一步可以观察到, 通过整合社交上下文, LeadLDA、ForumLDA、IATM 和 MV-TM 都优于 LCTM.

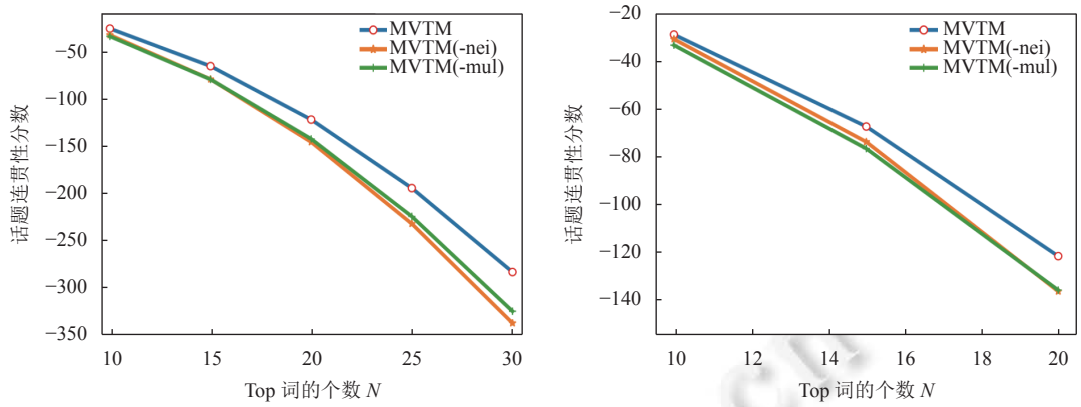
● 与基线模型相比, 我们的模型 MVTM 在所有 3 个评估数据集上都具有更好的连贯性得分. 实验结果首先表明, 整合社交上下文信息对主题检测来说是有帮助的. 此外, 在所有的基线模型中, IATM 和 PCFTM(-seq) 的效果最好. IATM 建模用户间的动态交互, PCFTM(-seq) 捕获网络结构与帖子文本之间的非线性关系. 然而它们同等对待用户间的转发和评论关系, 忽视了异构的社交上下文. 与 IATM 和 PCFTM(-seq) 相比, MVTM 考虑了社交网络的异构性. 它建模了用户的不同邻居和不同交互方式对主题检测的影响. 此外, MVTM 有效地捕获了不同视图间主题语义的潜在关联. 从实验结果上也可以看出, MVTM 取得了更好的连贯性得分. 这表明建模社交网络中的异构交互关系和不同的用户影响对主题检测是有帮助的.

● 注意到当 *K* 固定时, *N* 越小, 主题连贯性得分越高. 这可能是当 *N* 较大时, 由于用户表达的随意性, 主题模型混入了更多的常用词和噪声词. 常用词是指会在很多文档中出现的词, 包含较少的信息量. 噪声词是指与主题无关的词. 主题连贯性评价指标会对这两类词进行惩罚, 如公式 (21) 所示. 图 4–图 6 通过绘制 3 个数据集上连贯性得分相对于 *N* 的变化情况, 直观地展示了上述观察结果.



