DOI:10.11784/tdxbz202005079

基于在线社交网络事件库多因素耦合的流行度预测方法

于 海[1]，吕晴晴[2]，时 鹏[3]，王 铮[1]，胡长军[1]
(1. 北京科技大学计算机与通信工程学院，北京 100083；2. 中国电子科技集团公司第十五研究所，北京 100083；
3. 北京科技大学国家材料服役安全科学中心，北京 100083)

摘 要：随着近些年新一代信息网络技术的高速发展，人们可以通过社交网络平台更快、更广地获知社会上的各类事件，这使得事件在社交网络中的传播逐步呈现出高速化、扩大化的特点. 针对这种情况，为了更好地管理社交网络中的事件，提高对网络事件信息的治理水平，有必要对社交网络信息传播进行分析. 流行度预测是在线社交网络事件信息传播分析中的研究重点. 对事件流行度的预测能够为网络事件发生、发展、高峰、终结等提供深刻的见解. 尽管流行度预测已经被广泛研究，但是事件相关信息与流行度相关联的因素缺少即时可用的指标数据、指标数据差异化等问题使得有效地预测事件流行度仍然没有得到较好的解决. 有鉴于此，本文设计了一种基于在线社交网络事件库多因素耦合的流行度预测方法. 具体来说，首先提出了一种基于社交网络事件库的多因素指标获取方法，利用事件库对于社交网络数据的统一存储，从多源异构数据中提取各因素指标. 其次提出了一种多因素耦合的流行度预测方法，通过分组嵌入得到因素指标的可相互结合的低维表示，在预测中实现多因素指标的综合利用. 最后将 Twitter7 数据集中 3 000 个主题标签包含的推文作为实验对象进行平均准确率计算. 实验结果表明：与已有的深层神经网络模型(DNN)、支持向量回归模型(SVR)、SH 流行度预测模型等相比，本研究所提出预测方法在预测准确度上具有明显的优越性.

关键词：流行度预测；多因素耦合；累积性因素；固有性因素
中图分类号：TP311.5 **文献标志码**：A **文章编号**：0493-2137(2020) 12-1272-09

Popularity Prediction Method Based on the Multi-Factor Coupling of the Online Social Network Event Base

Yu Hai[1]，Lü Qingqing[2]，Shi Peng[3]，Wang Zheng[1]，Hu Changjun[1]
(1. School of Computer & Communication Engineering，University of Science and Technology Beijing，Beijing 100083，China；
2. China Electronics Technology Group Corporation 15th Research Institute，Beijing 100083，China；
3. National Center for Materials Service Safety，University of Science and Technology Beijing，Beijing 100083，China)

Abstract：With the rapid development of the new generation of information network technology in recent years，people become aware of all kinds of social events rapidly and extensively through the social network platform，which speeds up and expands the diffusion of events in the social network. Because of this situation，to manage the events in the social network and improve the governance level of network event information effectively，information diffusion in the social network needs to be analyzed. Popularity prediction is the focus of online social network event information diffusion analysis. Popularity prediction can provide profound insights into the occurrence，development，peak，and end of network events. Although popularity prediction has been widely investigated，the lack of instant and available indicator data for factors associated with event-related information and popularity and the difficulty in differentiating indicator data hinder the effective prediction of event popularity. For this reason，this study designs a

收稿日期：2020-05-28；**修回日期**：2020-06-29.
作者简介：于 海(1981—)，男，博士研究生，高级工程师，yuhaionline@163.com.
通信作者：王 铮，wangzheng@ustb.edu.cn.
基金项目：国家重点研发计划资助项目(2017YFB0803302).
Supported by the National Key Research and Development Program of China(No. 2017YFB0803302).

popularity prediction method based on the multi-factor coupling of the online social network event base. First，a multi-factor indicator acquisition method based on the social network event database，which uses the event database to uniformly store social network data and extract each factor indicator from multi-source heterogeneous data，is proposed. Second，a multi-factor coupling method for popularity prediction is proposed，with the low-dimensional representation of factor indicators that can be combined is obtained by grouping and embedding to realize the comprehensive utilization of multi-factor indicators during prediction. Finally，the tweets contained in 3 000 subject tags in the Twitter 7 data set are utilized as the experimental subjects to calculate the average accuracy. The experimental results show that，compared with the existing deep neural network，support vector regression，and SH popularity prediction models，the prediction method proposed in this study has superior prediction accuracy.

