

基于社交圈层和注意力机制的信息热度预测

郑作武 邵斯绮 高晓沅 陈贵海

(上海交通大学计算机科学与工程系 上海 200240)

摘 要 社交网络现已成为现实世界中信息传播与扩散的主要媒介,对其中的热点信息进行建模和预测有着广泛的应用场景和商业价值,比如进行信息传播挖掘、广告推荐和用户行为分析等.目前的相关研究主要利用特征和时间序列进行建模,但是并没有考虑到社交网络中用户的社交圈层对于信息传播的作用.本文提出了一种基于社交圈层和注意力机制的热度预测模型 SCAP(Social Circle and Attention based Popularity Prediction),首先对社交圈层进行定义,通过自动编码器提取用户历史文本序列的特征,对不同用户的社交圈层进行聚类划分,得到社交圈层特征.进而对于一条新发布的文本信息,通过长短期记忆网络与嵌入层提取其文本特征、用户特征和时序特征,并基于注意力机制,捕获到不同社交圈层对于该文本信息的影响程度,得到社交圈层注意力特征.最后将文本特征、用户特征、时序特征和社交圈层注意力特征进行特征融合,并通过两个全连接层进行建模学习,对社交信息的热度进行预测.在推特、微博和豆瓣等四个数据集上的实验结果表明,SCAP模型的预测表现相比于多个对比模型总体呈优,在不同数据集上均方误差(MSE)分别降低了0.017,0.022,0.021和0.031,F1分数分别提升0.034,0.021,0.034和0.025,能够较为准确地预测社交信息的热度.本文同时探究了不同实验参数对于模型的影响效果,如用户历史文本序列的数量、社交圈层的数量和时间序列的长度,最后验证了模型输入的各个特征和注意力机制的引入对于模型预测性能提升的有效性,在推特数据集中,引入社交圈层和注意力机制,模型的MSE指标分别降低了0.065和0.019.

关键词 社交网络;热度预测;社交圈层;注意力机制;用户偏好

中图法分类号 TP18 DOI号 10.11897/SP.J.1016.2021.00921

Social Circle and Attention Based Information Popularity Prediction

ZHENG Zuo-Wu SHAO Si-Qi GAO Xiao-Feng CHEN Gui-Hai

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240)

Abstract Social network has become the main medium for people to discuss what happens in the real world with the popularity of the Internet. On social platforms, such as Twitter, Weibo, and WeChat, user generated content (UGC) can form a rich data stream and disseminate it, so that users can have immediate insight into the hot information that are happening and the information generated around them. Modeling and predicting hot information on social networks have a wide range of application scenarios and commercial values, such as information dissemination mining, advertising recommendation, and user behavior analysis, which has attracted much attention in the field of data mining and social network analysis. In recent years, methods of popularity prediction mainly can be divided into two categories, including feature based methods and time series based methods, from which it has been extensively studied, such as feature engineering, utilizing stochastic process to model temporal sequence of information. However, these techniques either

收稿日期:2020-04-30;在线发布日期:2020-12-16. 本课题得到国家重点研发项目(2020YFB1707903)、国家自然科学基金(61872238, 61972254)、CCF-腾讯科研基金(RAGR20200105)、腾讯广告犀牛鸟专项研究计划(FR202001)、华为云项目(TC20201127009)资助. 郑作武,博士研究生,中国计算机学会(CCF)学生会会员(59662G),主要研究方向为社交网络分析、信息热度预测. E-mail: waydrow@sjtu.edu.cn. 邵斯绮,本科生,主要研究方向为社交网络、自然语言处理. 高晓沅(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为数据工程、数据库索引、网络优化. E-mail: gao-xf@cs.sjtu.edu.cn. 陈贵海,博士,教授,中国计算机学会(CCF)会士,主要研究领域为计算机网络、分布式计算、无线网络.

require a long-term observation of the information or hand-craft features that are expensive to extract. Besides, they fail to consider the role of social circles of users when information is spread on social networks. Different users may care about different information and are affected by different user groups at the same time, while users in the same social circle tend to pay attention to similar information and have closer connections. It will further encourage users in the same circle to influence each other. In this paper, we propose SCAP, a Social Circle and Attention based Popularity prediction model, which includes an autoencoder based social circle detection model and an attention based information popularity prediction model. Firstly, an autoencoder is used to extract user preference from user sequential behavior patterns, which can guide the exploration for different social circles and learning for social circle embedding by the clustering algorithm. Then for a new UGC, text embedding, user embedding, and temporal feature are extracted based on long short-term memory (LSTM) layer and embedding layer. Furthermore, we capture the influence on UGC of different social circles through attention mechanism, which indicates the weight of different social circles. Finally, using social circle attention, text embedding, user embedding, and temporal feature, the popularity of UGC is predicted by two fully connected layers. We conduct extensive experiments on 4 real world datasets from Twitter, Weibo, and Douban Event platforms. The experimental results show that the performance of SCAP is better than baseline methods. We first evaluate the effectiveness from regression and classification perspectives. The mean square error (*MSE*) of SCAP on different datasets decreases by 0.017, 0.022, 0.021, and 0.031, respectively, and the *F1*-score increases by 0.034, 0.021, 0.034, and 0.025, respectively. We also consider the impacts of different parameters on the model, including the number of users' recent sequences, the number of social circles, the length of temporal sequences. In addition, the effectiveness of different features and attention mechanism are validated in ablation study. For example, in Twitter dataset, the *MSE* decreases by 0.065 and 0.019 when integrating social circle and attention mechanism, respectively.

Keywords social networks; popularity prediction; social circle; attention mechanism; user preference

1 引 言

近年来,随着互联网的普及,社交网络逐渐成为讨论现实世界中发生事情的主要媒介.在社交平台上,如推特(Twitter)、微博、微信等,用户生成内容(User Generated Content, UGC)可以形成丰富的数据流并进行传播,从而让用户立即洞察正在发生的热点事件.在社交网络中,用户生成内容的热度预测在营销、舆论检测、广告投放等场景中有着广泛的应用需求,是最近几年数据挖掘与社交网络分析领域备受关注的研究课题.

目前有很多学者都对热度预测问题做出了相关研究,但是之前的研究工作往往更多集中于全网的热点^[1-4],例如, Fedoryszak 等人^[1]将事件建模为随时间推移的热点信息聚类集合,并构建了一个实时热点事件检测系统;Chen 等人^[2]提出了一种用

于子事件检测的编码-记忆-解码器框架,该模型以数据驱动的方式学习适合相似度度量的文档和子事件表示,利用无监督深度神经网络进行热点事件的检测;Liao 等人^[3]提出了一种时间过程和内容特征深度融合的方法,结合卷积神经网络(Convolutional Neural Networks, CNN)和循环神经网络(Recurrent Neural Network, RNN)建模文章热度变化的时间过程,并利用层次注意网络和特征嵌入技术预测网络文章的热度.另一方面,也有针对于单个用户的信息分析,比如针对同一用户发布信息的相似性,通过利用长短期记忆网络(Long Short-Term Memory, LSTM)与 CNN 模型学习序列信息之间的联系,实现用户文本内容热度预测任务^[5].针对社交网络中用户的关注和取消关注行为,结合空间特征(如网络结构)和时序特征(如用户历史发布序列),基于注意力机制进行预测^[6].

然而之前关于热度预测的研究并没有结合社交

圈层的概念. 社交圈层是社交网络中用户基于个人兴趣、关注领域、个人观点以及用户社交关联等用户特征形成的用户子集. 对于社交网络中的用户来说, 不同用户可能会关心不同的信息, 同时受到不同用户群体的影响, 而同一社交圈层中的用户则倾向于关注相似信息, 并拥有更加密切的联系, 这也会进一步促使同圈层内的用户互相影响, 如群体思维和群体极化均是群体影响的实例, 即观点类似的群体一定程度上会强化和影响个体态度^[7].

社交圈层的划分有多种方式, 如基于语义和偏好的划分、基于网络结构的划分等. 针对基于语义和偏好的划分, 可以将用户按领域、偏好等划分成不同的圈层, 比如对于一条深度学习领域的最新科研发现内容, 可能会在计算机相关科研人员和相关专业的学生朋友圈和群聊中被多次转发传播, 即该圈层中用户更加倾向计算机领域的内容, 发布的信息也与其相关. 所以对于某条文本信息的热度, 进行预测时可以考虑不同圈层对其的影响, 从而获得更加精准的预测结果; 而基于网络结构的划分则关注于用户的朋友关系网络. 社交网络中, 用户之间具有关注与被关注的好友关系, 具有好友关系的用户之间相似性也会更高. 随着网络中这种关系的不断扩展, 会形成不同的用户子集, 即用户社区, 这也是社交圈层的一种体现. 同一社区内的用户连接较紧密, 互动更加频繁, 不同社区之间连接较稀疏. 在热度预测问题中, 考虑社区的影响往往需要对社交网络的结构进行分析, 提取网络结构特征, 但是由于现实社交网络的规模庞大, 提取结构特征通常需要较高的时间复杂度^[8], 因此, 本文关注于基于语义和偏好的社交圈层在热度预测问题中的应用.

基于上述思想, 本文提出了基于社交圈层和注意力机制的信息热度预测模型 SCAP (Social Circle and Attention based Popularity Prediction), 该模型主要分为基于自动编码器的用户社交圈层划分模型和基于注意力机制的社交信息热度预测模型两部分. 首先, 在社交圈层划分模型中, 利用自动编码器 (AutoEncoder) 对不同用户发布的历史文本序列进行特征学习, 得到用户偏好的特征表示, 然后将其通过 K -Means++ 聚类算法划分为不同的社交圈层, 得到社交圈层特征, 作为热度预测模型的输入向量. 在热度预测模型中, 使用 LSTM 层和嵌入层提取社交信息的文本特征、用户特征和时序特征. 进而, 通

过注意力机制, 捕获该文本信息对重点社交圈层的关注度, 得到社交圈层注意力特征. 最后结合文本特征、用户特征、时序特征和社交圈层注意力特征, 通过两个全连接层进行建模学习, 得到用户生成内容的热度预测结果. 本文在多个真实的数据集上进行实验, SCAP 模型在热度预测方面, 相比于多个对比模型, 在均方误差 (Mean Square Error, MSE)、平均绝对误差 (Mean Absolute Error, MAE)、F1 分数和准确率 (Accuracy) 多个指标上均取得了较好的提升. 进而探究了不同实验参数, 如用户历史序列数量、社交圈层聚类个数和时间序列长度, 对于模型的影响. 同时通过模型简化测试和注意力机制可视化, 验证了模型各输入特征和注意力机制的引入对于模型性能提升的有效性.