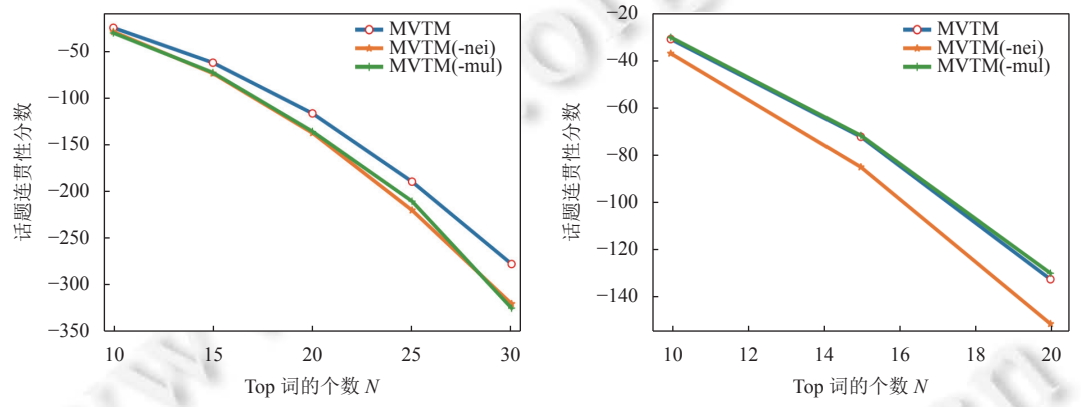
(a) *K*50 下的连贯性得分 (b) *K*100 下的连贯性得分

图 4 MVTM 与其两个变体在 5 月数据集上的表现



(a) K50 下的连贯性得分 (b) K100 下的连贯性得分

图 5 MVTM 与其两个变体在 6 月数据集上的表现



(a) K50 下的连贯性得分 (b) K100 下的连贯性得分

图 6 MVTM 与其两个变体在 7 月数据集上的表现

● 不同数据集上的实验结果表明, MVTM 模型在 6 月表现最好, 5 月、7 月次之. 考虑到不同数据集中可能包含不同密度的交互, 我们得出如下分析: 社交网络中交互密度越大, MVTM 的性能越好. 交互密度是指网络中交互关系数 (转发数+评论数) 与用户数的比值. 从表 1 中可以看到, 6 月份数据集中交互密度最大, 其他两个月较低. 拥有更多的交互密度, 模型可以捕获到更丰富的社交上下文, 从而得到更好的连贯性得分.

为了进一步说明 MVTM 模型对于面向社交媒体主题挖掘任务的有效性及其原因, 本文还进行了 MVTM 模型与其两个变体在 3 个数据集上的相关对比实验.

4.4.2 本文方法与其变体的比较

为了继续研究模型中的邻居级注意力与交互级注意力是否对微博主题检测有帮助, 我们将 MVTM 与其两个变体 MVTM(-nei) 和 MVTM(-mul) 进行比较. 这两个变体分别在 MVTM 中去除了邻居级注意力和交互级注意力, 最终根据主题连贯性的效果来判断注意力机制是否有作用. 在主题数 K=50 和 K=100 设置下, 图 4–图 6 分别展示了 3 个模型根据其得到的主题-词分布计算的主题连贯性得分. 通过观察图 4–图 6, 可以得出以下发现.

● 邻居级注意力的分析. 在大多数情况下, MVTM 的一致性得分高于 MVTM(-nei). 其原因可能是 MVTM(-nei) 忽略了用户邻居对主题推理的影响. 社交媒体中很多用户节点包含了重复的无效信息, 同等对待他们会在主题-词分布中引入常用词或噪声词. 而 MV-TM 通过邻居级的注意力机制有效地降低了这部分用户的权重.

● 交互级注意力的分析. 3 个评估数据集的主题一致性结果表明, MVTM 方法优于 MVTM(-mul). MVTM(-mul)

平等地对待用户之间的不同交互, 忽略了它们在主题语义方面的潜在相关性. 而 MVTM 引入了交互级注意力, 并设计了多视图神经变分推理技术, 有效地解决了这一问题. 交互级注意力捕捉交互关系之间的相互影响, 自适应地平衡不同主题语义之间的一致性和独立性. 尽管如图 6(b) 所示, 当 K 为 100 时, MVTM 的一致性得分略低于 MVTM(-mul), 但 MVTM 在大多数的情况下更有效.

● 以上两个因素的主题连贯性得分的平均增长百分比如表 6 所示, 可以看出, 这两种注意力都带来了主题连贯性的提高. 这表明考虑用户邻居和不同交互方式对主题检测是有意义的. 此外可以注意到, 由于交互级注意力的平均连贯性增长一般高于邻居级注意力, 原因可能是 AMHCN 中不同的用户交互 (如重新发布和评论) 对主题检测的影响更大, 而交互层面的注意力则有效地捕获了它们不同的重要性.

