基于时间演化图卷积网络的舆情热点内容预测

文　雅[1]，杨　频[1]，廖　珊[2]，代金鞘[1]，贾　鹏[1]
(1. 四川大学网络空间安全学院，成都 610211；2. 中国电子科技集团公司第三十研究所，成都 610041)

摘　要：有效预测舆情事件的热点内容有利于提高对舆论导向的把控能力和对公众诉求的预判能力. 然而，现有的舆情预测工作大多关注事件整体趋势指标或情感极性的演变预测，鲜有针对舆情事件热点内容的预测研究. 为解决以上问题，本文提出一种基于时间演化图卷积网络的舆情热点内容预测方法：以舆情事件的热点词作为预测对象，首先，通过演化图卷积网络学习各时间片词语的空间关联关系；然后，使用门控循环单元捕捉各时间片词语特征的时序变化；最后，通过全连接层进行输出，实现对舆情事件热点词的预测. 以微博上两个不同的舆情突发事件的相关文本作为数据集，与两种现有热点词预测方法开展对比实验. 实验结果表明，该方法在两个数据集上的精确率分别达到 51.21%和 50.98%，召回率分别达到 50.17%和 48.15%，F_1值分别达到 50.68%和 49.52%，均高于两种对比方法，能够更好地完成舆情事件中热点词的预测.

关键词：舆情预测；热点词预测；时间演化图卷积网络
中图分类号：TP391.1　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2023.033001

A temporal evolving graph convolutional network for Public opinion prediction in emergencies

WEN Ya[1], *YANG Pin*[1], *LIAO Shan*[2], *DAI Jin-Qiao*[1], *JIA Peng*[1]
(1. College of Cybersecurity, Sichuan University, Chengdu 610211, China;
2. The 30th Research Institute of China Electronics Technology Group Corporation, Chengdu 610041, China)

Abstract: Public opinion prediction is one of the key solutions to improve the ability to guide public opinion in emergencies. However, most of the existing public opinion prediction work focuses on the trend indicator or sentiment polarity of events ,while little attention paid to the prediction of hot words and topics in specific events. In this paper, a temporal evolving graph convolutional network for public opinion prediction in emergencies is proposed, in which the hot words associated with specific events are taken as the object of public opinion prediction. Our approach combines evolving graph convolutional network with gated recurrent unit: the former is used to learn the dynamic spatial correlation between words and the latter is used to capture the temporal changes of words, the hot words of an emergency in the next time period is then predicted through full connection layer output. To validate the proposed method, we selected discussion texts related to two emergencies on Weibo as the dataset, and conducted comparative experiments with two existing hot word prediction methods. The results show that our method achieved higher precision, recall, and F_1-score in both emergencies, with precision of 51.21%

收稿日期：2022-11-01
基金项目：四川省科技厅重点研发项目(2021YFG0156)
作者简介：文雅(1997—)，女，硕士研究生，主要研究领域为舆情分析与预测. E-mail：tanya_scu@163.com
通讯作者：杨频. E-mail：yangpin@scu.edu.cn

and 50.98%, recall of 50.17% and 48.15%, and F_1-scores of 50.68% and 49.52%, respectively. These results demonstrate that our proposed method is effective in predicting public opinion during emergencies

Keywords: Public opinion prediction; Hot words prediction; Temporal evolving graph convolutional network

1 引　言

舆情指舆论情况,是指在舆情因变事项(下文简称舆情事件)发生、发展和转变过程中,民众所持有的看法、观点和态度等[1]. 随着互联网和自媒体的发展,以前只会从事件发源地慢慢扩散流传的舆情事件,现在则很快通过网络散播并被全国各地人民知晓[2]. 网络舆情的传播速度快,传播规模大,参与门槛低[3].特别是当恶性事件发生后,如果不能及时了解民众诉求,尽快进行舆论引导,事件舆论可能会加速发酵升级,给群众带来恐慌,甚至影响民众对政府的信任度[2].因此,有效开展舆论引导工作具有重要意义[4].

及时发现和实时监控舆情可以对舆论引导工作起到帮助[4]. 目前,针对舆情的发现和监控已有广泛研究,其中 Nielsen、Goonie 和 PALAS 等舆情监控系统都可以帮助企业和政府对舆情进行发现和监控[3]. 但是这些系统主要用于发现未知舆情,并对已知舆情的舆论走向进行分析,对舆情舆论的未来发展预测较少. 面对复杂多变的网络舆情环境,为了更有效地开展舆论引导工作,需要防患于未然,加强对公众诉求的预判能力,预防突发的舆论危机,那么一个关键的解决途径是对舆情事件下一个阶段的发展进行有效预测.

如果能够在舆情事件发展过程中,对其未来走向进行有效预测,就能够更加准确地预判群众对事件的看法,及时调整舆论引导策略[4]. 目前国内外在舆情预测方面已有较多研究,然而,现有的相关研究工作,通常只对舆情事件整体的趋势指标或者情感极性进行预测. 这类预测对象在一定程度上确实能够反映大众对事件的态度,如大众对事件关注度的高低,对事件的态度是积极还是消极等. 但是这类预测对象难以捕捉舆情事件发展过程中群众关注点和具体诉求的变化,即舆情热点内容的变化. 及时了解舆情热点内容变化的意义主要表现在:有助于提高政府和企业对舆情事件中群众的关注点和具体诉求的预判能力,能够在苗头性倾向性问题上掌握主动权,在群众的不满和对立等负面情绪升级之前,进行适当的良性引导,摆脱被动滞后,使舆论引导更加主动和精准[4]. 因此,为了达到更好的舆论引导效果,可以通过预测舆情事件中热点内容随时间的发展变化及时获取群众的关注点和具体诉求的变化,进而为舆论引导策略的调整提供参考.

基于上述分析,本文提出了一种基于时间演化图卷积网络(Temporal Evolving Graph ConvolutionalNetwork,T-EGCN)的舆情热点内容预测方法. 以舆情事件每个时间片的热点词作为内容预测对象,通过预测一个舆情事件发展过程中热点词的变化来体现热点内容的变化. 具体来说,本文首先搜集社交媒体上针对某个舆情事件的讨论文本,通过主题模型筛选得到每个时间片的热点词,以此代表该舆情事件不同时间片群众的关注点和具体诉求,即舆情热点内容. 然后,根据每个时间片的热点词的热度权重为阈值选取每个时间片的候选热点词语,以关联关系为边转化为图结构,形成候选热点词空间关联关系图. 接着,将词语特征和图结构输入到演化图卷积网络(EvolveGCN)[5],通过候选热点词的动态空间关联关系,为下一时间片的热点词预测提供丰富的前序时间片词语关系变化信息. 然后,使用门控循环单元(GRU)学习带有空间信息的词语特征,实现对候选热点词时序关系的捕捉,使用词语的时序信息丰富预测特征. 最后,使用全连接层进行热点词预测输出. 实验表明,本文方法相比已知的两种热点词预测方法,能更准确地预测舆情事件下一时间片的热点词,完成舆情热点内容预测.