本文的主要贡献在于: (1) 提出了一个用户社交圈层划分模型, 可以根据社交网络中用户的偏好特征对社交圈层自动划分, 并将其应用到热度预测领域; (2) 利用注意力机制分析不同社交圈层对于用户和文本的影响, 进而结合文本特征、用户特征和时序特征来实现用户生成内容的热度预测任务; (3) 在多个真实数据集上的对比实验验证了 SCAP 模型的有效性, 可以较好地预测社交信息的热度.

本文第 2 节介绍相关的研究工作; 第 3 节给出问题描述和模型介绍, 主要包括社交圈层划分模型和热度预测模型; 第 4 节介绍实验环境、实验设计和结果分析; 第 5 节给出总结和未来研究方向.

2 相关工作

社交平台的信息热度预测问题一直以来广受研究者关注, 目前学术界主流的方法主要分为基于特征驱动、时序过程建模和协同过滤的三类方法^[8]. 另外, 近年来, 深度学习模型被广泛应用于热度预测问题中. 2.1 节介绍了目前主流的热度预测方法, 2.2 节介绍了注意力机制在热度预测模型中的应用.

2.1 热度预测问题

2.1.1 基于特征驱动的方法

早期的热度预测模型主要基于特征进行建模. 该类模型尝试选择和提取文本信息在传播过程中的有效特征, 包括内容特征、用户特征、网络结构特征等, 然后训练传统机器学习方法来进行热度预测. 内容特征包括从文本中直接提取的简单特征, 包括文本长度、特定词汇、表情使用等^[9-10], 以及通过自然

语言处理方法提取的文本内容中的话题信息和情感信息^[11-13]. 用户特征包括用户属性和历史行为, 用户属性是指可以直接从用户首页获取的信息, 包括用户的注册时长、关注者和粉丝的数量等, 由此考虑用户影响力对于热度的影响^[14-15]. 历史行为是指用户的历史文本和转发行为、点赞行为、评论行为等. 根据用户行为的连续性, 使用这些特征来预测用户的历史偏好^[16-18]和下一个时期用户特定行为的可能性^[19-20]. 网络结构特征是从用户关系网络中提取的, 例如转发网络、朋友关系网络等. 这些特征显示了社交网络中用户之间的交互行为与信息的传播过程. Cha 等人^[21]将热度与信息的传播速度建立联系. EPAB 模型基于时序特征、用户特征和网络结构特征提出一种基于贝叶斯网络的热度预测模型, 提出了早期模式的概念, 以建立早期特征信息与未来热度变化之间的关系^[22]. 然而, 该类模型高度依赖于特征工程, 特征的选择需要很多专家知识, 如果所选特征不够准确, 会导致模型性能不良. 并且, 由于不同数据集所能提取到的特征不同, 该类模型的迁移能力相对较弱.

2.1.2 基于时序过程的方法

在热度预测任务中, 基于时序特征的点过程模型是一种有效的模型. 该类模型思想主要借鉴了金融学和传染病学模型, 社交平台上的文本信息会激发一个连续的转发序列, 序列中的每次转发都可以视作是逐点发生的, 因此可以通过点过程来对社交网络中的行为动力学进行统计学建模, 然后进行热度的预测. 例如, 强化泊松过程(Reinforced Poisson Processes, RPP)模型使用非均匀泊松过程模型来模拟每次转发的当前状态, 并考虑了“富者愈富”机制与时间松弛效应^[23]. SEISMIC 模型基于自激点过程理论, 假定过去的转发操作将影响该过程的未来演变, 采用了双重随机过程刻画信息的传染性与转发事件的到达时间^[24]. Yu 等人^[25]基于韦伯分布(Weibull Distribution)进行建模, 并将节点行为动力学的参数与其邻居节点的行为特征之间的强相关性用于参数学习. 但是, 时间序列模型存在一定的缺陷, 该类模型要求数据集包含时序信息, 对于不满足该条件的数据集无法进行建模. 同时, 该类模型仅考虑时间序列信息, 对于未观察到转发操作的文本信息, 时间序列模型将无法进行预测.

2.1.3 基于协同过滤的方法

基于协同过滤的方法思想来源于推荐系统. 与

基于特征和时间序列的方法不同, 协同过滤方法主要集中于用户交互, 在已知其中一个用户的历史转发操作后, 它使用未知用户与已知用户的相似性来预测未知用户是否会转发消息. 协同过滤方法使用矩阵分解(Matrix Factorization)来填充并预测缺失的转发信息, 但是矩阵分解模型仅关注用户的转发操作, 会忽略其余可用信息; 并且模型不能准确地区分负面数据和缺失数据, 即无法确定用户不转发某条文本信息是由于不想转发还是没有看到该条文本. 因此, 研究者对其进行了一系列改进, 对于前一个缺陷, ContextMF 模型将用户的转发可能性与用户的偏好协同进行建模, 即偏好相似的用户有更高的概率转发相同文章^[26]; Li 等人^[27]将文本信息空间转换到关键词空间, 且具有相同关键词的文本信息更可能被同一用户转发. 对于后一个缺陷, OCCF 模型将所有缺失数据视为负值, 但通过在示例之间增加权重来区分负面数据和缺失数据^[28].

2.1.4 深度学习算法的应用

近几年来, 随着深度学习算法的深入发展, 神经网络被广泛应用于热度预测领域. 其中, 循环神经网络 RNN 及其变种的长短期记忆网络 LSTM、门控循环单元网络(Gated Recurrent Unit, GRU)由于在时间序列输入上表现出色, 在社交网络的热度预测问题中得到了充分地应用. Ye 等人^[29]提出了基于上下文嵌入的 LSTM 递归网络模型, 同时利用历史时间序列数据和上下文信息来实现热度预测任务. DeepCas 模型则通过随机游走得到用户网络信息, 并将其输入 GRU 神经网络结构中进行预测^[30]. DeepHawkes 模型使用 RNN 将整个转发过程进行编码, 同时考虑用户影响的时效效应与网络结构影响^[31]. Lamprindis 等人^[32]通过网络结构、主题、词汇等因素提取用户偏好, 利用多任务 GRU 网络, 对新闻标题的热度进行预测. 研究者还将 RNN 与点过程模型结合起来进行建模^[33-35]. 此外, 图神经网络(Graph Neural Network, GNN)可以对网络中节点间的依赖关系进行建模, CoupledGNN 模型对用户之间的级联影响进行了研究, 将其应用于微博热度的预测中^[36]; Sankar 等人^[37]利用 GNN 与变分自编码器提出了 Inf-VAE 模型, 对社会同质性和时序影响共同建模, 以预测用户的转发行为.

但是, 目前各类热度预测模型输入的特征以文本信息、时序信息以及用户信息为主, 没有考虑到用户社交圈层的影响. 本文利用用户历史文本序列对

用户社交圈层进行划分,并结合社交圈层特征、文本特征和用户特征综合地进行热度预测。

2.2 注意力机制

注意力机制(Attention Mechanism)借鉴人类的视觉注意力特点,Bahdanau 等人^[38]提出其核心思想是通过学习注意力分布,来从众多信息中提取出对当前任务目标更关键的信息.近年来,注意力机制被广泛应用于热度预测任务的建模中.例如,基于时序特征和时序注意力机制的深度时间背景网络,用以预测社交媒体中图片的受欢迎程度^[39].Chen 等人^[40]将注意力机制使用在文本内容编码器、用户编码器和时间序列编码器中,学习文本,用户和时间序列表示,最后通过融合层将这三种编码器产生的特征进行组合,并输出预测结果.Zhang 等人^[41]首先应用内部注意力机制来分别学习文本内容和图像内容的嵌入,再应用交互注意力机制来学习用户在文本和图像内容上的不同注意力.Wang 等人^[5]使用显式和隐式结合的双重注意力模型,前者用于表示发布内容的图像和文本信息,后者引入用户特征的平均值来表示用户环境,通过学习用户环境来完成预测任务.Song 等人^[42]利用注意力机制提取文章中具有吸引力的句子,进而确定主题的热度.本文

使用注意力机制来学习社交圈层对于文本信息传播的影响,最终实现热度预测任务。

3 社交信息热度预测模型 SCAP

在本章节中,将介绍本文提出的基于社交圈层和注意力机制的热度预测模型 SCAP(Social Circle and Attention Based Popularity Prediction),主要包括两个部分:基于自动编码器(AutoEncoder)的用户社交圈层划分模型和基于注意力机制(Attention Mechanism)的社交信息热度预测模型.模型的主要框架如图 1 所示,本文首先基于用户发布的历史文本序列利用自动编码器进行特征学习,得到用户偏好的特征表示,进而通过 K -Means++ 聚类算法自动划分不同的社交圈层,作为热度预测模型的输入向量.其次利用长短期记忆网络(Long Short-Term Memory, LSTM)和嵌入层提取社交信息的文本特征、用户特征和时序特征.进而通过注意力机制,捕获该文本信息对不同社交圈层的关注度,得到社交圈层注意力特征.最后结合文本特征、用户特征、时序特征和社交圈层注意力特征,用两个全连接层进行建模,最终得到社交信息的热度预测结果。