Keywords：popularity prediction；multi-factor coupling；cumulative factor；inherent factor

在线社交网络是人们参与、传播和讨论事件的重要载体[1-2]. 通过微信、微博、Facebook、Twitter 等线上社交平台，人们可以随时随地分享事件信息. 这使得社交网络信息传播的方式发生了重大变化. 人们从过去的被动接收信息逐步转向为主动发布信息. 事件信息通过人们之间的关系进行传播扩散. 为了有效管理社交网络中的事件，治理事件相关信息，感知事件发展趋势，有必要对社交网络信息传播进行分析[3-4]. 在面向在线社交网络的事件信息传播分析中，流行度演化分析至关重要[5]. 例如，在本次 Covid-19 疫情中，借助流行度演化分析能够及时发现热点事件并做出应急处理，能够快速掌握舆情发展态势并进行形势研判. 同时，还能够科学评估官方通告、防控知识等可靠信息和谣言、误传新闻等虚假信息的传播情况等. 流行度演化分析涉及演化模式、影响因素、预测等多个维度的问题. 其中，对未来流行度预测是目前流行度演化分析的热点研究主题，它是基于在线社交网络事件发展进行决策的基础，受到了各国学者的高度关注.

1　流行度预测

流行度可以理解为某条网络信息在给定时刻的被关注程度. 流行度预测是对在线社交网络信息发展趋势的测定. 流行度预测的主要目标是预测在线社交网络事件相关信息未来的流行度指标[6]. 目前，大多数流行度预测都是利用截至当前时刻为止的单一因素指标进行预测. Szabo 等[7]基于 YouTube 视频和 Digg 帖子构建了经典的 SH 模型，该模型通过对前期和后期流行度作对数处理发现了历史流行度之间的强相关性，并以此预测未来流行度. Bao 等[8]对社交网络结构与流行度相关关系进行了研究，进而扩展了 SH 模型进行流行度预测. 然而，事件相关信息的流行度在随时间演化的过程中会受到各种信息传播驱动因素的影响，如信息自身吸引力、用户兴趣、用户活跃度等，并且其中部分因素本身也会随着时间动态变化. 这导致基于单一因素指标的预测方法效果不佳，越来越多的研究者将注意力转移到对于多因素的利用上. He 等[9]基于用户评论的双向图表示设计了一个预测算法，该算法利用时序因素和社交影响力因素对事件信息的影响进行流行度预测. Kong 等[10]基于机器学习方法，利用 Twitter Hashtag 中的内容、网络结构、社交等因素进行流行度预测. 虽然基于多因素指标的预测能够在一定程度上摆脱单因素指标所带来的局限，但是目前仍有两方面因素制约着社交网络事件信息流行度预测中对于多因素的利用.

首先，事件相关信息的与流行度相关联的因素常隐含在社交网络大数据中，缺少即时可用的指标数据，并且难以直接通过简单统计得到时间序列形式的指标数据，那么对于多因素来说如何得到量化、可用的指标数据成为了一个主要挑战. 其次，即使有了多因素指标数据，由于其含义、范围、时域分布特征等方面存在差异，如何对其进行综合利用以实现对流行度的精准预测成为了另一个主要挑战.