本文的主要贡献有:(1)提出了 T-EGCN 模型,在 EvolveGCN 的基础上融入 GRU,通过 EvolveGCN 捕捉空间特征变化,利用 GRU 学习时间特征变化,该模型能够同时捕获空间动态性和时间动态性,是图卷积网络(GCN)在时空预测任务上的一个扩展方案,可以应用于时间和空间信息均存在动态变化的时空预测任务;(2)提出了一种网络舆情热点内容预测方法,使用基于数据量的动态时间分片方法、词语相对热度计算方法和候选热点

词筛选方法对数据进行预处理，并使用 T-EGCN 模型进行预测，实现了一种利用舆情事件前序时间片段的词语信息，预测后续时间片段该舆情事件热点词的方法；(3) 通过实验验证了本文所提方法能够在已知前序多个时间片的事件数据的基础上，预测下一时间片事件的热点词，方法预测效果优于近年的两种预测未来热点词的方法，在本文的两个舆情事件数据集上，预测精确率分别达到 51.21%和 50.98%，召回率分别达到 50.17%和 48.15%，F_1 值分别达到 50.68%和 49.52%.

2　相关工作

目前，舆情预测的研究工作中，选取的研究对象主要是趋势指标[6]或情感极性[7]. 张虹等[8]以热点事件的网络论坛点击率和回复数为预测对象，提出了一种基于小波分析和神经网络建模的非线性事件序列的预测方法. 杜慧等[9]针对热度趋势指标缺乏统一衡量指标的问题，提出了一种基于因果模型的主题热度算法，以定量评估的主题热度作为预测对象，实现了一种基于多峰高斯曲线拟合热度变化进行主题热度预测的方法. 崔彦琛等[10]针对舆情预测研究中情感分析预测研究不足的问题，提出了一种构建事件专属情感词典对情感极性进行定量分析的方法，以定量评估的情感极性值为预测对象，实现了一种基于 ARIMA 模型的舆情事件情感分析预测方法. 程铁军等[11]以百度指数作为热度趋势指标，利用模态分解在非线性噪声序列数据处理方面的优势，提出了一种结合 BP 神经网络和模态分解对事件百度指数进行预测的方法，增强舆情预测模型的泛化能力和非线性预测能力. 这些针对趋势指标或情感极性的预测研究，能够在一定程度上反映舆情变化，但是并不能细粒度地反映舆情中群众关注点和具体诉求的变化，即舆情热点内容的变化.

而目前针对热点内容的预测研究，通常只利用前序一个时间片的词语空间关联关系，如语义关系或者共现性关系，对下一时间片的热点词进行预测，未考虑前序多个时间片的词语空间关联关系对热点词预测的影响. 岳丽欣等[12]提出一种基于 word2vec 语义关系的热点词预测方法，将与当前热点主题词的 word2vec 词语相似度最高的词语作为预测的未来热点词，实现了对美国干细胞研究领域热门研究方向的未来热点词预测. Li 等[13]提出了一种基于词共现概率的关键词信息熵算法，将上一时间片中信息熵高的词组预测为下一个时间片的热点词，最后通过新冠肺炎事件作为例子，说明了该算法在预测流行病事件话题的未来热点词上的可行性.

由于舆情热点内容会随着舆情事件发展逐渐变化，所以每个时间片中词语出现的频率是会不断变化的，词语空间关联关系也会随之改变. 因此，如果要通过前序多个时间片的词语空间关联关系来预测热点词，就需要同时捕捉时间和空间特征.

Zhao 等[14]提出了一种基于时间图卷积网络(Temporal Graph Convolutional Network，T-GCN)的时空预测模型，使用 GCN 学习节点的空间关联关系特征，使用 GRU 学习节点的时间特征，成功将节点的时间依赖性和空间依赖性有机结合在一起. Pareja 等[5]针对 GCN 难以挖掘图的动态演化特征的问题，提出了演化 GCN 结构的 EvolveGCN 模型，使用 RNN 演化图节点在图空间上的时序变化，能够为不同的时间节点输入不同的节点空间关联关系图结构，并获取随着关系图动态变化的节点嵌入.

3　基于 T-EGCN 的舆情热点内容预测方法

本文提出的基于 T-EGCN 的舆情热点内容预测方法的框架如图 1 所示，方法架构包括数据预处理、候选热点词提取、词语关系图构建和 T-EGCN 模型预测等 4 个部分. 首先，通过关键词搜索从社交媒体上爬取某一舆情事件在演化生命周期内的全部原创发表文本，并对其进行过滤清洗和分片. 接着，对每个时间片的词语，通过转赞评计算得到其内容影响力，结合词频-逆文档频率(TF-IDF)进行热度量化，得到各个时间片的词语相对热度排序，并通过主题模型动态筛选候选主题词形成候选热点词典. 然后，对每个时间片，分析候选热点词之间的语义相似度和共现性关系，进而结合内容影响力为每个时间片构造出一个候选热点词空间关联关系图. 最后，使用 EvolveGCN 和 GRU 分析词语关系共同进行热点词预测输出.

3.1　数据预处理

社交媒体上能够发表的信息载体有文本、图片和表情等[15]. 本文主要通过社交媒体上舆情事件的文本数据分析舆情热点. 考虑到本文研究的重点为热点词，所以对数据进行去重后，使用正则表

达式过滤掉了 tag、表情符号、链接和评论的回复前缀. 由于文本中的名词、动词等是具有实际意义、能够更好地反应舆论热点的词语[16],所以使用 jieba 库(https://pypi.org/project/jieba/) 进行分词,并做词性分析,保留名词、动词、形容词等有效词语,并去除过滤后为空的文本.