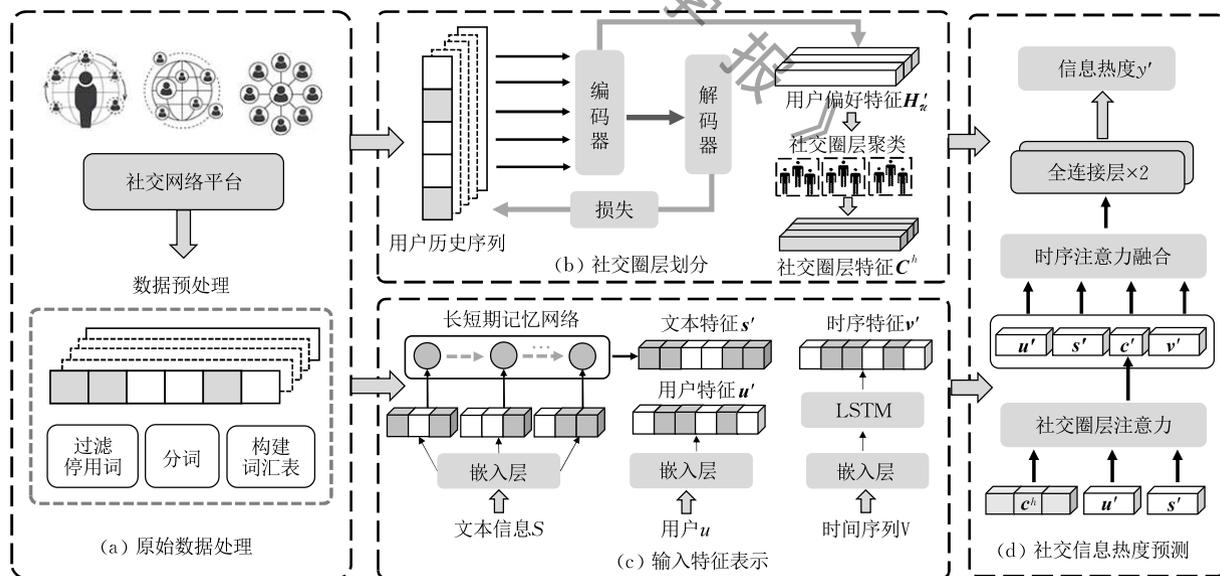


图 1 基于社交圈层和注意力机制的热度预测模型(SCAP)框架图

3.1 相关定义

为了方便阅读和理解,本节将主要介绍相关模型用到的定义及符号表示.在社交网络中,用户集合 $U = \{u_1, u_2, \dots, u_{n_u}\}$,其中 n_u 表示用户的数量.对于每一个用户 $u \in U$,其历史文本序列定义如下:

定义 1. 用户历史文本序列.某个用户 u 发布的 M 条历史信息序列 $h_u = \{S_1, S_2, \dots, S_M\}$,其中每条文本 $S_i = \{\omega_1, \omega_2, \dots, \omega_{n_w}\}$, $\omega_j \in \mathcal{V}$,即文本 S_i 是由词汇表 \mathcal{V} 中的单词 ω_j 组成的长度为 n_w 的单词序列。

定义 2. 社交圈层. 给定用户集合 \mathcal{U} , 社交圈层 $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$ 且 \mathcal{C} 是 \mathcal{U} 的一个划分, 即 \mathcal{U} 的非空子集的集合, 对于每一个 $c_i = \{u_j | u_j \in \mathcal{U}\}$, 其中的用户具有相似的兴趣和特征, k 表示社交圈层的数量.

定义 3. 热度时间序列. 对于社交网络中用户 u_i 发表的信息文本 S_i , 根据给定一段时间 T 内的总交互次数 (本文采用转发次数), 作为 S_i 最终的信息热度 y_i . 为了方便计算, 给定时间间隔 Δt , 将时间 T 离散化为多个时间段, 并统计每个时间段内的转发次数, 则给定任意时间段 t , 热度时间序列 V_i 表示为 $\{y_i^1, y_i^2, \dots, y_i^t\}$.

定义 4. 信息热度预测问题. 给定社交网络中 u_i 发布的一条信息文本 S_i , $\{u'_i, S'_i, v'_i, C^h\}$ 是一个四元组, 分别表示用户特征, 文本特征, 时间序列特征和社交圈层 C 的特征, 则该条文本信息的热度 $y_i = f(u'_i, S'_i, v'_i, C^h)$, 预测问题目标即学习映射函数 f 以预测热度 y_i .

3.2 社交圈层划分模型

本节介绍如何基于用户的历史文本序列进行特征学习, 并得到社交圈层的划分. 首先, 对于用户的历史文本序列进行处理, 表示为文本向量. 其次利用 LSTM 自动编码器学习历史文本序列的压缩表示, 作为用户的偏好特征向量, 最终通过 K -Means++ 聚类得到社交圈层.

3.2.1 历史文本序列特征表示

首先, 对于文本信息进行分词, 构建词汇表, 得到每条文本的向量表示. 其次, 由于每条文本包含的单词数量并不相同, 在模型训练过程中为了处理不同长度文本的情况, 本文使用了自然语言处理 (Natural Language Processing, NLP) 中的截断或填充 (Padding) 操作, 以得到相同维度的向量^[43], 并通过对数据分布的统计合理设置向量维度. 这样对于用户 u 的每条文本信息向量 $S_i = \{x_1, x_2, \dots, x_L\}$, $S_i \in \mathbb{R}^L$, 其中 L 表示序列的最大长度, 则用户 u 的历史文本序列表示矩阵 $\mathbf{h}_u = \{S_i\}_{i=1}^M$, $\mathbf{h}_u \in \mathbb{R}^{M \times L}$, M 为历史文本序列的数量.

3.2.2 基于自动编码器的用户偏好特征提取

本文通过对用户的历史文本序列矩阵进行特征挖掘提取, 来得到用户的偏好特征表示. 长短期记忆网络 LSTM^[44] 是一种循环神经网络, 其在处理序列数据、文本序列建模、语言理解等方面有着良好的表现^[45], 可用于用户文本序列的处理. 这里需要注意的是, 由于用户偏好特征的学习没有对应的训练标签, 同时用户不同历史文本之间可能存在一定的联

系, 为了综合地对多条序列文本学习, 同时增加模型的可扩展性, 本文采用自动编码器的思想进行自我监督学习, 即构建一个使用编码器—解码器 LSTM 架构^[46-47] 的模型, 来实现序列数据的自动编码和特征提取, 模型的输入和输出是相同的用户历史序列文本矩阵.

对于给定的用户 u , 模型的输入序列矩阵为 $\mathbf{h}_u = \{S_i\}_{i=1}^M$, $u \in \mathcal{U}$, 经过编码器 $\Phi: \mathbf{h}_u \rightarrow \mathbf{h}'_u$, 即将 \mathbf{h}_u 通过 LSTM 层, 该层位于 i 位置的隐含状态 hd_i 表示如下:

$$hd_i = \text{LSTM}(S_i, hd_{i-1}) \quad (1)$$

经过迭代更新, 可以得到用户历史文本序列的嵌入向量表示 $\mathbf{h}'_u \in \mathbb{R}^{K_C}$, K_C 为 \mathbf{h}'_u 向量的维度. 进而传入解码器 $\Psi: \mathbf{h}'_u \rightarrow \mathbf{h}_u$, 即重建原始用户历史文本序列矩阵 \mathbf{h}_u , 过程与编码器相反. 这样自动编码器的两层网络定义如下:

$$x' = \text{LSTM}(\mathbf{h}_u) \quad (2)$$

$$\mathbf{h}'_u = g_1(x') \quad (3)$$

$$y' = \text{LSTM}(\mathbf{h}'_u) \quad (4)$$

$$\hat{\mathbf{h}}_u = g_2(y') \quad (5)$$

其中, g_1 和 g_2 分别是编码器和解码器的激活函数, 本文使用了 ReLU (Rectified Linear Unit) 激活函数. 需要注意的是, 编码器和解码器的 LSTM 层的参数和权重是不同的. 最后根据式 (2)~(5), 该自动编码器的损失函数 (Loss Function) 定义为

$$\mathcal{L}_u = \|\hat{\mathbf{h}}_u - \mathbf{h}_u\|^2 \quad (6)$$

利用式 (6) 通过标准的反向传播程序来训练神经网络. 训练完成后, 将解码器移除, 只保留编码器, 该编码器可用于将输入序列矩阵编码为固定长度的向量, 本文将该模型中编码器的输出 \mathbf{h}'_u 作为用户 u 的偏好特征向量表示. 相较于传统的用户偏好学习算法, 该模型不需要人工进行复杂特征的提取 (如用户的朋友网络、信息转发链等), 利用用户历史的文本序列即可自动挖掘用户的偏好特征.

3.2.3 基于 K -Means++ 的社交圈层划分

K -Means 聚类是一种无监督的分类算法, 本文利用上述编码器模型的输出 \mathbf{h}'_u , 得到用户集合 \mathcal{U} 的偏好特征矩阵 $\mathbf{H}'_u = \{\mathbf{h}'_{u_i} | u_i \in \mathcal{U}\}_{i=1}^{n_u}$ 作为聚类算法的输入, 最终得到社交圈层 $\mathcal{C} = \{c_1, c_2, \dots, c_k\}$, 所以聚类的个数为 k .

首先需要选择初始的 k 个聚类中心, 即簇中心, 而 K -Means 算法通过随机选择的方式确定初始簇中心, 其产生的聚类结果依赖于初始簇中心的选择,

如果初始簇中心选择不好,就会陷入局部最优解.为了解决这一问题,本文利用改进的 K -Means++ 算法,首先在输入集合 \mathbf{H}'_u 中随机选择一个样本点作为第一个簇中心 \mathbf{c}_1^h ,计算剩余样本点与当前所有簇中心的最短距离,定义为

$$D(\mathbf{h}'_{u_i}) = \min\{dist(\mathbf{h}'_{u_i}, \mathbf{c}_1^h), \dots, dist(\mathbf{h}'_{u_i}, \mathbf{c}_n^h)\} \quad (7)$$

则某样本点被选为下一个簇中心的概率

$$P(\mathbf{h}'_{u_i}) = \frac{D(\mathbf{h}'_{u_i})}{\sum_{\mathbf{h}'_{u_j} \in \mathbf{H}'_u} D(\mathbf{h}'_{u_j})} \quad (8)$$

这样可以得到 k 个初始聚类中心.进而采用标准的 K -Means 算法,聚类得到社交圈层 C ,以及与之对应的聚类中心集合 $\mathbf{C}^h = \{\mathbf{c}_i^h\}_{i=1}^k$,对于每一个 $\mathbf{c}_i^h \in \mathbf{C}^h$,本文将 \mathbf{c}_i^h 作为社交圈层 c_i 的特征向量, $\mathbf{c}_i^h \in \mathbb{R}^{K_C}$, $\mathbf{C}^h \in \mathbb{R}^{k \times K_C}$.