针对上述挑战，本研究提出了一种基于事件库多因素耦合的流行度预测方法. 社交网络事件库是一种面向社交网络事件分析的统一数据平台，为获取各种流行度演化相关联的指标提供了便利，本研究提出了基于事件库获取多因素指标的方法. 基于深度学习的嵌入方法为时间序列数据的降维与融合提供了可能. 为了实现多种因素指标的综合利用，本研究提出了一种分组嵌入的方法，根据因素的物理意义与特征对其进行分组，然后分别采用不同的神经网络进行嵌入得到这些因素指标的地位表示，并在此基础上提出了预测方法. 与前述多因素预测方法[9-10]相比，本研究没有局限于影响因素的划分，并且重点对多因素指标数据进行了分析. 所提出的方法在因素选取、因素抽象化定义、因素分组、因素指标数据获取和指标数据综合利用等方面实现了创新设计. 实验表明本研究提出的方法相比现有模型具有更高的准确性.

2 预测方法

为了在事件流行度预测中能够利用多种多样的因素,首先需要对事件流行度因素进行分析,给出了这些因素的抽象表示,为获取多因素量化指标及构建预测模型奠定基础. 之后,在多因素指标中,以在线社交网络事件库为基础给出了多因素指标获取方法,利用了在线社交网络事件库作为统一数据平台对于附有时间属性的原始数据的全面存储,从中按需抽取多各种指标. 在此基础上,基于深度学习中的嵌入表示思想,提出了多因素分组嵌入方法,得到多因素指标的低维表示,从而能够在流行度预测中将这些因素耦合起来. 最后,给出了基于深度神经网络的流行度预测方法.

2.1 符号与问题定义

在深入研究所提出的预测方法之前,本文首先介绍数学符号,所涉及到的数学符号如表 1 所示.

表 1 符号表

Tab.1 Symbol table

符号	定义	符号	定义
y	流行度演化时间序列	i	时间步
n	个数	g	传播结果子图变化序列
G	传播结果子图	u	参与用户影响力变化序列
U	参与用户影响力	c	事件话题语义
h	事件代表性帖子集合	user	用户
t	参与讨论时刻	τ	某一时刻
V	节点集合	E	边集合
θ	嵌入表示	d	嵌入表示的维数
ω	窗口大小	N	每个节点游走的次数
L	每个节点游走的步数	ϕ	传播子图嵌入表示
R	实数	$\boldsymbol{\Phi}$	初始化矩阵
T	霍夫曼树	O	节点乱序洗牌
W	随机游走算法	μ	参与用户影响力因素表示
$\boldsymbol{\Theta}$	分组表示向量	$\boldsymbol{M}$	向量表示矩阵
p	卷积核数量	S	卷积层输出
I	指示函数	δ	调节因子
A	流行度预测平均准确率		

接下来介绍本文的问题定义. 即基于给定的历史流行度和附加的观测信息来预测在线社交网络未来某时刻事件的流行度.

2.2 事件流行度因素分析

在线社交网络上的流行度反映了用户参与事件讨论的热烈程度,同时也反映了事件相关信息在网络上的影响范围. 事件的流行度可以用单位时间内评论数、帖子(推文/微博)数、转发数或点赞互动数来定义,是一个增量. 无论采用哪种具体量,其随时间相对变化趋势是基本一致的,在本研究中采用最常用的帖子数作为流行度指标. 为便于研究,通常将流行度随时间的演化表示为通过固定时间间隔采样得到的时间序列 $y=\{y(1),y(2),\cdots,y(n)\}$,其中 $y(i)$ ($i=1,2,\cdots,n$) 表示第 i 时间步的流行度. 对于流行度的预测即预测某一时刻的流行度值,通常是在给定历史流行度时间序列的基础上,对未来某一时间步进行预测. 历史流行度是一种最常用的作为预测依据的指标.