表 6 建模邻居级注意力和用户级注意力所带来的相关分数平均增长率 (%)

Factor	5月		6月		7月	
	*K*50	*K*100	*K*50	*K*100	*K*50	*K*100
邻居级注意力	0.7	9.8	17.8	8.5	14.4	14.9
交互级注意力	10.0	1.8	18.3	11.8	14.8	−1.7

综合以上分析, MVTM 可以生成更加一致的主题. 通过对比实验与消融实验, 我们将其原因归结为以下两点: (1) MVTM 整合了异构社交上下文, 区分对待不同的交互方式. 具体地, 通过建模不同邻居和交互方式的重要性来学习高质量的特定视图的嵌入表示. (2) 多视图神经变分推理通过尽可能地重构特定视图的嵌入表示, 捕捉到了不同视图之间的复杂关联. 通过自适应地平衡多视图的一致性和独立性, MVTM 最终生成更加连贯的主题.

4.5 案例分析

为了对模型检测的主题有直观上的感受与理解, 我们提取出模型挖掘出的关于 “互联网”主题的前 10 个词. 我们在基线模型中选择效果最好的 IATM 作为对比, 利用词云做出可视化, 效果分别如图 7、图 8 所示. 图中每个词代表检测出的相应的主题词, 字体大小代表出现概率, 字体越大表明在当前主题下出现概率越大, 即与当前的主题越相关. 从图 7 和图 8, 可以得到以下观察结果.

图 7 IATM 描述主题“互联网”的词云

图 8 MVTM 描述主题“互联网”的词云

● IATM 建模用户的动态交互, 学习社交网络中边的嵌入表示. IATM 检测出的主题词中概率最大的是“公司”, 这是与“互联网”主题相关的词. 接着是“投资”, 这是与互联网不相关的. 再往后的是“互联网”“产品”“谷歌”“经理”“基金”“谷歌”“百度”“注重”. 其中, “经理”“基金”和“注重”是与主题不相关的. 这是由于互联网经常与一些企业相关, 企业的相关帖子中经常出现“经理”“基金”等词, 导致 IATM 在建模用户交互时错误地将这些词混入.

● MVTM 区分对待不同的交互方式, 并且建模了不同用户的重要性. 它首先检测出的是“谷歌”, 与主题相关. 接着是“公司”“谷歌”“百度”“互联网”“苹果”“产品”“微软”“审查”“国内”. 其中只有“审查”“国内”与主题无关. 可以看到, MVTM 提取的词与主题更加相关, 且不相关词出现的位置更靠后. MVTM 通过邻居级注意力和交互级注意力建模异构社交上下文, 学习更好的用户表示. 同时, 多视图神经变分推理动态地平衡不同交互语义之间的关系,

推断出更连贯的主题.

4.6 参数调节

在 MVTM 中, 有两个超参数对模型的性能有重要影响. 一个是计算视图嵌入时用户嵌入与交互关系嵌入的平衡因子 α; 另一个是损失函数中表示学习的损失函数与多视图神经变分推理的损失函数之间的平衡因子 λ. 我们对这两个参数做进一步的实验分析, 在 5 月份数据集上对这两个参数采取不同的值, 分析得到的主题连贯性得分, 观察参数对模型的影响. 具体地, 两个超参数的取值设置在 [0, 1] 之间, 步长设置为 0.1, 根据主题-词分布按照概率从大到小排序, 取前 10 个词语 (N=10) 计算主题连贯性得分. 实验结果如图 9、图 10 所示.