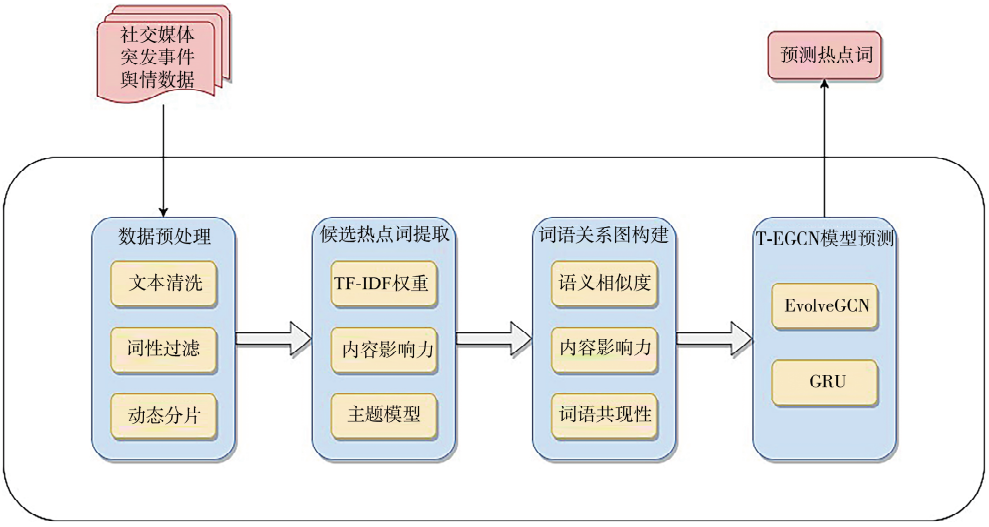


图 1　方法框架

Fig. 1　The framework of the method

在完成数据的清洗和过滤工作后,要对其进行分片. 目前的舆情预测研究使用的时间分片方法通常是均等时长分片,即在舆情事件的整个生命周期内,按均等时间长度切分数据进行分片分析[3, 9-12],如每 M 分钟,每 H 小时,每 D 天等,该方法的优点是划分方式简单,适用于舆情的事后分析工作.

然而,要在舆情事件发展过程中不断进行热点内容预测,显然需要实时性. 但由于舆情事件的文本量在不同时刻变化较大[17],例如当事件出现新发展时,讨论量会激增;根据作息规律,人们在凌晨的发帖量总是较少;爆发期博文数平均高于产生期和衰退期等. 这些情况导致均等时长分片的方法存在实时分析时不同时间片的数据量差异较大,而且无法及时感知舆情事件的突发新变化的问题,难以满足本文舆情预测的实时分析需求.

因此,本文提出了一种基于数据量的动态时间分片方法来实时和均衡地划分数据,其流程如图 2 所示. 首先以小时为单位捕获数据,设第 t 个小时获取的数据量为num_t,此时正在划分第 k 个时间片的数据,其已获取的数据量为$slice_k$,每个时间片的最小数据量阈值为 MIN,当$slice_k$达到 MIN 时,将$slice_k$划分为一个时间片的数据. 本文参考事件舆情产生期的第一个数据量激增周期数据设置阈值 MIN.

3.2　候选热点词提取

热点词是通过算法挑选出的能代表每个时间片中网民观点的词语[18,19]. 目前的研究当中,通常使用的选方法有两种,分别是基于主题模型的提取方法[12, 17, 20] 和基于词频等权重排序的选取方法[13, 18, 21,22]. 由于目前没有统一的挑选方法,本文同时考虑主题模型和权重排序,更全面地提取候选热点词.

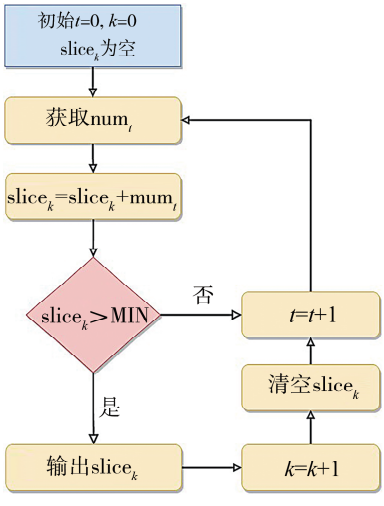


图 2　分片流程图

Fig. 2　The flow chart of partition

本文结合内容影响力和 TF-IDF,获取词语的权重排序. 其中,内容影响力代表一条文本对群众

的影响. 一般来说，文本的转赞评在很大程度上体现了文本的影响力[18]. 不同影响力的文本对包含在其中的词语的权重贡献程度应该不同，然而传统的 TF-IDF 忽略了这一点[21].

本文在 TF-IDF 的基础上加入内容影响力作为 TF-IDF 的权重. 具体过程如下：对第 t 个时间片的文本 d，其内容影响力 $P(d)$ 的计算公式为

$$P(d)=\frac{RP_d+RT_d+L_d+1}{RP_{D_t}+RT_{D_t}+L_{D_t}+N_{D_t}} \tag{1}$$

其中，RP_{D_t}、RT_{D_t} 和 L_{D_t} 分别表示第 t 个时间片中的所有文本的回复总数、转发总数和点赞总数；N_{D_t} 表示第 t 个时间片中的文本总数；RP_d、RT_d 和 L_d 表示文本 d 的回复数、转发数和点赞数. 该公式可以使转赞评越多的文本，其 $P(d)$ 越大.

然后，将 $P(d)$ 加入到 TF-IDF 当中. 对文本 d 中出现的词 j，其权重表示为 $w_{j,t}$，$w_{j,t}$ 的计算公式为

$$w_{j,t}=\sum_{d=1}^{D_t}P(d)\cdot tf\text{-}idf_{d,j} \tag{2}$$

其中，D_t 表示第 t 个时间片中的所有文本；$P(d)$ 表示文本 d 的内容影响力；$tf\text{-}idf_{d,j}$ 表示文本 d 中词 j 的 TF-IDF 值.

通过式(2)得到第 t 个时间片中所有词语的权重后，进行降序排列，便得到词语的权重排序. 接下来即可设置阈值，选取排序靠前的词语作为候选热点词.

本文选取在短文本上表现较佳的 GSDMM 主题模型[23,24]构建每个时间片的热点词集并设置候选热点词选取阈值，具体方法如下：对第 t 个时间片的数据，通过 GSDMM 主题模型抽取每个主题的前 N 个主题词加入热点词集，根据式(2)计算得到的词语权重排序，取热点词集中权重最低的词的权重为阈值，选出该时间片所有权重高于该阈值的词形成候选热点词典.

3.3　词语关系图构建

对第 t 个时间片的所有候选热点词，根据词语的语义和共现性，将它们关联起来，构建候选热点词空间关联关系图，其结构参考图 3.