除此以外,从驱动信息传播的角度,与流行度演化密切相关的因素还有很多. 按照因素的作用来源又可分为结构性因素、内容性因素和影响力因素. 其中,结构性因素指的是与传播网络、传播级联结构相关的因素,内容性因素则是与事件相关信息的内容相关的因素,影响力因素指的是反映参与信息传播用户的影响力的因素. 按照这些因素是否伴随流行度一同演化,可分为累积性因素和固有性因素. 累积性因素随时间推移会产生变化,固有性因素则反映了内在属性不易随事件变化. 下面给出 4 种能够用于预测的代表性的因素及其抽象化定义. 对于预测中使用的具体指标表示,则在第 2.4 节中给出了如何通过嵌入方法得到.

1) 传播结果子图

传播结果子图指的是事件相关信息在社交网络上形成的子图快照. 传播结果子图反映着信息在网络上传播情况,能够用于流行度预测. 例如,当子图中连接到“外部”的边变多时,则很可能引起信息广泛传播到子图外,从而引起流行度上升. 随着时间推移和信息的传播,传播结果子图也会发生变化,将其记作 $g=\{G(1),G(2),\cdots,G(n)\}$,其中 $G(i)(i=1,2,\cdots,n)$ 表示第 i 时间步的传播结果子图. 传播结果子图属于结构性、累积性因素.

2) 参与用户影响力

参与用户影响力指的是已经参与到事件讨论中的用户影响力情况. 直观地,参与讨论的用户影响力水平越高,易受其影响的用户数量越大,那么在未来时刻,当信息传播到其粉丝群体,会引起流行度上升. 随着时间推移,越来越多用户参与进来,参与用户影响力指标也会发生变化. 将参与用户影响力记作 $u=\{U(1),U(2),\cdots,U(n)\}$,其中 $U(i)(i=1,2,\cdots,n)$ 表示第 i 时间步的用户影响力. 由于热点事件的参与用户数量十分庞大,为简便起见,对于具体的用户影响力指标 U 可采用统计量作为特征,如最大值、最小值、平均值等. 参与用户影响力按照分类属于影响力因素,同时也是累积性因素.

3）事件话题语义

事件话题语义指的是事件引发讨论的话题本身的含义. 部分类型的话题(例如疫情、火灾等公共突发事件)在社交网络上更容易引起传播或争议，引发大量讨论，造成流行度上升. 事件话题本身的语义比较固定，将事件话题语义记作 c. 其嵌入表示方式将在第 2.4 节中详细介绍. 事件话题语义是内容性、固有性因素.

4）事件代表性帖子特征

在线社交网络上关于事件的讨论通常会存在一些代表性的帖子. 这些代表性的帖子对于吸引用户参与讨论至关重要，因此也是流行度预测因素之一. 这些帖子生成后几乎不会修改，反复被新加入讨论用户阅读. 可将代表性帖子的集合记作 h. 对于事件代表性帖子的建模可以考虑语义，也可以通过统计特征. 为了减少因素较多时考虑帖子语义带来的计算量过大问题，本研究采用了统计特征，通过代表性帖子长度和配图数量来进行建模. 事件代表性帖子特征属于内容性、固有性因素.

在流行度预测中，可利用的因素不限于上述 4 种，本方法支持根据实际需要在现有因素上进行扩展. 对因素进行定义，根据因素的具体性质分类，并采取符合该因素属性特征的嵌入方式，从而能够将其利用到流行度预测中.

2.3　基于事件库的多因素指标获取

实现在线社交网络事件库所采用的具体编程语言、开发框架、数据存储系统可以有很多种选择，在归纳事件库的本质的基础上，本节关注其抽象化、概念化的模型，以及如何服务于利用多因素进行流行度预测. 在线社交网络事件库是一个利用在线社交网络数据为分析应用提供支持的平台，主要包括多源采集模块、底层数据存储模块、无二义性的统一数据表示模型、易于访问的数据接口等，具体情况见图 1.