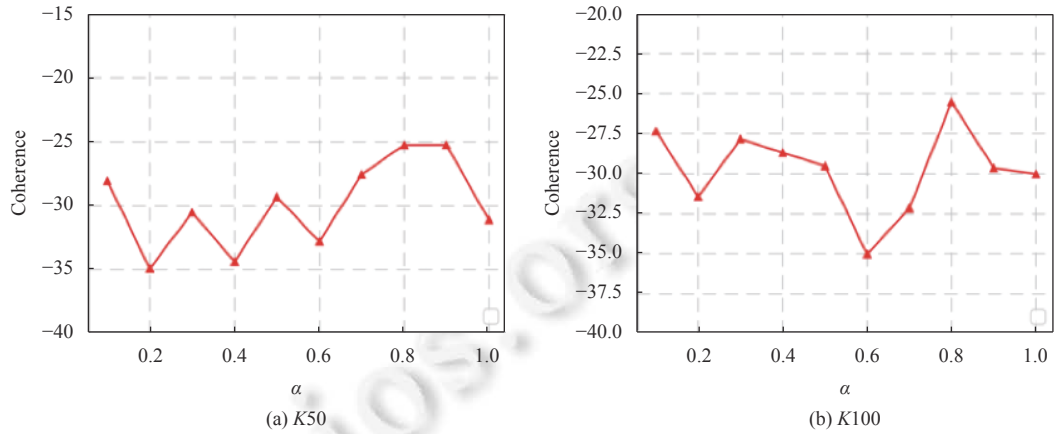


图 9 α 对实验性能的影响

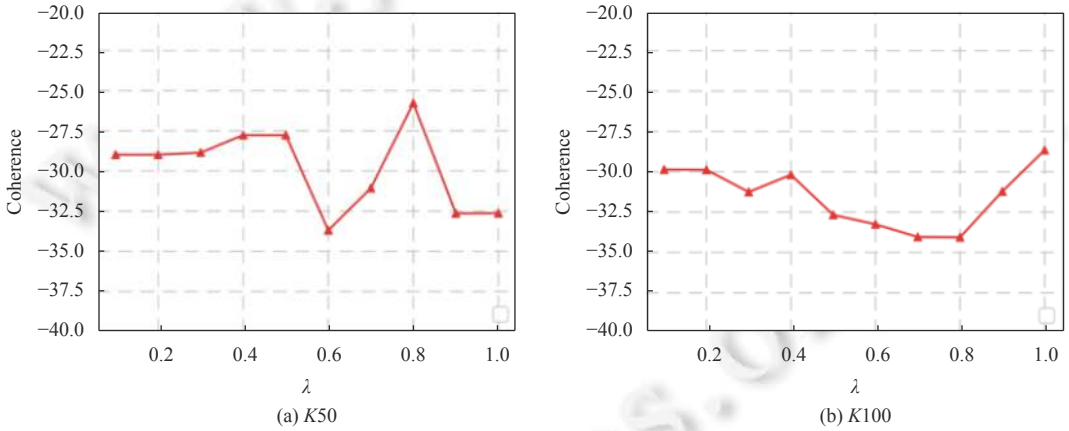


图 10 λ 对实验性能的影响

图 9 展示了 α 从 0.1 变化到 1.0 的主题连贯性得分变化情况. 从 K50 和 K100 两幅图中可以看到, 模型的性能先出现震荡, 然后 α 在取到 0.8 时连贯性达到最高值. 这表明用户嵌入表示与交互嵌入表示在这一比例下结合能更好地为主题检测提供社交上下文支持. 图 10 展示了 λ 从 0.1 变化到 1.0 的主题连贯性得分变化情况. 我们可以看到, 当 λ 取值较小时, 主题连贯性得分较低. 这是因为此时变分自编码器的损失函数在总体损失函数中占比较低, 使得主题推断模块不能很好地更新参数. 当 λ 较大时, 主题连贯性得分会达到最优值. K=50, 当 λ 取 0.8 时主题的连贯性最佳. K=100, 当 λ 取 1 时, 主题的连贯性最佳.

5 总结与展望

针对前人工作仅考虑帖子的文本内容, 或者对同构情境下的社交上下文进行建模, 使得不容易发现更加紧凑

的主题, 本文从社交网络的异构性出发, 提出编码异构社交上下文的多视图主题模型 (MVTM) 进行社交媒体短文本主题检测. 考虑社交网络中不同用户对主题的注意力以及用户间多种交互方式对主题检测的影响, 将社交网络按照交互关系划分为多个视图, 利用邻居级注意力和交互级注意力分别建模网络中不同邻居用户和不同交互方式间的相互作用, 学习特定视图的嵌入表示. 通过多视图驱动的神经变分推理捕获了不同视图级主题语义间的复杂关联, 通过自适应性地平衡多视图语义间的一致性和独立性, 我们的模型可以生成更连贯的主题. 在真实的 3 个月新浪微博数据集上的实验结果表明, 本文提出的 MVTM 对社交媒体领域短文本主题检测的有效性, 并且消融实验证明模型中的邻居级注意力和交互级注意力机制对主题连贯性的提高是有帮助的.