社交圈层划分模型的算法描述如算法 1 所示.

算法 1. 基于自动编码器的社交圈层划分算法.

输入: 用户集合 \mathcal{U} , 历史文本序列 \mathbf{h}_u , 聚类个数 k

输出: 社交圈层特征矩阵 $\mathbf{C}^h = \{\mathbf{c}_1^h, \mathbf{c}_2^h, \dots, \mathbf{c}_k^h\}$

1. FOREACH $u \in \mathcal{U}$ DO
2. FOREACH $S_i \in \mathbf{h}_u$ DO
3. 将文本 S_i 向量化生成 $\mathbf{S}_i \in \mathbb{R}^L$
4. END FOR
5. END FOR
6. 生成用户历史序列矩阵 $\mathbf{h}_u \in \mathbb{R}^{M \times L}$
7. WHILE *training* DO
8. 根据式(6)最小化损失函数
9. END WHILE
10. 计算用户 u 的偏好向量 $\mathbf{h}'_u = \text{encoder}(\mathbf{h}_u)$
11. 计算用户集合 \mathcal{U} 偏好特征矩阵 $\mathbf{H}'_u = \{\mathbf{h}'_{u_i}\}_{i=1}^{i=n_u}$
12. 随机初始化聚类中心集合 $\mathbf{C}^h = \{\mathbf{c}_i^h\}$
13. WHILE 聚类中心个数 $n \neq k$ DO
14. 根据式(7)、(8)计算 $D(\mathbf{h}'_{u_i}), P(\mathbf{h}'_{u_i})$
15. 按照 $P(\mathbf{h}'_{u_i})$ 选择聚类中心 \mathbf{c}_i^h 放入 \mathbf{C}^h
16. END WHILE
17. 聚类中心集合 $\mathbf{C}^h = K\text{Means}(\mathbf{H}'_u)$
18. RETURN 社交圈层特征矩阵 \mathbf{C}^h

3.3 特征的嵌入表示

3.3.1 文本特征表示

在社交网络中,对于用户 u 新发布的某一条信息文本 $S = \{\omega_i\}_{i=1}^{i=n_w}$,采取与历史文本序列特征处理相同的思路,可以得到初始的 $\mathbf{S} = \{x_1, x_2, \dots, x_L\}$, $\mathbf{S} \in \mathbb{R}^L$,对应于定义 4 中的文本特征向量.其次将 \mathbf{S} 通过嵌入层,每一个 $x_i \in \mathbf{S}$ 转化为嵌入向量:

$$\mathbf{x}'_i = \mathbf{W}_X x_i \quad (9)$$

其中, \mathbf{W}_X 是词嵌入矩阵, $\mathbf{x}'_i \in \mathbb{R}^{K_S}$,在获得文本嵌入

矩阵 $\{\mathbf{x}'_i\}_{i=1}^L$ 后,本文利用 LSTM 层对其进行特征提取,生在隐含状态序列.在每一个时间步 t ,每个 LSTM 单元对应一个输入门 i_t ,一个忘记门 f_t ,一个输出门 o_t 和一个细胞状态 c_t ,则隐含状态 h_t 可通过如下方式计算:

$$i_t = \sigma(\mathbf{W}_{X_i} \mathbf{x}'_t + \mathbf{W}_{H_i} h_{t-1} + b_i) \quad (10)$$

$$f_t = \sigma(\mathbf{W}_{X_f} \mathbf{x}'_t + \mathbf{W}_{H_f} h_{t-1} + b_f) \quad (11)$$

$$o_t = \sigma(\mathbf{W}_{X_o} \mathbf{x}'_t + \mathbf{W}_{H_o} h_{t-1} + b_o) \quad (12)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \tanh(\mathbf{W}_{X_c} \mathbf{x}'_t + \mathbf{W}_{H_c} h_{t-1} + b_c) \quad (13)$$

$$h_t = o_t \circ \tanh(c_t) \quad (14)$$

其中,“ \circ ”表示哈达玛积(Hadamard Product), $\mathbf{W}_X, \mathbf{W}_H, b$ 是 LSTM 层的待学习参数, σ 表示 sigmoid 激活函数.经过循环计算每个时间步,可以得到隐含状态的序列 $\{h_t\}_{t=1}^L$,最终文本特征的嵌入矩阵表示 $\mathbf{S}' = \{h_1, h_2, \dots, h_L\}$, $\mathbf{S}' \in \mathbb{R}^{L \times K_S}$.

3.3.2 用户特征表示

对于社交信息文本的发布者,本文使用了用户 ID、性别、关注者数量、粉丝数量、发布信息数量共 5 个特征进行特征拼接作为用户初始特征表示 $\mathbf{u} = \{u_x^1, u_x^2, u_x^3, u_x^4, u_x^5\}$, $\mathbf{u} \in \mathbb{R}^5$,其中,用户 ID、关注者数量、粉丝数量、发布信息数量均为实数表示,性别采用 0/1 表示.将 \mathbf{u} 通过嵌入层,得到用户特征的嵌入矩阵 $\mathbf{U}' \in \mathbb{R}^{5 \times K_U}$.

3.3.3 时序特征表示

与文本特征的提取类似,LSTM 可以挖掘到时间序列的特征,其中的细胞状态保证了可以学习到序列间的长期依赖信息.给定时间段 t 和对应的热度时间序列 $V = \{y^1, y^2, \dots, y^t\}$,将其通过嵌入层,得到嵌入矩阵 $\mathbf{V} \in \mathbb{R}^{t \times K_T}$.进而将其输入到 LSTM 中,将最终的隐含状态序列作为时序特征表示,即 $\mathbf{V}' = \{h_1, h_2, \dots, h_t\}$, $\mathbf{V}' \in \mathbb{R}^{t \times K_T}$.

最终,本文得到了文本特征的嵌入矩阵 \mathbf{S}' ,用户特征的嵌入矩阵 \mathbf{U}' ,时序特征的嵌入矩阵 \mathbf{V}' ,以及在社交圈层划分模型中计算的社交圈层特征矩阵 \mathbf{C}^h , $\{\mathbf{S}', \mathbf{U}', \mathbf{V}', \mathbf{C}^h\}$ 将作为热度预测模型的输入.

3.4 基于注意力机制的热度预测模型

首先,输入的维度 $\mathbf{S}' \in \mathbb{R}^{L \times K_S}$, $\mathbf{U}' \in \mathbb{R}^{5 \times K_U}$, $\mathbf{V}' \in \mathbb{R}^{t \times K_T}$, $\mathbf{C}^h \in \mathbb{R}^{k \times K_C}$,其中 K_S, K_U, K_T 和 K_C 分别是文本嵌入向量、用户嵌入向量、时序嵌入向量和社交圈层特征向量的维度.在本文中,为了方便表示, $L = 50, K_S = 128, K_U = 128, K_T = 128, K_C = 128$.

3.4.1 社交圈层注意力机制建模

对于社交网络中的用户,其新发布的文本信息会受到不同社交圈层的影响,同时每个圈层对于其

信息热度的贡献值也是不同的. 所以, 本文提出的 SCAP 模型利用注意力机制, 捕获不同社交圈层对于信息的影响程度, 以获得实际的社交圈层表示. 在社交圈层注意力的计算过程中, 文本嵌入矩阵 \mathbf{S}' 和用户嵌入矩阵 \mathbf{U}' 将协同参与模型的训练.

为方便后续社交圈层注意力的计算, 借鉴卷积神经网络中的平均池化 (mean-pooling) 思想^[43], 首先将文本特征嵌入矩阵 $\mathbf{S}' \in \mathbb{R}^{L \times K_s}$ 和用户特征嵌入矩阵 $\mathbf{U}' \in \mathbb{R}^{5 \times K_u}$, 按如下公式转化为向量表示:

$$\mathbf{s}' = \frac{\mathbf{V}_1^s \mathbf{S}'}{L} \quad (15)$$

$$\mathbf{u}' = \frac{\mathbf{V}_1^u \mathbf{U}'}{5} \quad (16)$$

其中, $\mathbf{V}_1^s \in \mathbb{R}^{1 \times L}$, $\mathbf{V}_1^u \in \mathbb{R}^{1 \times 5}$ 是两个元素全为 1 的向量, 通过该操作后, 文本和用户特征表示均为向量形式.

其次, 社交圈层注意力的计算公式定义如下:

$$r_i^c = \mathbf{W}_C'' (\tanh(\mathbf{W}_C^h) \circ \tanh(\mathbf{W}_S^s \mathbf{s}') \circ \tanh(\mathbf{W}_U^u \mathbf{u}')) \quad (17)$$

其中, r_i^c 即表示社交圈层中第 i 个圈层的影响力大小, 式中的参数满足 $\mathbf{W}_C'' \in \mathbb{R}^{1 \times 128}$, $\mathbf{W}_C^h \in \mathbb{R}^{128 \times 128}$, $\mathbf{W}_S^s \in \mathbb{R}^{128 \times 128}$, $\mathbf{W}_U^u \in \mathbb{R}^{128 \times 128}$. 通过上式, 可以计算得到每个圈层对于文本和用户特征表示的相关程度. 因此, 文本特征和用户特征可以促进社交圈层注意力的计算, 提供潜在的引导, 同时挖掘不同的用户圈层对于最终信息热度的影响.