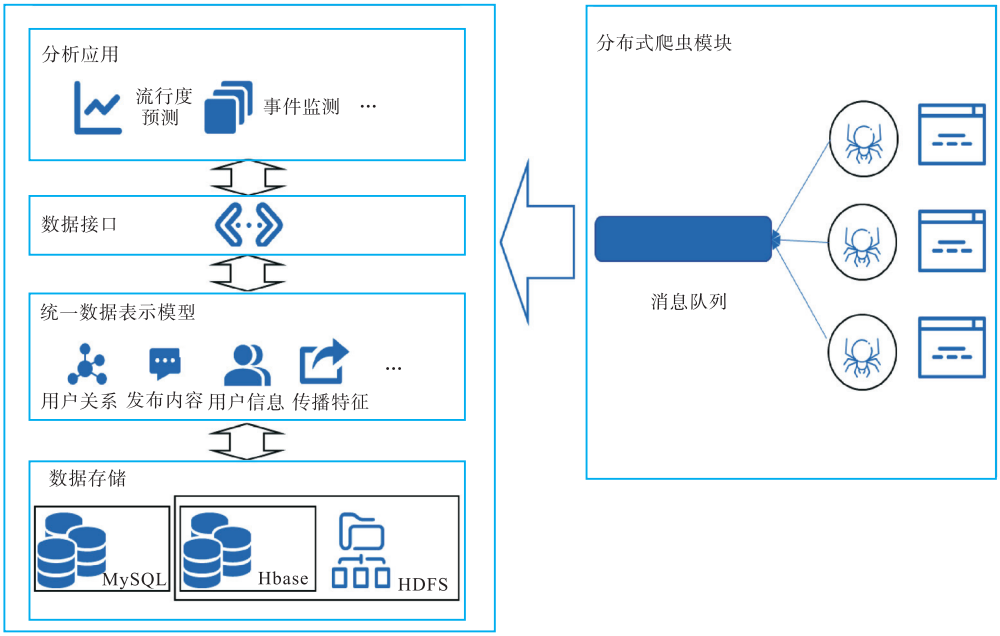

图 1　在线社交网络事件库主要组成部分

Fig.1　Main components of the online social network event base

(1) 对于多源数据采集主要利用目前已经较为成熟的分布式爬虫技术，利用消息队列收集不同进程/实例的爬虫程序所获取的数据. 其中爬虫程序有两种形式：解析网页获取数据和直接调用在线社交网络开放 API(例如 Twitter 的 Streaming API)获取数据. 无论对于哪种形式，都遵循一个重要的获取数据原则，即尽可能保留所有原始数据属性. 例如，网页中解析得到的时间戳、作者、链接地址等信息尽量统统保留.

(2) 存储层目前同样已经有较为成熟的技术，包括传统的关系型数据库以及近年来兴起的非关系型数据库. 在对数据进行建模后，可以选用的数据库存储系统以及存储方案有很多种. 在评估可选方案时，需要考虑的主要因素包括：①社交网络数据的异构性、数据量大的特点使得关系型数据库难以适用，需要支持较为灵活的数据结构以及对数据进行分片存储；②支持上层应用所需的复杂查询操作，并且能够通过索引提高查询效率. 多源异构社交网络数据的存储本身是一个值得深入研究的问题，在这里不做展开讨论. 采用了混合存储方案，利用 HDFS 分布式文

件系统存储爬虫模块获得的原始数据,利用能够分布式存储海量数据的列数据库 Hbase 存储用户发布内容(推文、微博帖子、评论等),以关系型数据库 MySQL 作为补充,存储分析结果、用户数据,应对查询要求较为复杂的情况,例如查询操作较多、需要关系查询和范围查询.

(3) 统一数据表示模型提供了社交网络数据的无二义性表示,屏蔽了具体的数据格式与底层存储方式之间的差异,使得设计数据操作与调用接口的时候无需考虑数据在具体格式、存储方式、编程语言方面的差异,关注为实现分析应用所需的抽象化的互操作逻辑. 事件中数据所包含的核心概念如图 2 所示.