图 3 中节点 $W_{i,t}$ 和 $W_{j,t}$ 表示第 t 个时间片的第 i 和 j 个候选热点词，$S_i=\{s_a,s_b,s_c,s_d\}$ 代表第 t 个时间片的文本中出现词 $W_{i,t}$ 的文本集合，$S_j=\{s_a,s_b,s_e,s_k\}$ 代表第 t 个时间片的文本中出现词 $W_{j,t}$ 的文本集合，文本和词是多对多的关系. 本文提出一种结合内容影响力的候选热点词相关性计算方法，来确定图中词与词之间的边的权值，即相关性 $r_{i,j,t}$. 如图 3 所示，词 $W_{i,t}$ 和词 $W_{j,t}$ 的公共文本集合为 $s_{\text{same}}=\{s_a,s_b\}$，其内容影响力权重为 $h_{s_{\text{same}}}$，不同文本集合为 $s_{\text{diff}}=\{s_c,s_d\}\cup\{s_e,s_k\}$，其均值 word2vec 文本相似度[25]为 $sim_{s_{diff}}$，内容影响力权重为 $h_{s_{\text{diff}}}$，则词 $W_{i,t}$ 和词 $W_{j,t}$ 的相关性 $r_{i,j,t}$ 的计算公式为

$$r_{i,j,t}=h_{s_{\text{same}}}+h_{s_{\text{diff}}}\cdot sim_{s_{\text{diff}}} \tag{3}$$

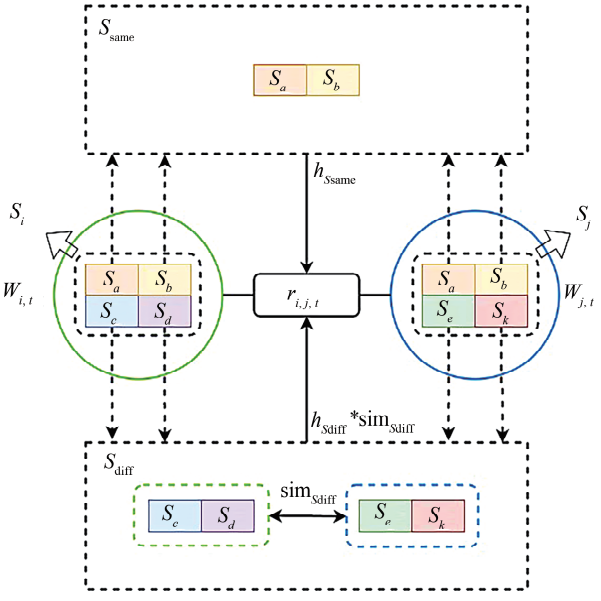


图 3　词语空间关联关系图

Fig. 3　Word spatial relevance diagram

内容影响力权重 h_s 的计算公式为

$$h_s=\frac{RP_s+RT_s+L_s+N_s}{RP_S+RT_S+L_S+N_S} \tag{4}$$

其中 S 表示 S_i 和 S_j 的并集；s 代表 s_{same} 或者 s_{diff}，N_S 表示 S 包含的文本总数；RP_S、RT_S 和 L_S 表示 S 包含的文本的回复数、转发数和点赞数；N_s 表示 s_1 或者 s_2 包含的文本总数，RP_s、RT_s 和 L_s 表示 s_{same} 或者 s_{diff} 包含的文本的回复数、转发数和点赞数. 转赞评越多的文本，其内容影响力权重越大，词语之间的相互影响程度越高.

以式(3)计算得到的相关性特征值作为图结构中节点之间的边的权值，可以表现出第 t 个时间片的候选热点词之间的语义相似性和共现性关系，即空间关联关系.

3.4　T-EGCN 模型预测

T-EGCN 模型利用 EvolveGCN 学习随着时间片变化的舆情事件候选热点词关联关系图的拓扑结构，获得空间维度特征. 再将带有空间信息的特征输入 GRU，提取时间维度特征，最后通过全连

接层进行预测输出,模型结构如图 4 所示.

图 4 的左侧部分展示了 T-EGCN 模型的整体结构,它由前后依次连接的 T-EGCN 单元组成,每个 T-EGCN 单元内部包含一个 EvolveGCN 层以及一个 GRU 层. 模型的输入为节点特征矩阵O_t和通过候选热点词空间关联关系图得到的带权重的邻接矩阵A_t,EvolveGCN 层通过图卷积的过程来学习空间依赖性,GRU 层通过重置门R_t和更新门Z_t来学习时间依赖性.

图 4 的右侧部分展示了 T-EGCN 模型的内部构成. 在 EvolveGCN 层中,使用基于频域的 GCN 来聚合候选热点词节点的邻居信息,加入权重参数矩阵W_t,用于记忆和传递图空间的时序变化,共同计算节点的嵌入矩阵X_t,具体计算过程如下.

$$\widetilde{A}_t = A_t + I \tag{5}$$

$$\widetilde{D}_t = \mathrm{diag}\left(\sum_j \widetilde{A}_{tij}\right) \tag{6}$$

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

$$X_t = \sigma(\hat{A}_t O_t W_t) \tag{8}$$

其中,A_t为带权重的邻接矩阵;I 为单位矩阵;diag 表示加入自环图的度矩阵计算函数,得到对角线为对应节点度加 1,其余数值为 0 的度矩阵$\widetilde{D}_t$;O_t是节点特征矩阵;W_t为 EvolveGCN 权重参数矩阵;σ 代表 ReLU 激活函数;卷积得到的节点嵌入矩阵X_t为 GRU 层的输入.