对于社交圈层的注意力分配概率分布 α^c ,

$$\alpha^c = \text{Softmax}(r^c) \quad (18)$$

最终, 基于式(18), 可以得到基于不同注意力分配权重的社交圈层注意力特征表示

$$\mathbf{c}' = \sum_{i=1}^k \alpha_i^c \cdot \mathbf{c}_i^h \quad (19)$$

3.4.2 时序注意力特征融合

在信息传播的过程中, 不同于文本特征、用户特征和社交圈层特征, 时序特征对于传播建模的影响力是不断变化的. 在信息发布的早期阶段, 由于观测周期较短, 无法提取到有效的时序特征用以预测, 所以模型主要依赖于其它特征信息. 而随着时间的推移, 观测到的热度时间序列更多, 时序特征在预测模型中的影响程度应该变大. 同时, 在信息传播的后期, 由于热度变化趋于稳定, 时序特征的影响也会发生变化. 所以对于文本特征、用户特征、时序特征和社交圈层特征, 在进行特征融合时, 应设置不同的权重.

首先将时序特征嵌入矩阵 $\mathbf{V}' \in \mathbb{R}^{t \times K_T}$ 根据如下公式转化为向量表示:

$$\mathbf{v}' = \frac{\mathbf{V}_1^v \mathbf{V}'}{t} \quad (20)$$

其中, $\mathbf{V}_1^v \in \mathbb{R}^{1 \times t}$ 是一个元素全为 1 的向量. 其次使用注意力机制实现动态权重调整, 模型输入为文本特征 \mathbf{s}' , 用户特征 \mathbf{u}' , 时序特征 \mathbf{v}' 和社交圈层特征 \mathbf{c}' , 则计算公式如下:

$$r_i' = \mathbf{W}_i \tanh(\mathbf{W}_i^s \mathbf{s}' + \mathbf{W}_i^u \mathbf{u}' + \mathbf{W}_i^v \mathbf{v}' + \mathbf{W}_i^c \mathbf{c}') \quad (21)$$

$$\alpha' = \text{Softmax}(r') \quad (22)$$

则最终的特征组合为

$$\mathbf{m}' = \alpha_s' \mathbf{s}' + \alpha_u' \mathbf{u}' + \alpha_v' \mathbf{v}' + \alpha_c' \mathbf{c}' \quad (23)$$

3.4.3 信息热度预测

得到特征组合后, 利用两个全连接层, 最终生成热度预测结果, 计算过程如以下公式所示:

$$y' = f(\mathbf{m}') = \mathbf{W}_M'' g'(\mathbf{W}_M' \mathbf{m}' + \mathbf{b}_M') + \mathbf{b}_M'' \quad (24)$$

其中, $\mathbf{W}_M' \in \mathbb{R}^{128 \times 128}$ 和 $\mathbf{b}_M' \in \mathbb{R}^{128}$ 是第一层的学习参数, $\mathbf{W}_M'' \in \mathbb{R}^{128}$ 和 $\mathbf{b}_M'' \in \mathbb{R}$ 是第二层的学习参数, g' 是 ReLU 激活函数, y' 即信息热度.

本文将社交信息的热度预测看作是一个回归问题, 即能够定量地分析信息的热度, 而不仅仅是当作分类问题来计算信息是否会变得流行. 所以, 在 SCAP 模型的训练过程中, 采用了均方误差 (Mean Square Error, MSE) 作为模型的优化指标, 则损失函数可以定义为

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N (y_i - y_i')^2 \quad (25)$$

其中, N 表示训练样本的数量. 本文提出的 SCAP 热度预测模型的算法描述如算法 2 所示.

算法 2. SCAP 热度预测模型.

输入: 用户集合 U , 训练数据 D

输出: 参数集合 Θ

1. 初始化训练轮数 $Epochs = 100$, Batch 大小 $B = 128$, 随机初始化参数集合 Θ
2. 根据算法 1 计算社交圈层特征矩阵 \mathbf{C}^h
3. WHILE $epoch \leq Epochs$ DO
4. FOREACH $batch \subseteq D$ DO
5. 根据式(9)~(14)提取文本特征的嵌入 \mathbf{S}'
6. 将文本特征向量化 $\mathbf{s}' = (\mathbf{V}_1^s \mathbf{S}') / L$
7. 提取用户特征的嵌入并向量化得到 \mathbf{u}'
8. 利用 LSTM 层提取时序特征向量 \mathbf{v}'
9. 根据式(17)~(19)计算圈层注意力特征 \mathbf{c}'
10. 根据式(21)~(23)计算特征组合嵌入 \mathbf{m}'
11. 由式(24)计算热度预测的结果 y'
12. 更新参数集合 $\Theta = \arg \min \frac{1}{B} \sum_{i=1}^B (y_i - y_i')^2$
13. END FOR
14. END WHILE
15. RETURN Θ

4 实验分析

本章节主要介绍实验工作,包括实验数据集、评价指标、对比模型、实验环境和结果分析等方面。

4.1 数据集

本文使用了推特(Twitter)^[22]、微博^[22]和豆瓣^[5]三个平台的数据集:(1)推特是一个全球流行的社交网络平台,用户可以发布状态,也可以与他人的信息进行互动,如点赞、评论和转发,本文只保留了信息为英文的推特文本,数据集内容的时间范围为2017年8月13日至2017年9月10日;(2)豆瓣活动是中国流行的社区网站,用户在该平台注册认证后可以发起活动,吸引感兴趣的人们参加,此平台的活动主要按城市进行划分,本文选取了上海和北京两个城市的活动信息作为两个数据集,内容的时间范围为2007年1月1日至2012年11月19日;(3)微博是中国流行的社交网络平台,用户可以发布信息,并与感兴趣的内容互动,数据集内容的时间范围为2017年8月10日至2017年12月22日。豆瓣和微博平台上的信息内容主要为中文。

在正式实验前,需要对数据集中的文本信息进行预处理操作,如分词、词干化、删除停用词、构建词汇表等。对于英文数据集,使用NLTK自然语言工具包^①可以方便地完成文本预处理;对于中文数据集,使用Jieba工具包^②分词,并利用了百度中文停用词表^③以完成预处理操作。得到对应词汇表后,即可将每条信息文本统一为向量形式输入。另外,由于需要使用用户的历史文本序列,而不同的用户历史序列长度差异可能很大,为了保证模型输入信息完整,本文只保留了数据集中至少发布10条信息的用户,并且每个用户的历史信息按照时间顺序进行排列。为了获得时间序列特征,本文将每条信息发布12小时后的热度作为最终热度,同时按1小时间隔划分时间段,最终统计得到对应的热度时间序列。

为了评估模型的预测效果,将数据集切分为训练集、验证集和测试集,前70%作为训练数据集,对于剩下的数据集,随机选取1/3作为验证数据集来确定最优参数,2/3作为测试数据集来评估预测精度。数据集的统计信息如表1所示。

表1 数据集统计信息描述

数据集	推特	微博	豆瓣上海	豆瓣北京
用户数量	12485	10621	883	1109
总样本数量	529059	410733	39906	51953
词汇表数量	45291	147942	85847	104225

4.2 评价指标

为了对热度预测模型的表现进行综合的评价,本文选取了两个回归指标和两个分类指标。对于回归指标,本文采用了回归任务中常用的均方误差(Mean Square Error, MSE)和平均绝对误差(Mean Absolute Error, MAE),根据本文的定义,对于样本 i , y_i 表示其真实热度值, y'_i 表示模型的预测热度值,则上述指标可以定义为

$$MSE = \frac{1}{N_T} \sum_{i=1}^{N_T} (y_i - y'_i)^2 \quad (26)$$

$$MAE = \frac{1}{N_T} \sum_{i=1}^{N_T} |y_i - y'_i| \quad (27)$$

其中, N_T 表示测试集的大小。另外,由于不同的信息文本的热度差别很大,直接进行训练容易导致模型很难拟合,所以,本文将原始的热度值 y_i 按如下公式进行缩放^[5]:

$$y_i = \log(r_i^n + 1) \quad (28)$$

其中, r_i^n 表示样本 i 的总转发人数或活动的总参加人数。

对于分类评价,即预测信息文本是否会变得流行,为此需要确定流行度阈值 p ,首先将数据集中的样本按热度 y_i 由高到低排序后,按照1:4的比例划分为流行信息和普通信息^[22],从而确定阈值 p 的大小,由此可以对每个样本标注0-1标签(0表示普通信息 $y_i < p$,1表示流行信息 $y_i \geq p$)。分类评价的指标为F1分数(F1-Score)和准确率(Accuracy),定义如下:

$$F1 = \frac{2TP}{2TP + FP + FN} \quad (29)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (30)$$

其中,TP是真正例(True Positive),FP是假正例(False Positive),TN是真负例(True Negative),FN是假负例(False Negative),F1分数是精确率(Precision)和召回率(Recall)的调和均值,是一种混合度量,准确率则是对模型整体性能的评估。

4.3 对比模型

为了评估SCAP模型的预测表现,将与以下模型方法进行实验对比,包括基于特征的方法(EPAB)、基于时间序列的方法(SEISMIC)、深度学习的方法(UMAN、DTCN、DFTC和Inf-VAE):

① <http://www.nltk.org/index.html>

② <https://github.com/fxsjy/jieba>

③ <https://github.com/goto456/stopwords>

(1) EPAB^[22]. 一种基于贝叶斯网络的热度预测模型,主要使用了时序特征、用户特征和网络结构特征等信息,利用贝叶斯网络进行建模;

(2) SEISMIC^[24]. 一种基于自激点过程理论的时序模型,采用了双重随机过程刻画信息的传染性与转发事件的到达时间,以预测信息最终的转发数量;

(3) UMAN^[53]. 一种基于循环神经网络与注意力机制的热度预测模型,主要利用了用户的历史序列信息和对应的热度值,融合长期特征和短期波动情况,进行用户生成内容的热度预测;

(4) DTCN^[39]. 一种基于时序特征和时序注意力机制的深度学习模型,以预测社交媒体中图片的热度,本文利用用户的历史序列等内容,将其扩展应用于文本内容热度的预测;

(5) DFTC^[3]. 一种深度融合时间过程和内容特征的方法,结合卷积神经网络和循环神经网络建模文章热度变化的时间过程,并利用层次注意网络和特征嵌入技术预测网络文章的热度;

(6) Inf-VAE^[37]. 一种基于图神经网络与变分自编码器的传播预测模型,利用注意力机制和好友关系对社会同质性和时序影响共同建模,在给定种子用户的情况下,预测受影响的用户集合,本文将其应用于用户的转发行为预测.