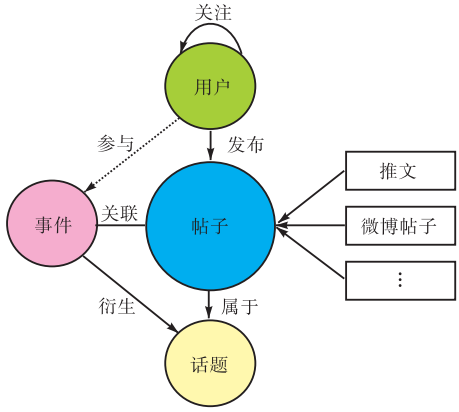

图 2 事件中数据的核心概念

Fig.2 Core concept of data in the event

其中,用户与事件之间的关系并非直接关联,而是用户发表关于某一事件的帖子而间接产生的. 对于一个事件,通常会随着事件的发展和在线讨论的进行,衍生出一到多个话题. 图 2 中所展示的均为高层抽象概念,对于帖子这一概念给出了两个典型子类,其他的概念在实际中均可根据数据来源和分析应用需求. 另外,图 2 中并未展示概念内部的属性. 对于用户来说,用户之间的关注/好友关系是网络结构分析的基础,用户概念的内部主要属性包括昵称、粉丝数、关注数以及性别年龄等用户个人信息. 对于帖子来说,其主要属性包括内容、发布时间戳等,时间戳是流行度预测、信息传播分析必不可少的属性. 话题既可以通过话题模型进行提取,也可以利用在线社交网络平台内置的话题机制.

在统一数据表示模型的基础上,根据分析应用的需求中包含的数据操作设计接口,保证数据的高可访问性.

(4) 在多因素指标获取时,首先,在线社交网络数据库采集、存储了来自各个数据源的数据,经过必要的预处理,得到统一数据表示模型的实例形式的数据. 然后对所需因素分别进行查询和处理.

① 根据帖子时间戳,划分固定的时间步并进行统计,即可得到历史流行度时间序列.

② 将用户发布帖子时间戳中最早的时刻作为用户 user 参与讨论时刻 t,可得到二元组 (user,t),则对于 τ 时刻的传播子图 $G(\tau)=(V,E)$,有 $V=\{\text{user}\mid(\text{user},t)\wedge t<\tau\}$,即 τ 时刻前已经参与讨论的用户组成了节点集合,而边集合 E 则根据 V 中用户的关注关系得到. 在实际处理中,可以先不考虑时刻而将用户全集构建网络结构图,以键值对的形式存储节点与其关注节点集合,在构建传播子图的时候能够复用这一结构从而提高效率.

③ 对于用户影响力因素的获取与传播子图类似,同样以 (user,t) 作为线索,按照时间步进行统计即可. 在实际处理中可采用增量处理,复用上一时间步的结果.

④ 事件话题语义因素与事件代表性帖子特征两个因素属于固有因素,因此省去了需要以时间为线索进行处理的过程,可以直接通过查询获取.

2.4 多因素分组嵌入

正如第 2.2 节中所分析的,各个指标因素指标之间在性质和数学表示上各不相同,在事件流行度预测中,对于多种相关因素指标的利用需要面对这种各因素指标之间的异构性. 对于传统的回归分析的流行度预测方法来说,新增一项考虑因素即为模型增加一个变量. 然而对于相当部分因素,无法对其进行简单量化,例如传播结果子图是以图结构形式存在的,事件话题语义则包含了自然语言文字描述. 如果分别对这些因素进行建模和量化,本质上要面对的是人工特征提取的问题,必然会引入传递性误差,从而导致这些指标在预测中无法发挥作用. 考虑到以上因素,采用基于深度学习的方法,首先将这些异构因素按照特征进行分组,通过适当的神经网络结构进行嵌入,可以从真实数据中学习得到结合、可利用的低维表示,之后通过基于神经网络的模型将这些低维表示的因素与历史流行度结合到一起,实现对未来流行度的预测. 这一方法的优点在于无需人工设计如何将因素映射到量化的、可用于预测的指标,避免对因素的间接利用导致的失真.