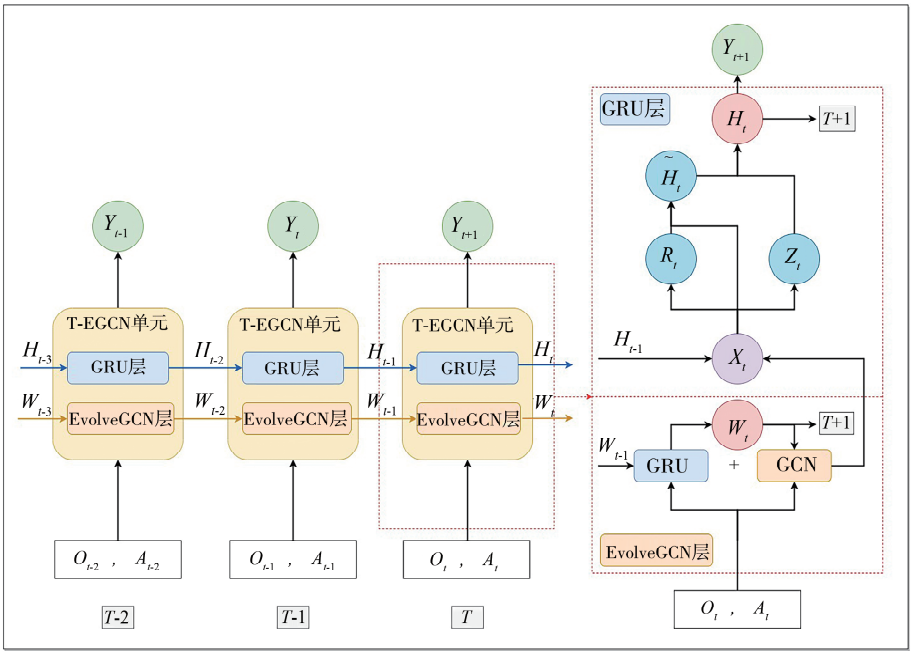


图 4　T-EGCN 模型结构

Fig. 4　Architecture of T-EGCN model

如图 4 所示,EvolveGCN 层中也包含能够传递图结构时序信息的 GRU 单元. 但是与 GRU 层相比,两者的输入I_t不相同:在 GRU 层中,输入I_t是第t 个时间片中 EvolveGCN 层卷积得到的节点嵌入矩阵X_t,学习的是词语节点特征的时序变化;而在 EvolveGCN 层的 GRU 单元中,输入I_t是第t 个时间片的候选热点词节点特征矩阵O_t,学习的是词语图空间的时序变化. 在第t 个时间片中,GRU 层和 EvolveGCN 层的 GRU 单元的重置门R_t、更新门Z_t、候选隐藏状态$\widetilde{C}_t$和最终隐藏状态C_t的计算公式为

$$R_t = \sigma(U_{rx} I_t + U_{rh} C_{t-1} + B_r) \tag{9}$$

$$Z_t = \sigma(U_{zx} I_t + U_{zh} C_{t-1} + B_z) \tag{10}$$

$$\widetilde{C}_t = \tanh(U_{cx} I_t + U_{ch}(R_t * C_{t-1}) + B_h) \tag{11}$$

$$C_t = Z_t * C_{t-1} + (1 - Z_t) * \widetilde{C}_t \tag{12}$$

其中,U 为可学习的权重参数;B 为可学习的偏差参数;I_t为当前时间片t 的输入;C_{t-1}为上一时间片的隐藏状态;σ 代表 sigmoid 激活函数;tanh 代表 tanh 激活函数;$*$ 代表哈达玛积.

因此,EvolveGCN 层的权重参数W_t和 GRU 层的隐藏状态H_t的参数传递和计算过程如下.

$$W_t = GRU(I_t = O_t, C_{t-1} = W_{t-1}) \tag{13}$$

$$H_t = GRU(I_t = X_t, C_{t-1} = H_{t-1}) \tag{14}$$

可以看到,两个 GRU 网络结构的输入输出存

在差异，本文利用 GRU 的特性来捕捉和学习不同维度的时序变化. 最后，将 GRU 层的H_t作为最终节点特征表示，输入到全连接层当中，得到下一时间片的热点词预测Y_{t+1}.

4　实验与分析

本节将基于微博舆情事件数据集，展开验证实验. 首先，详细介绍所用的数据集、评价指标和相关实验设置. 然后，基于对比实验，深入讨论分析本文方法的优点和不足. 其中，为了验证本文方法在时间分片上的数据均衡性和实时性，基于事件数据集，开展本文分片方法与常规分片方法的对比实验；为了验证本文方法在未来热点词预测上的有效性，基于事件数据集，开展本文方法与该领域新工作的对比实验；为了以更直观的方式展现本文方法在预测热点词方面的效果，基于事件数据集，绘制热点词云进行对比分析.

4.1　数据集

(1) 数据集 *A*. 爬取自新浪微博上关于 2018 年“女孩乘滴滴遇害”事件发布时间介于 2018 年 8 月 25 日 9 时至 8 月 31 日 24 时生命周期内[18]的中文原创微博文本和热门微博评论文本，共 38 668 条. 通过过滤清洗，得到结果不为空且不重复的文本 29 521 条. 按照本文基于数据量的动态时间分片方法，设最小数据量阈值 $MIN=500$，将数据划分为 43 个时间片，由于数据向前递补，所以最后一个时间片数据仅有 431 条，不足 500 条，为避免数据量不均衡问题，去掉最后一个时间片的数据，最终数据的时间介于 2018 年 8 月 25 日 9 时至 8 月 30 日 19 时，一共有 29 090 条，分为 42 个时间片.

(2) 数据集 *B*. 爬取自新浪微博上关于 2021 年“三只松鼠模特妆容争议”事件发布时间介于 2021 年 12 月 25 日 13 时至 12 月 29 日 24 时生命周期内的中文原创微博文本和热门微博评论文本，一共有 22 769 条. 通过过滤清洗，得到结果不为空且不重复的文本 17 900 条. 按照本文基于数据量的动态时间分片方法，设最小数据量阈值 $MIN=200$，将数据划分为 51 个时间片，由于数据向前递补，所以最后一个时间片数据仅有 191 条，不足 200 条，为避免数据量不均衡问题，去掉最后一个时间片的数据，最终数据的时间介于 2021 年 12 月 25 日 13 时至 12 月 29 日 21 时，一共有 17 709 条，分为 50 个时间片.

4.2　评估指标

由于方法的最终目标是预测未来时间片的热点词，就实验结果而言，对词语的判定为：若在第 t 个时间片的预测热点词集中则“是第 t 个时间片的热点词”，不在词集中则“不是第 t 个时间片的热点词”. 所以可以看作分类预测，采用分类预测算法常用的评价指标，即精确率(*Precision*)、召回率(*Recall*)和两者的调和平均(F_1-*score*)对算法进行评估，计算公式如下.

$$Precision=\frac{TP}{TP+FP} \tag{15}$$

$$Recall=\frac{TP}{TP+FN} \tag{16}$$

$$F_1\text{-}score=2\cdot\frac{Precision\cdot Recall}{Precision+Recall} \tag{17}$$

其中，*TP* 代表真阳性，即预测结果和实际都为热点词的词语数量；*FP* 代表假阳性，即预测结果为热点词，但实际不是热点词的词语数量；*TN* 代表真阴性，即预测结果和实际都为非热点词的词语数量；*FN* 代表假阴性，即预测结果为非热点词，实际却是热点词的词语数量.

4.3　实验设置

本文实验设置如下：

(1) 将数据集按时序关系切割成 7∶3 的两部分，T-EGCN 模型每次利用前序三个时间片的图信息进行学习，预测下一个时间片的热点词. 因此，最开始的三个时间片不能作为预测输出，数据集 *A* 构成 26 个训练集和 12 个测试集，数据集 *B* 构成 31 个训练集和 15 个测试集.