本文在用户序列上采用自注意力机制有效学习了不同用户对主题推断的不同重要性. 然而, 用户文本信息聚合自原始消息和转发消息等相关内容, 用户注意力机制只能建模用户间的不同权重, 无法学习每个用户的多个帖子的不同权重, 使得局部范围内可能存在较多垃圾信息; 同时, 社交媒体平台上存在许多无转发、评论等交互关系的安静用户, 他们的帖子同样是主题信息的重要来源. 然而由于缺乏在社交网络中的结构信息, 这些用户将面临更严重的数据稀疏问题. 对此, 我们下一步将探索: (1) 帖子级的注意力机制, 在用户文本信息聚合时降低主题无关消息或简单回应消息的权重, 进一步减小噪声帖子对主题推断的影响; (2) 如何深入整合社交网络中安静用户的信息来挖掘社交媒体上短文本的主题.

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] Wang Y, Li J, Chan HP, King I, Lyu MR, Shi SM. Topic-aware neural keyphrase generation for social media language. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 2516–2526. [doi: 10.18653/v1/P19-1240]

[3] 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: Association for Computational Linguistics, 2019. 608–618. [doi: 10.18653/v1/P19-1058]

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

[5] Hong LJ, Davison BD. Empirical study of topic modeling in Twitter. In: Proc. of the 1st Workshop on Social Media Analytics. Washington: ACM, 2010. 80–88. [doi: 10.1145/1964858.1964870]

[6] 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 Information Retrieval. Dublin: Springer, 2011. 338–349. [doi: 10.1007/978-3-642-20161-5_34]

[7] 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]

[8] Tang J, Zhang M, Mei QZ. One theme in all views: Modeling consensus topics in multiple contexts. In: Proc. of the 19th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. Illinois: ACM, 2013. 5–13. [doi: 10.1145/2487575.2487682]

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

[10] 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]

[11] Lu HY, Xie LY, Kang N, Wang CJ, Xie JY. Don't forget the quantifiable relationship between words: Using recurrent neural network for short text topic discovery. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco: AAAI Press, 2017. 1192–1198.

[12] Turian J, Ratinov L, Bengio Y. Word representations: A simple and general method for semi-supervised learning. In: Proc. of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: Association for Computational Linguistics, 2010. 384–394.

[13] Mikolov T, Yih WT, Zweig G. Linguistic regularities in continuous space word representations. In: Proc. of the 2013 Conf. of the North American Chapter of the Association for Computational Linguistics. Atlanta: Association for Computational Linguistics, 2013. 746–751.

[14] Sridhar VKR. Unsupervised topic modeling for short texts using distributed representations of words. In: Proc. of the 1st Workshop on

Vector Space Modeling for Natural Language Processing. Denver: Association for Computational Linguistics, 2015. 192–200. [doi: 10.3115/v1/W15-1526]

[15] Hu WH, Tsujii J. 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: Association for Computational Linguistics, 2016. 380–386. [doi: 10.18653/v1/P16-2062]

[16] Wang YM, Li XM, Zhou XT, Zhou XT, Ouyang JH. Extracting topics with simultaneous word co-occurrence and semantic correlation graphs: Neural topic modeling for short texts. In: Proc. of the 2021 Findings of the Association for Computational Linguistics (EMNLP 2021). Punta Cana: Association for Computational Linguistics, 2021. 18–27. [doi: 10.18653/v1/2021.findings-emnlp.2]