如图 3 所示,将传播结果子图与参与用户影响力分为累积性因素分组,利用张量连接方法,以时间步为对齐依据,在每个时间步内对二者的表示向量进行拼接,之后对所有时间步的拼接后表示向量利用 LSTM 进行进一步处理,得到该分组的嵌入表示. 对于事件话题语义和事件代表性帖子特征两个固有性

因素，分别处理得到表示向量后，同样采用张量连接方法，拼接表示向量得到该分组的嵌入表示. 分组嵌入表示结果经过张量连接后与历史流行度数据一起作为最终预测模型的输入.

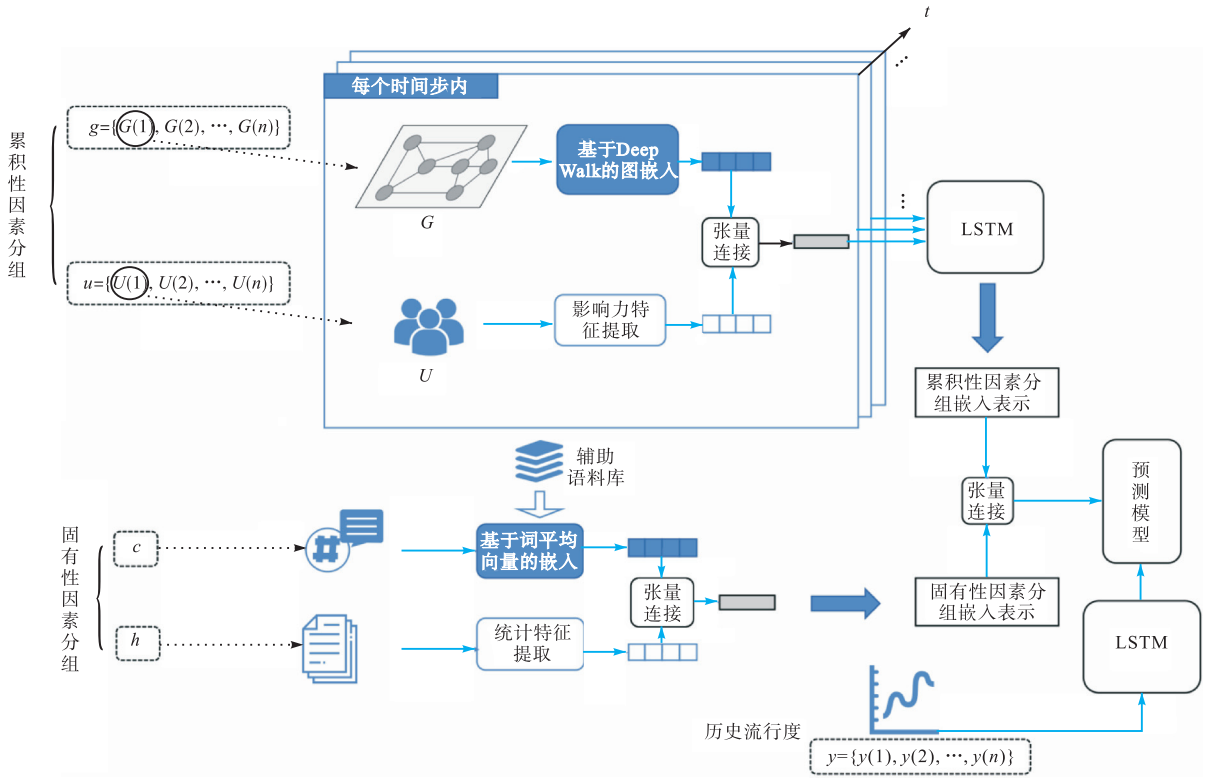

图 3　多因素分组嵌入

Fig.3　Multi-factor grouping embedding

对于分组内各个因素，按照其特点设计嵌入表示方法.