(2) T-EGCN 模型输入的节点特征值和输出的节点预测值为词语在对应时间片的相对权重，优化器为 Adam，学习率为 0.005，EvolveGCN 层和 GRU 层的隐藏层单元数目均为 200.

(3) 将整个数据集中出现的词语进行汇总，在第 t 个时间片中，对词汇表的每个词，根据真实热点词集，出现在热点词集中的词语标记为 1，其余词语标记为 0，作为样本的真实标签. 以各类算法提取出的预测热点词集为预测结果，对词汇表的每个词，出现在预测热点词集中的词语标记为 1，其余词语标记为 0，作为样本的预测标签.

(4) 本实验所用 GPU 服务器的显卡型号为 NVIDIA GeForce RTX 3090，显存为 24 G，编程语言为 python，深度学习框架为 pytorch.

4.4　实验结果与分析

本小节在 GSDMM 主题模型提供的对舆情事

件的主题提取分析结果的基础上,开展实验.

4.5.1　时间分片对比实验　为了验证本文所提出的基于数据量的动态时间分片方法比均等时长的时间分片方法更具有数据均衡性和实时性,对清洗后的数据数量进行分片统计. 对照组为每小时分片和每 X 小时分片,不同数据集选择的 X 数值不同的原因是:在对应数据集上选取的 X 时间长度和本文方法在对应数据集上划分的总数据片数最相近,在数据集 A 上 X 选择 4,划分结果为 39 个时间片,在数据集 B 上 X 选择 2,划分结果为 52 个时间片. 最终在两个数据集上得到微博条数随不同时间分片方法变化的规律如图 5 和图 6 所示.

从数据均衡性上来看,如图 5 和图 6 所示,当按每小时和每 X 小时切分文本时,不同时刻的数据量波动较大,特别是按每小时分片,某些凌晨时刻的博文数接近 0,导致时间序列分析时会出现信息断层问题,而本文方法每个分片的数据量分布明显更加均衡,分片数量也更合理.

从实时性上来看,如图 5 和图 6 所示,每 X 小时分片的数据量峰值相较本文基于数据量的动态时间分片,有一定的滞后性. 例如图 5 中数据集 A 的第 32 个小时,事件数据量激增,讨论较为激烈,但是每 4 个小时分片的方法会在第 36 个小时后,才将 32~36 这 4 个小时的数据划分为一个时间片进行分析,而本文基于数据量的动态时间分片方法能够感知数据量的激增,在第 32~36 小时期间几乎每个小时就会汇总划分一个时间片,能够及时捕捉由于事件出现新进展而出现的讨论峰值.

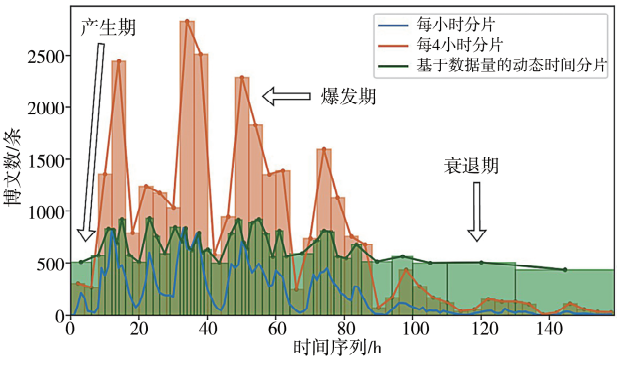


图 5　数据集 A 中微博条数随不同分片方法变化的统计图

Fig. 5　Statistics of the number of blog posts changing with different partition methods in dataset A

由上述分析可知,均等时长的时间分片方法存在信息量不均等,捕捉数据激增点的滞后性明显的缺点. 而本文提出的基于数据量的动态时间分片方法能够使每个时间片数据量相当,信息量均衡,并且能够及时获取事件的数据激增点,实时性较好.

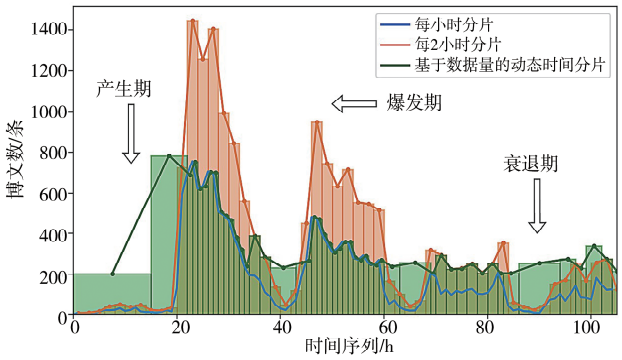


图 6　数据集 B 中微博条数随不同分片方法变化的统计图

Fig. 6　Statistics of the number of blog posts changing with different partition methods in dataset B

4.5.2　舆情热点词预测对比实验　针对上述两个舆情事件数据集,将本文提出的基于 T-EGCN 的舆情热点内容预测方法和两个近年的未来热点词预测方法以及 EvolveGCN 模型在网络舆情热点内容预测上的性能进行对比,相关对比方法描述如下.

(1) 基于 word2vec 的方法[12]:计算当前时间片词语与热点词典词语的 word2vec 词向量间距,筛选与每个主题词语义距离最近的前三个词汇作为下一时间片的预测热点词.

(2) 基于信息熵的方法[13]:通过关键词关联规则挖掘出共现频率较高的关键词组合,引入信息熵公式计算关键词组合的信息熵,选取信息熵较高的关键词组合,将其作为下一时间片的预测热点词.

(3) EvolveGCN[5]:本文提出的 T-EGCN 模型的 EvolveGCN 层,EvolveGCN 能够学习随着时序图关系变化的节点嵌入,将卷积得到的节点特征输入到全连接层中,获得下一时间片的预测热点词.

实验结果如表 1 所示. 由表 1 可知,在同一舆情数据集下,本文方法的预测精确率、召回率和 F_1 值均为最高. EvolveGCN 模型仅考虑空间拓扑结构,预测效果略低于本文加入 GRU 后同时考虑时间和空间特征的模型结构. 说明相较其他方法,本文方法能够更好完成舆情事件中热点词的预测.