为了保证实验的客观公正,对于以上对比模型,均采用与本文 SCAP 模型同样的数据预处理操作,对于 UMAN 和 DTCN 模型,采用相同的 LSTM 层处理用户历史序列,在预测信息热度时,结合文本特征、用户特征和时序特征协同进行预测工作.

4.4 实验环境

本文将数据集划分为训练集、验证集和测试集,其中验证集的作用即确定最优的模型参数.在本文中,文本序列的最大长度 $L=50$,文本嵌入向量维度 $K_S=128$,用户嵌入向量维度 $K_U=128$,社交圈层向量维度 $K_C=128$,社交圈层模型所用的用户历史序列文本数量 $M=6$,圈层聚类个数 $k=8$,热度时间序列长度 $t=3$.

实验机器配置为 Ubuntu 18.04 LTS,CPU Intel Xeon(2.20GHz),内存 12GB,显卡为 NVIDIA Tesla P100 16GB.本文采用 Keras 机器学习框架实现 SCAP 模型,实验环境版本 Python 3.6.9,Keras 2.3.1, Tensorflow 1.15.2, Numpy 1.18.2, Sklearn 0.22.2,模型训练采用自适应矩估计 Adam(Adaptive Moment Estimation)优化器^[48],学习率默认为 0.001,训练批(Batch)大小为 128.

4.5 结果分析

本文实验的目的是验证基于社交圈层和注意力机制的热度预测模型 SCAP 的有效性,实验验证主要包括如下内容:(1)利用 MSE 和 MAE 指标对模型进行回归评价;(2)利用 F1 分数和准确率指标对模型进行分类评价;(3)分析模型参数对于预测的影响;(4)验证模型各特征和机制的引入对于模型预测性能提升的有效性;(5)可视化社交圈层注意力对于模型的影响.

4.5.1 热度预测的回归实验

本文首先进行了热度预测的回归分析实验,实验指标为 MSE 和 MAE,实验结果如表 2 所示,其中粗体表示最优结果,加下划线表示次优结果.

表 2 在不同数据集下各模型的实验结果

	推特				微博				豆瓣上海				豆瓣北京			
	MSE	MAE	F1	Acc												
EPAB	0.812	0.664	0.881	0.846	0.833	0.770	0.761	0.850	0.815	<u>0.626</u>	0.846	0.870	0.815	0.686	0.839	0.884
SEISMIC	1.134	0.879	0.762	0.831	1.196	0.903	0.752	0.787	1.092	0.906	0.830	0.738	1.075	0.895	0.787	0.722
UMAN	<u>0.723</u>	0.648	<u>0.890</u>	0.917	<u>0.737</u>	<u>0.682</u>	<u>0.880</u>	<u>0.912</u>	<u>0.742</u>	0.628	<u>0.861</u>	<u>0.887</u>	<u>0.761</u>	<u>0.652</u>	<u>0.874</u>	<u>0.902</u>
DTCN	0.840	0.741	0.745	0.868	0.892	0.804	0.796	0.824	0.929	0.776	0.842	0.780	0.893	0.739	0.794	0.751
DFTC	0.745	<u>0.629</u>	0.863	<u>0.922</u>	0.759	0.713	0.854	0.885	0.734	0.638	0.839	0.871	0.792	0.688	0.845	0.880
Inf-VAE	0.747	0.650	0.849	0.861	0.793	0.732	0.846	0.877	0.786	0.691	0.840	0.862	0.805	0.707	0.858	0.882
SCAP	0.706	0.602	0.924	0.929	0.715	0.644	0.901	0.928	0.721	0.610	0.895	0.913	0.730	0.616	0.899	0.923

(1)不论是基于 MSE 指标还是 MAE 指标,本文提出的 SCAP 模型在四个数据集上的实验结果均优于对比模型,取得了最好的结果.例如,在推特、微博、豆瓣上海和豆瓣北京数据集上,SCAP 模型的 MSE 指标,相比于对比模型中表现最好的 UMAN 模型,分别降低了 0.017,0.022,0.021 和 0.031;

(2)实验中表现最好的 SCAP 模型与 UMAN 模型均利用了用户的历史文本序列进行分析,这说明了用户历史信息特征对于热度预测是非常有效的. DFTC 模型融合了时序信息和内容特征,在推特数据集上取得了较好的表现,而 SCAP 模型进一步考虑了不同社交圈层对于信息的影响,取得了更好

的效果;

(3) EPAB 模型考虑了不同种类的特征,包括时序特征、用户特征和网络结构特征,利用贝叶斯网络建模,来对热度进行预测,取得了较好的结果,MAE 指标在豆瓣上海数据中为 0.626,但是由于其贝叶斯网络的结构未知,需要从特征中学习,这限制了模型的预测表现;

(4) DTCN 模型表现较差的原因是,其将所有用户的文本序列信息作为整体的序列进行输入,没有考虑到不同用户的历史序列反映了不同的偏好特征,影响了实验结果;

(5) SEISMIC 模型需要时序信息的输入来对信息的演变过程进行建模,同时做了一些参数假设,由于一条文本内容发布后,充足的时间序列信息需要较长的观测周期才能获取,同时仅利用时序信息会限制模型对于信息传播热度的预测结果;

(6) Inf-VAE 模型考虑了用户社交同质性关系和时序影响,以进行用户转发预测.然而其是面向用户层面的预测模型,没有考虑文本和用户的相关特征,限制了该模型的表现.同时需要利用图神经网络对社交网络的结构进行分析,也会对模型的运行时间产生一定的影响.

4.5.2 热度预测的分类实验

本文将热度预测作为分类问题,对各模型进行了评估,实验指标为 $F1$ 分数和准确率,实验结果如表 2 所示.

(1) 相比于回归实验,分类实验的结果差别不大.在 $F1$ 分数和准确率指标上,SCAP 模型在四个

数据集上的表现依然非常优异,均取得了最好成绩,相比于对比模型中表现最好的 UMAN 模型, $F1$ 分数分别提升了 0.034,0.021,0.034 和 0.025,准确率分别提升了 0.012,0.016,0.026 和 0.021;

(2) 值得注意的是,各个模型在不同数据集上的表现是不同的.总体上来看,各模型在推特数据集上的表现略优于其它三个数据集,造成这种情况的原因可能是推特数据集中的用户数量和样本数据更多,同时推特中信息的流行程度更高,信息的传播速度更快,提取到的特征相对更为显著有效,实验结果有一定的差异.

4.5.3 SCAP 模型参数分析

在本文提出的 SCAP 模型中,用户的历史文本序列数量、社交圈层的聚类个数和热度时间序列长度是可以动态调整的参数,为了评估各参数对模型预测性能的影响,进行了如下实验.

(1) 用户历史文本序列数量的影响

本文选取了 UMAN 模型作为对比模型,因为 UMAN 模型在回归和分类实验中其表现较好,同时可以处理序列数据.与本文 SCAP 模型不同,UMAN 模型着重挖掘每个用户的历史序列文本之间的影响与联系,包括内容序列的相关性和热度序列的相关性,即同一用户发布的两条信息之间存在的内容和热度差异.而 SCAP 模型更侧重于挖掘用户历史文本序列中的偏好特征,并且考虑不同的社交圈层对于新发布内容的影响程度,并结合文本特征、用户特征和时序特征协同进行预测.实验的评价指标为 MSE 和 $F1$ 分数,实验结果如图 2 和图 3 所示.

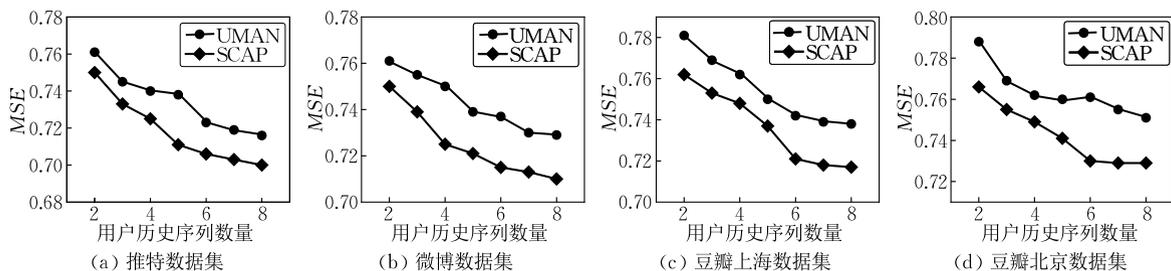


图 2 模型在不同用户历史序列数量下的 MSE 变化情况

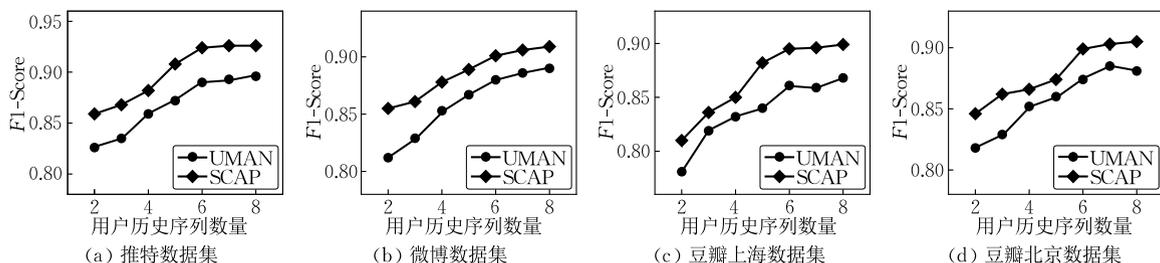


图 3 模型在不同用户历史序列数量下的 $F1$ 分数变化情况

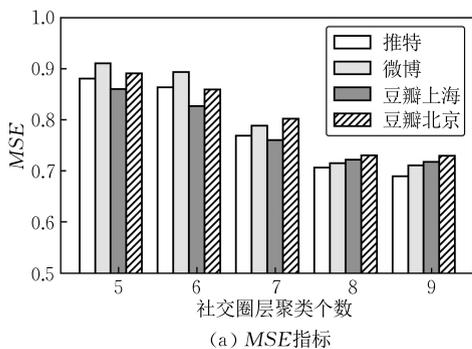
① 不论是 MSE 还是 $F1$ 分数指标,模型的预测表现均随着用户历史序列数量的增加而提升,这是因为更多的用户历史序列可以帮助模型提取到更为有效的用户特征,促进社交圈层的划分效果,从而提升实验结果;