[17] Lyu B, Chen L, Zhu S, Yu K. LET: Linguistic knowledge enhanced graph transformer for Chinese short text matching. In: Proc. of the 35th AAAI Conf. on Artificial Intelligence. AAAI, 2021. 13498–13506. [doi: 10.1609/aaai.v35i15.17592]

[18] Bi B, Tian YY, Sismanis Y, Balmin A, Cho J. Scalable topic-specific influence analysis on microblogs. In: Proc. of the 7th ACM Int'l Conf. on Web Search and Data Mining. New York: ACM, 2014. 513–522. [doi: 10.1145/2556195.2556229]

[19] 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. Berlin: Association for Computational Linguistics, 2016. 2114–2123. [doi: 10.18653/v1/P16-1199]

[20] 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.

[21] 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]

[22] Gui L, Leng J, Pergola G, Zhou Y, Xu RF, He YL. Neural topic model with reinforcement learning. 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. 3478–3483. [doi: 10.18653/v1/D19-1350]

[23] Nan F, Ding R, Nallapati R, Xiang B. Topic modeling with wasserstein autoencoders. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 6345–6381. [doi: 10.18653/v1/P19-1640]

[24] Wang R, Hu XM, Zhou DY, He YL, Xiong YX, Ye CC, Xu HY. Neural topic modeling with bidirectional adversarial training. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 340–350. [doi: 10.18653/v1/2020.acl-main.32]

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

[26] Lim KW, Chen CY, Buntine WL. Twitter-network topic model: A full Bayesian treatment for social network and text modeling. In: Proc. of the 27th Annual Conf. on Neural Information Processing Systems: Topic Models Workshop. 2013. 1–5.

[27] Wang X, Ji HY, Shi C, Wang B, Ye YF, Cui P, Yu PS. Heterogeneous graph attention network. In: Proc. of the 2019 World Wide Web Conf. San Francisco: ACM, 2019. 2022–2032. [doi: 10.1145/3308558.3313562]

[28] Park C, Kim D, Han JW, Yu H. Unsupervised attributed multiplex network embedding. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 5371–5378. [doi: 10.1609/aaai.v34i04.5985]

[29] Cen YK, Zou X, Zhang JW, Yang HX, Zhou JR, Tang J. Representation learning for attributed multiplex heterogeneous network. In: Proc. of the 25th ACM SIGKDD Int'l Conf. on Knowledge Discovery & Data Mining. Anchorage: ACM, 2019. 1358–1368. [doi: 10.1145/3292500.3330964]

[30] Miao YS, Grefenstette E, Blunsom P. Discovering discrete latent topics with neural variational inference. In: Proc. of the 34th Int'l Conf. on Machine Learning. Sydney: JMLR.org, 2017. 2410–2419.

[31] Srivastava A, Sutton C. Autoencoding variational inference for topic models. In: Proc. of the 5th Int'l Conf. on Learning Representations. Toulon: ICLR, 2017.

[32] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: Proc. of the 1st Int'l Conf. on Learning Representations. Scottsdale: ICLR, 2013.

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

[34] 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.

[35] Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 2014, 15(1): 1929–1958.

[36] 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. Red Hook: Curran Associates Inc., 2009. 288–296.

[37] Mimno D, Wallach HM, Talley E, Leenders M, McCallum A. Optimizing semantic coherence in topic models. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computational Linguistics, 2011. 262–272.

[38] 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]

[39] 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]

[40] Ding Y, Wei H, Pan ZS, Liu X. Survey of network representation learning. Computer Science, 2020, 47(9): 52 –69 (in Chinese with English abstract). [doi: 10.11896/jsjkx.190300004]

附中文参考文献:

[39] 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]

[40] 丁钰, 魏浩, 潘志松, 刘鑫. 网络表示学习算法综述. 计算机科学, 2020, 47(9): 52–69. [doi: 10.11896/jsjkx.190300004]

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.

刘宏宇(1997－), 女, 硕士, 主要研究领域为社会媒体话题检测.

王浩成(1997－), 男, 硕士, 主要研究领域为社会媒体话题检测.

王博(1979－), 男, 博士, 副教授, 主要研究领域为自然语言处理, 个性化推荐, 心理计算.