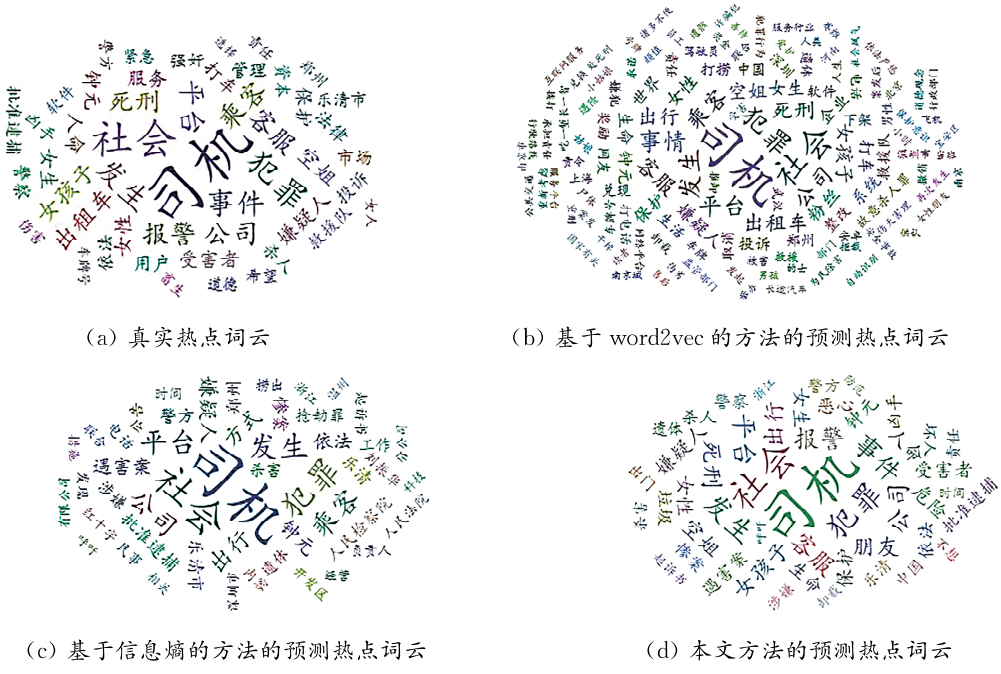
为了更加直观地分析不同方法的预测效果,绘制数据集 A 在第 30 和 40 个时间片的词云进行展示,如图 7 和图 8 所示. 第 30 个时间片事件处于爆发期,凶手刚被逮捕,此时群众愤怒情绪高涨,舆

论主要是希望判处凶手死刑、追责滴滴卸载滴滴软件、希望社会保护女性安全和怀疑警方执法不力；第 40 个时间片件处于衰退期，因为事件相关人员都被惩处，事件有了交代，群众的负面情绪得到了平复，和事件相关的微博在慢慢减少，更多的关注重点转移到了对如何避免此类事件再发生的建议上.

表 1　方法预测性能对比

Tab. 1　Comparison of methods prediction performance

方法\数据集	数据集 A			数据集 B		
	Precision/%	Recall/%	F1-score/%	Precision/%	Recall/%	F_1-score/%
基于 word2vec 的方法[12]	17.86	47.10	25.90	15.32	44.07	22.73
基于信息熵的方法[13]	21.34	21.08	21.20	18.60	19.63	19.10
EvolveGCN[5]	40.79	39.96	40.37	41.57	39.26	40.38
基于 T-EGCN 的方法(本文方法)	51.21	50.17	50.68	50.98	48.15	49.52



(a) 真实热点词云　　(b) 基于 word2vec 的方法的预测热点词云

(c) 基于信息熵的方法的预测热点词云　　(d) 本文方法的预测热点词云

图 7　第 30 个时间片的真实热点词云与预测热点词云对比图

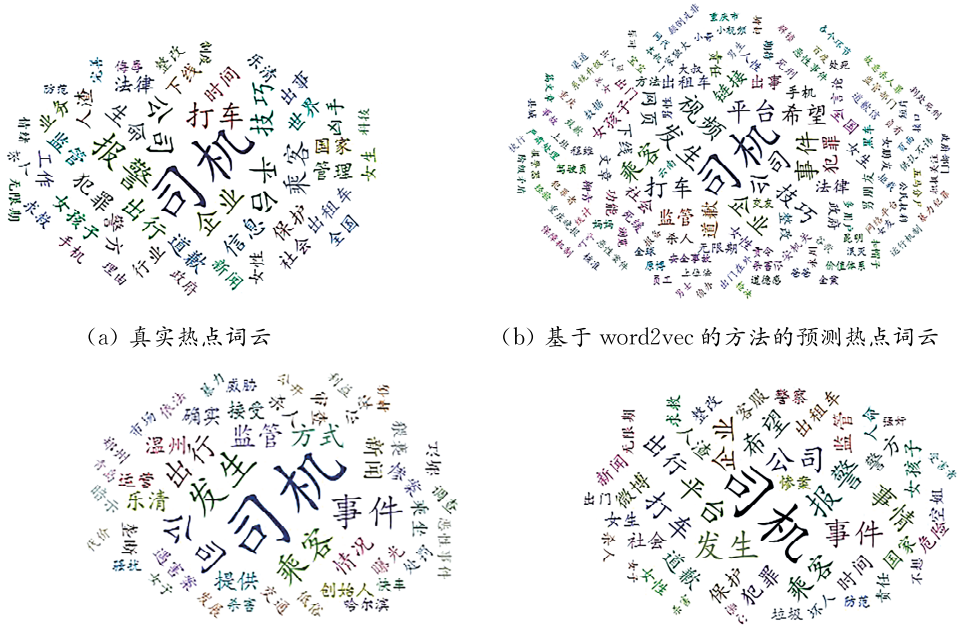
Fig. 7　Comparison of real and predicted hot word cloud in the 30th time period

从表 1 可以看出，基于 word2vec 的方法的精确率偏低，但是召回率比较高. 这是因为该方法会为第 t 个时间片的每个热点词寻找第 $t+1$ 个时间片的三个预测热点词进行对应，预测中存在很大冗余，所以精确度较低而召回率较高. 这一点也可以从图 7b 和图 8b 中可以看出，基于 word2vec 的方法能够找到较多真实的热点词，但是其预测的热点词数量远远多于真实的热点词. 造成该情况的主要原因可能是该方法主要研究科技文献领域的热点词预测，其预测工作通常以年为单位，热点词会有长期逐渐替代的过程，通过领域专家进行人工筛选，可以较为准确地去除冗余词. 而舆情事件变化时间间隔短，随着事态发展，热点内容随时间的波动较大，难以得到领域专家的先验知识对预测得到的热点词进行筛选，所以该方法在舆情热点词预测上的性能不佳.