② 随着用户历史序列数量的增加,模型表现的提升逐渐放缓,这是由于用户历史序列越多,各序列间的相似度就越高,产生了冗余信息,对特征提取的帮助作用相对减小;

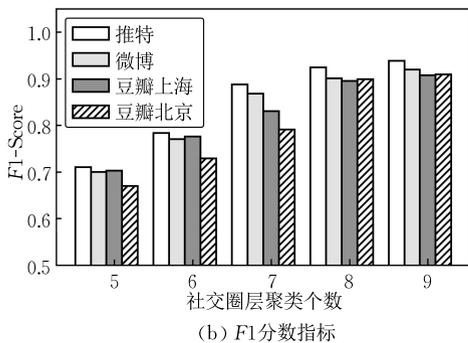
③ 值得说明的是,更多的用户历史序列会带来模型复杂度的增加,因为需要更多的 LSTM 单元去处理序列信息,而循环神经网络的训练相对比较慢,每个时间步要依赖于前一个时间步的信息,导致模型的运算时间变长,所以在实际应用中,按照具体的需求,应根据模型的预测表现和运行时间做出权衡.

(2) 社交圈层聚类个数的影响

图 4 展示了 SCAP 模型在不同社交圈层聚类个数下的表现. 与针对用户历史序列数量的实验结果类似,在 MSE 和 $F1$ 分数评价指标下,SCAP 模型的表现随着社交圈层聚类个数的增加而提升,这是因为模型可以为新发布的文本信息捕获到更多的相似社交圈层特征. 同时由于更多的社交圈层之间偏好特征可能有所重叠交叉,模型表现的提升幅度也会逐渐减小.



(a) MSE 指标



(b) $F1$ 分数指标

图 4 SCAP 模型在不同社交圈层聚类个数下的表现

(3) 热度时间序列长度的影响

对于热度时间序列长度的影响实验,结果如图 5 所示. SCAP 模型的预测表现 (MSE 指标) 随时间序列长度的增加而提升,这是因为随着观测周期的增长,模型可以提取到足够且有效的时序特征. 同时,模型表现的提升幅度逐渐变小,原因是在社交网络中,文本信息转发数量的增加往往随着时间的增长而放缓,比如一个用户发布新的社交动态后,短期内会收到大量的互动信息,但是在后期关注度会越来越低,其转发数量变化趋于稳定,时序特征带来的帮助减小.

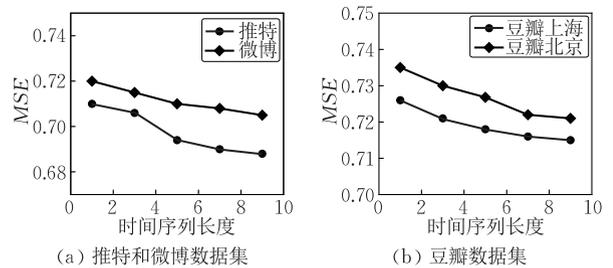


图 5 SCAP 模型 MSE 指标随时间序列长度的变化

4.5.4 模型简化测试

为了验证 SCAP 模型各特征和注意力机制的有效性,本文分别构建了以下模型:(1) SCAP-NC 模型表示不考虑社交圈层特征;(2) SCAP-NS 模型表示去除文本特征;(3) SCAP-NU 模型表示去除用户特征;(4) SCAP-NT 模型表示去除时序特征;(5) SCAP-NA 模型表示去除圈层注意力机制. 以上模型均利用其余特征的融合向量作为输入. 实验的结果如表 3 所示.

表 3 模型简化测试的实验结果(采用 MSE 指标)

模型	推特	微博	豆瓣上海	豆瓣北京
SCAP-NC	0.771	0.782	0.778	0.796
SCAP-NS	0.719	0.723	0.734	0.745
SCAP-NU	0.712	0.719	0.728	0.741
SCAP-NT	0.720	0.727	0.732	0.739
SCAP-NA	0.725	0.731	0.734	0.756
SCAP	0.706	0.715	0.721	0.730

首先,社交圈层特征对于模型预测性能的提升是非常明显的,相比于 SCAP-NC 模型,在不同数据集中,SCAP 的 MSE 指标分别降低了 0.065, 0.067, 0.057 和 0.066;而相比于 UMAN 模型,SCAP-NC 模型的 MSE 指标分别提高了 0.048, 0.045, 0.036 和 0.035,这说明在热度预测任务中,对用户的历史序列进行挖掘,考虑社交圈层特征可

以获得更加精准的预测结果. 其次, 文本特征、用户特征和时序特征均为模型表现的提升带来了一定的帮助. 另外, 注意力机制的引入减小了预测的误差, 相比于 SCAP-NA 模型, SCAP 模型在不同数据集中 MSE 分别降低了 0.019, 0.016, 0.013 和 0.026, 表明利用注意力机制, 可以提取到不同社交圈层的影响大小, 从而验证了模型各特征和机制的有效性.

4.5.5 社交圈层注意力机制分析

为了更加直观地展示社交圈层注意力机制在模型预测中发挥的作用, 本文从推特和豆瓣北京两个数据集中选取了两条文本信息, 对其进行文本预处理后, 将模型计算得到的圈层注意力权重可视化, 字词颜色越深表示模型的关注度越高, 实验结果如图 6 所示. 以第二条文本为例, 模型对于“管弦乐团”、“国家大剧院”和“音乐会”赋予了更高的权重, 能够有效地反映出其领域和特征偏好信息, 验证了本文提出的社交圈层注意力的有效性.

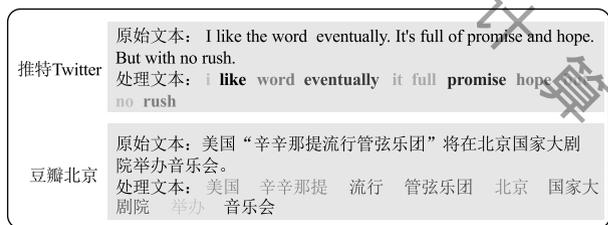


图 6 社交圈层注意力机制可视化

5 总结

本文提出了一种基于社交圈层和注意力机制的信息热度预测模型 SCAP, 该模型首先利用自动编码器对用户发布的历史文本序列进行特征学习, 得到用户偏好的特征表示, 然后将其通过 K-Means++ 聚类算法划分为不同的社交圈层, 得到社交圈层的特征向量; 其次, 利用 LSTM 长短期记忆网络和嵌入层提取社交信息的文本特征、用户特征和时序特征; 进而通过注意力机制, 捕获到不同社交圈层对于该文本信息的影响程度, 得到社交圈层注意力特征; 最后将文本特征、用户特征、时序特征和社交圈层注意力特征进行融合, 通过两个全连接层进行建模学习, 对社交信息的热度进行预测.

实验结果表明, SCAP 模型在推特、微博、豆瓣多个数据集上的预测表现相比于对比模型总体呈优, 能够较为准确地预测社交信息的热度. 实验同时探究了不同实验参数对于模型的影响效果, 以及验

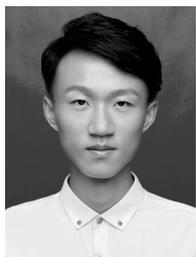
证了输入特征、社交圈层机制和注意力机制对于模型性能提升的有效性, 说明了用户的历史行为和序列信息对于社交信息传播影响的重要性. 因此, 在后续的研究中, 计划对不同用户间的行为关联和差异进行更进一步的研究.

参 考 文 献

- [1] Fedoryszak M, Frederick B, Rajaram V, et al. Real-time event detection on social data streams//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Anchorage, USA, 2019: 2774-2782
- [2] Chen G, Xu N, Mao W. An encoder-memory-decoder framework for sub-event detection in social media//Proceedings of the 27th ACM International Conference on Information and Knowledge Management. Turin, Italy, 2018: 1575-1578
- [3] Liao D, Xu J, Li G, et al. Popularity prediction on online articles with deep fusion of temporal process and content features//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Hawaii, USA, 2019: 200-207
- [4] Mu L, Jin P, Zheng L, et al. Lifecycle-based event detection from microblogs//Proceedings of the 27th International World Wide Web Conference. Lyon, France, 2018: 283-290
- [5] Wang W, Zhang W, Wang J, et al. Learning sequential correlation for user generated textual content popularity prediction//Proceedings of the 26th International Joint Conferences on Artificial Intelligence. Stockholm, Sweden, 2018: 1625-1631
- [6] Wu H, Hu Z, Jia J, et al. Mining unfollow behavior in large-scale online social networks via spatial-temporal interaction//Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020: 254-261
- [7] Myers D. Social Psychology. The 8th Edition. Beijing: Posts & Telecom Press, 2006
- [8] Gao X, Cao Z, Li S, et al. Taxonomy and evaluation for microblog popularity prediction. ACM Transactions on Knowledge Discovery from Data, 2019, 13(2): 1-40
- [9] Cheng J, Adamic L, Dow P A, et al. Can cascades be predicted?//Proceedings of the 23rd International World Wide Web Conference. Seoul, Korea, 2014: 925-936
- [10] Jenders M, Kasneci G, Naumann F. Analyzing and predicting viral tweets//Proceedings of the 22nd International World Wide Web Conference. Rio de Janeiro, Brazil, 2013: 657-664
- [11] Liao Xiang-Wen, Zheng Hou-Dong, Liu Sheng-Hua, et al. Learning influences and susceptibilities for sentiments from users' behaviors. Chinese Journal of Computers, 2017, 40(4): 955-969(in Chinese)

- (廖祥文, 郑候东, 刘盛华等. 基于用户行为的情感影响力和易感性学习. 计算机学报, 2017, 40(4): 955-969)
- [12] Naveed N, Gottron T, Kunegis J, et al. Bad news travel fast: A content-based analysis of interestingness on Twitter// Proceedings of the 3rd International Web Science Conference. Southampton, UK, 2011: 1-7
- [13] Huang Fa-Liang, Feng Shi, Wang Da-Ling, et al. Mining topic sentiment in microblogging based on multi-feature fusion. Chinese Journal of Computers, 2017, 40(4): 872-888(in Chinese)
(黄发良, 冯时, 王大玲等. 基于多特征融合的微博主题情感挖掘. 计算机学报, 2017, 40(4): 872-888)
- [14] Cha M, Haddadi H, Benevenuto F, et al. Measuring user influence in Twitter: The million follower fallacy//Proceedings of the 4th International AAAI Conference on Weblogs and Social Media. Washington, USA, 2010: 10-17
- [15] Weng J, Lim E P, Jiang J, et al. TwitterRank: Finding topic-sensitive influential twitterers//Proceedings of the 3rd ACM International Conference on Web Search and Data Mining. New York, USA, 2010: 261-270
- [16] Yuan N J, Zhong Y, Zhang F, et al. Who will reply to/retweet this tweet? The dynamics of intimacy from online social interactions//Proceedings of the 9th ACM International Conference on Web Search and Data Mining. San Francisco, USA, 2016: 3-12
- [17] Hoang B T, Chelghoum K, Kacem I. Modeling information diffusion via reputation estimation//Proceedings of the 27th International Conference on Database and Expert Systems Applications. Porto, Portugal, 2016: 136-150
- [18] Zhang J, Liu B, Tang J, et al. Social influence locality for modeling retweeting behaviors//Proceedings of the 23rd International Joint Conference on Artificial Intelligence. Beijing, China, 2013: 2761-2767
- [19] Liu Wei, He Min, Wang Li-Hong, et al. Research on microblog retweeting prediction based on user behavior features. Chinese Journal of Computers, 2016, 39(10): 1992-2006(in Chinese)
(刘玮, 贺敏, 王丽宏等. 基于用户行为特征的微博转发预测研究. 计算机学报, 2016, 39(10): 1992-2006)
- [20] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks//Proceedings of the 19th ACM International Conference on Information and Knowledge Management. Toronto, Canada, 2010: 1633-1636
- [21] Cha M, Mislove A, Gummadi K P. A measurement-driven analysis of information propagation in the Flickr social network//Proceedings of the 18th International World Wide Web Conference. Madrid, Spain, 2009: 721-730
- [22] Wu Q, Yang C, Gao X, et al. EPAB: Early pattern aware Bayesian model for social content popularity prediction// Proceedings of the 18th IEEE International Conference on Data Mining. Singapore, 2018: 1296-1301
- [23] Shen H, Wang D, Song C, et al. Modeling and predicting popularity dynamics via reinforced Poisson processes// Proceedings of the 28th AAAI Conference on Artificial Intelligence. Québec City, Canada, 2014: 291-297
- [24] Zhao Q, Erdogdu M A, He H Y, et al. SEISMIC: A self-exciting point process model for predicting tweet popularity// Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Sydney, Australia, 2015: 1513-1522
- [25] Yu L, Cui P, Wang F, et al. From micro to macro: Uncovering and predicting information cascading process with behavioral dynamics//Proceedings of the 15th IEEE International Conference on Data Mining. Atlantic City, USA, 2015: 559-568
- [26] Jiang M, Cui P, Liu R, et al. Social contextual recommendation //Proceedings of the 21st ACM International Conference on Information and Knowledge Management. Maui Hawaii, USA, 2012: 45-54
- [27] Li J, Qin J, Wang T, et al. A collaborative filtering model for personalized retweeting prediction//Proceedings of the 20th International Conference on Database Systems for Advanced Applications. Hanoi, Vietnam, 2015: 122-134
- [28] Pan R, Zhou Y, Cao B, et al. One-class collaborative filtering//Proceedings of the 8th IEEE International Conference on Data Mining. Pisa, Italy, 2008: 502-511
- [29] Ye J, Wang Z, Zhu W. From global to local: A context-embedded LSTM recurrent network for local content popularity prediction//Proceedings of the 7th International Conference on AI and Mobile Services. San Francisco, USA, 2018: 29-42
- [30] Li C, Ma J, Guo X, et al. DeepCas: An end-to-end predictor of information cascades//Proceedings of the 26th International World Wide Web Conference. Perth, Australia, 2017: 577-586
- [31] Cao Q, Shen H, Cen K, et al. DeepHawkes: Bridging the gap between prediction and understanding of information cascades//Proceedings of the 26th ACM International Conference on Information and Knowledge Management. Singapore, 2017: 1149-1158
- [32] Lamprinidis S, Hardt D, Hovy D. Predicting news headline popularity with syntactic and semantic knowledge using multi-task learning//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 659-664
- [33] Du N, Dai H, Trivedi R, et al. Recurrent marked temporal point processes: Embedding event history to vector//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1555-1564
- [34] Xiao S, Yan J, Yang X, et al. Modeling the intensity function of point process via recurrent neural networks//Proceedings of

- the 31st AAAI Conference on Artificial Intelligence. San Francisco, USA, 2017; 1597-1603
- [35] Wang Y, Liu S, Shen H, et al. Marked temporal dynamics modeling based on recurrent neural network//Proceedings of the 21st Pacific-Asia Conference on Knowledge Discovery and Data Mining. Jeju, Korea, 2017; 786-798
- [36] Cao Q, Shen H, Gao J, et al. Popularity prediction on social platforms with coupled graph neural networks//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, USA, 2020; 70-78
- [37] Sankar A, Zhang X, Krishnan A, et al. Inf-VAE: A variational autoencoder framework to integrate homophily and influence in diffusion prediction//Proceedings of the 13th International Conference on Web Search and Data Mining. Houston, USA, 2020; 510-518
- [38] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015
- [39] Wu B, Cheng W H, Zhang Y, et al. Sequential prediction of social media popularity with deep temporal context networks//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017; 3062-3068
- [40] Chen G, Kong Q, Mao W. An attention-based neural popularity prediction model for social media events//Proceedings of the 15th IEEE International Conference on Intelligence and Security Informatics. Beijing, China, 2017; 161-163
- [41] Zhang W, Wang W, Wang J, et al. User-guided hierarchical attention network for multi-modal social image popularity prediction//Proceedings of the 27th International World Wide Web Conference. Lyon, France, 2018; 1277-1286
- [42] Song Y Z, Shuai H H, Yeh S L, et al. Attractive or faithful? Popularity-reinforced learning for inspired headline generation //Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 2020; 8910-8917
- [43] Kim Y. Convolutional neural networks for sentence classification //Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar, 2014; 1746-1751
- [44] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Computation*, 1997, 9(8): 1735-1780
- [45] Tai K S, Socher R, Manning C D. Improved semantic representations from tree-structured long short-term memory networks//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics. Beijing, China, 2015; 1556-1566
- [46] Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using LSTMs//Proceedings of the 32nd International Conference on Machine Learning. Lille, France, 2015; 843-852
- [47] Xie J, Girshick R, Farhadi A. Unsupervised deep embedding for clustering analysis//Proceedings of the 33rd International Conference on Machine Learning. New York, USA, 2016; 478-487
- [48] Kingma D P, Ba J. Adam: A method for stochastic optimization //Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015



ZHENG Zuo-Wu, Ph. D. candidate. His research interests include social networks analysis and prediction of information popularity.

SHAO Si-Qi, undergraduate. Her research interests include social networks and natural language processing.

GAO Xiao-Feng, Ph. D., professor. Her research interests include data engineering, database index, and network optimization.

CHEN Gui-Hai, Ph. D., professor. His research interests include computer networks, distributed computing, and wireless networks.

Background

Popularity prediction of user generated content (UGC) in social networks is a hot topic in recent years, which has a wide range of applications in scenarios such as marketing, public opinion detection, and advertising recommendation. In general, there are two main categories of methods in this field, including feature based methods and time series methods, from which it has been extensively studied, such as constructing hand-craft features, utilizing stochastic process

to model temporal sequence of UGC. However, studies on the impact of social circles is missing. In this work, we tend to fill this gap by considering social circle features and the corresponding impact for UGC to predict the popularity. It gives a new perspective to model the relations between user and content, which is important in this filed since we can make prediction more accurate if we know the correlations of them.

In this paper, we propose a social circle and attention based popularity prediction model (SCAP). First of all, from user sequential behavior patterns, we extract user preference by a LSTM-based autoencoder, from which social circle embedding can be learned by clustering method. Then for a new UGC, text embedding, user embedding and temporal embedding are extracted based on LSTM layer and embedding layer. Furthermore, we capture the influence on UGC of different social circles through attention mechanism. Finally, using social circle attention, text embedding, user embedding, and temporal embedding, the popularity of UGC is predicted by two fully connected layer. We conduct extensive experiments on four real world datasets from Twitter, Weibo, and Douban Event platforms. The experimental results show that

the performance of SCAP is better than baseline methods with different metrics, such as mean square error (*MSE*) and *F1*-score. We also consider the impacts of different parameters on the model, including the number of users' recent sequences, the number of social circles, and the length of temporal sequences. In addition, the effectiveness of the social circle and attention mechanism are validated in ablation study.

This work was supported by the National Key R&D Program of China (2020YFB1707903), the National Natural Science Foundation of China (61872238, 61972254), the CCF-Tencent Open Fund (RAGR20200105), the Tencent Marketing Solution Rhino-Bird Focused Research Program (FR202001), and the Huawei Cloud (TC20201127009).