同时，从表 1 可以看出，基于信息熵的方法的精确率和召回率都较低. 通过分析图 7c 和图 8c 的热点词分布，可以推测出基于信息熵的方法在预测时通常能抓住最可能延续热度的词语，即第 t 个时间片信息熵最高的 1～3 个词组实际上确实非常可能是第 $t+1$ 个时间片的热点词，例如长期被提及的“司机”等词，以及第 30 个时间片前刚发生的罪犯被逮捕一事. 但是当需要预测的范围变大时，该方法的性能急剧下降. 造成该情况的主要原因可能是该方法主要研究流行病事件的热点词预测，

流行病事件的舆情周期较长,相比本文的需求,其分片的时间跨度较大,如 10 d 为一个时间片,所以该方法并未考虑多个前序时间片信息对热点词的影响,仅考虑了前一个时间片词语之间的关联关系,导致其在预测短期舆情事件时,对次级重要的词语的预判能力不足.

而本文方法中,T-EGCN 模型会对每个候选热点词进行单独判断,所以相比基于 word2vec 的方法,冗余词较少;同时本文方法使用 GRU 考虑前序多个时间片的词语信息,在预测时信息量更加丰富,所以相比基于信息熵的方法,对次级重要的词语的预判能力更强.



(a) 真实热点词云　　(b) 基于 word2vec 的方法的预测热点词云

(c) 基于信息熵的方法的预测热点词云　　(d) 本文方法的预测热点词云

图 8　第 40 个时间片的真实热点词云与预测热点词云对比图

Fig. 8　Comparison of real and predicted hot word cloud in the 40th time period

5　结　论

本文提出一种基于 T-EGCN 的舆情热点内容预测方法. 根据社交媒体上针对特定突发舆情事件的讨论文本,获得每个时间片中事件的热点词,通过热点词的变化反映大众对该事件的关注重心的变化. 该方法将热点词作为预测的对象,利用候选热点词之间的语义相似性和共现性关系,为每个时间段都构建一个对应的候选热点词相关关系图,再使用 EvolveGCN 与 GRU 进行时间维度和空间维度上的联合分析,预测下一时间片的热点词. 实验结果表明本方法能够对网络舆情事件的热点词进行有效预测,在舆情时间数据集上,模型预测精度高于近年的热点词预测方法,能够实现在特定舆情事件发展过程中对具体热点内容进行预判.

参考文献:

[1]　高承实, 陈越, 荣星, 等. 网络舆情几个基本问题的探讨[J]. 情报杂志, 2011, 30: 52.

[2]　杨志, 祁凯. 基于“情景-应对”的突发网络舆论事件演化博弈分析[J]. 情报科学, 2018, 36: 30.

[3]　彭思琪, 周安民, 廖珊, 等. 基于图注意力网络的舆情演变预测研究[J]. 四川大学学报: 自然科学版, 2022, 59: 013004.

[4]　程新斌. 对重大舆情与突发事件舆论引导研究的分析与对策[J]. 西南民族大学学报: 人文社会科学版, 2022, 43: 235.

[5]　Pareja A, Domeniconi G, Chen J, *et al*. Evolvegcn: Evolving graph convolutional networks for dynamic graphs [C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: AAAI, 2020.

[6]　游丹丹, 陈福集. 我国网络舆情预测研究综述[J]. 情报科学, 2016, 34: 156.

[7]　史伟, 薛广聪, 何绍义. 情感视角下的网络舆情研究综述[J]. 图书情报知识, 2022, 39: 105.

[8]　张虹, 钟华, 赵兵. 基于数据挖掘的网络论坛话题热度趋势预报[J]. 计算机工程与应用, 2007, 43: 159.

[9]　杜慧, 郭岩, 范意兴, 等. 基于因果模型的主题热度计算与预测方法[J]. 中文信息学报, 2016,

30：50.

[10] 崔彦琛，张鹏，兰月新，等. 面向时间序列的微博突发事件衍生舆情情感分析研究——以“6. 22”杭州保姆纵火案衍生舆情事件为例[J]. 情报科学，2019，37：119.

[11] 程铁军，王曼，黄宝凤，等. 基于 CEEMDAN-BP 模型的突发事件网络舆情预测研究[J]. 数据分析与知识发现，2021，5：59.

[12] 岳丽欣，刘自强，胡正银. 面向趋势预测的热点主题演化分析方法研究[J]. 数据分析与知识发现，2020，4：22.

[13] Li J，Tang H，Tan H. Research on the evolution and prediction of Internet public opinion of major pandemics-Taking the COVID-19 pandemic as an example [J]. J Phys，2021，1774：012038.

[14] Zhao L，Song Y，Zhang C，*et al*. T-gcn：A temporal graph convolutional network for traffic prediction [J]. IEEE Transactions on Intelligent Transportation Systems，2019，21：3848.

[15] Comito C，Forestiero A，Pizzuti C. Bursty event detection in Twitter streams[J]. ACM T Knowl Discov D，2019，13：1.

[16] 张孝飞，陈航行，张春花. 基于语义概念和词共现的微博主题词提取研究[J]. 情报科学，2021，39：142.

[17] Huang J，Peng M，Wang H，*et al*. A probabilistic method for emerging topic tracking in microblog stream [J]. World Wide Web，2017，20：325.

[18] 丁晟春，刘笑迎，李真. 融合评论影响力的网络舆情热点主题演化研究[J]. 现代情报，2021，41：87.

[19] 刘定一，沈阳阳，詹天明，等. 融合微博热点分析和 LSTM 模型的网络舆情预测方法[J]. 江苏大学学报：自然科学版，2021，42：546.

[20] 曾庆田，胡晓慧，李超. 融合主题词嵌入和网络结构分析的主题关键词提取方法[J]. 数据分析与知识发现，2019，3：52.

[21] 苏晓慧，张晓东，胡春蕾，等. 基于改进 TF-PDF 算法的地震微博热门主题词提取研究[J]. 地理与地理信息科学，2018，34：90.

[22] 张孝飞，陈航行，张春花. 基于语义概念和词共现的微博主题词提取研究[J]. 情报科学，2021，39：142.

[23] Yin J，Wang J. A dirichlet multinomial mixture model-based approach for short text clustering [C]//Proceedings of the 20th ACM SIGKDDInternational Conference on Knowledge Discovery and Data Mining. [S. l；s. n.]，2014：233.

[24] Mazarura J，De Waal A. A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text [C]//Proceedings of the 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference. [S. l]：IEEE，2016：1.

[25] 马思丹，刘东苏. 基于加权 Word2vec 的文本分类方法研究[J]. 情报科学，2019，37：38.

引用本文格式：

中　文：文雅，杨频，廖珊，等. 基于时间演化图卷积网络的舆情热点内容预测[J]. 四川大学学报：自然科学版，2023，60：033001.

英　文：Wen Y，Yang P，Liao S，*et al*. A temporal evolving graph convolutional network for Public opinion prediction in emergencies [J]. J Sichuan Univ：Nat Sci Ed，2023，60：033001.