1）传播结果子图因素表示

对于传播结果子图这一结构性因素，需反映其拓扑结构特征. 采用基于 DeepWalk 的图嵌入方法，对随机游走的节点序列进行基于 SkipGram 的嵌入. 在得到节点的嵌入表示向量后，对其进行平均，即为传播结果子图的嵌入表示向量，具体算伪代码如下.

算法　基于 DeepWalk 的图嵌入

Input：输出嵌入表示的维数 d，传播子图网络结构 $G=(V,E)$，在 SkipGram 中采用的窗口大小 ω，每个节点游走的次数 N，游走的步数 L

Output：$\phi\in R^d$ 传播子图嵌入表示

1：初始化矩阵 $\boldsymbol{\Phi}\in R^{|V|\times d}$ //用于保存全部节点嵌入表示，最后根据它生成 ϕ

2：根据 V 初始化 SkipGram 采用的霍夫曼树 T

3：for $i=0$ to N do

4：$O=\text{Shuffle}(V)$　//节点乱序洗牌

5：for each　$v_i\in O$　do

6：$W_{v_i}=\text{RandomWalk}(G,v_i,L)$ //在 G 上调用自 v_i 开始步数为 L 的随机游走算法

7：SkipGram($\boldsymbol{\Phi},W_{v_i},\omega,T$) //调用 SkipGram 模型，将节点的嵌入表示更新到 $\boldsymbol{\Phi}$

8：end for

9：end for

10：$\phi=\text{average}(\boldsymbol{\Phi})$ //平均后得到传播子图嵌入表示

在每个时间步内执行上述过程，即对于 g 中每个 $G(i)(i=1,2,\cdots,n)$ 执行基于 DeepWalk 的图嵌入算法，可得到各个时间步的传播结果子图组成的嵌入表示记作 $\theta_g=\{w_g(1),w_g(2),\cdots,w_g(n)\}$，其中 $w_g(i)(i=1,2,\cdots,n)$ 为 g 中第 i 时间步的传播结果子图的嵌入表示向量. 之后如图 3 所示，θ_g 将与参与用户影响力表示按照时间步进行对齐和拼接.

2）参与用户影响力因素表示

用户影响力的最直观建模方式即采用粉丝数量进行建模，因为用户的粉丝是当用户发布/转发事件相关信息后，潜在的次级转发群体. 对于一个事件来说，常常有大量用户参与讨论，其中不乏有粉丝数量十分庞大的“微博名人用户”加入. 考虑到庞大的粉丝数量为数据处理带来挑战，并且由于在线社交网络平台处于隐私和商业利益考虑，普遍禁止对用户的完整粉丝进行获取，在建模时不考虑具体的粉丝个体. 因此，对于每个时间步，在粉丝数量的基础上，构建特征为

$$\mu=[|V_{\text{celebrity}}|,\text{sum}(\eta_{\text{fans}}),\max(\eta_{\text{fans}}),\text{med}(\eta_{\text{fans}}),\text{mean}(\eta_{\text{fans}}),\ln(\text{sum}(\eta_{\text{fans}})),\ln(\max(\eta_{\text{fans}})),\ln(\text{med}(\eta_{\text{fans}})),\ln(\text{mean}(\eta_{\text{fans}}))] \tag{1}$$

其中的元素分别代表截至该时间步内参与讨论的“微博名人用户”数量、参与用户粉丝总数、参与用户粉丝数的最大值、参与用户粉丝数的中位数、参与用户粉丝数的均值、参与用户粉丝总数的对数、参与用户粉丝数最大值的对数、参与用户粉丝数的中位数的对数、参与用户粉丝数的均值的对数. 其中，引入对数特征可以减轻不同信息中粉丝数的数量级相差很大的问题，粉丝数最小值对于促进信息传播、反映流行度未来趋势参考意义不大，因此未予考虑. 用户影响力因素为：$\theta_u=\{\mu_u(1),\mu_u(2),\cdots,\mu_u(n)\}$，$\mu_u(i)(i=1,2,\cdots,n)$ 为 u 中第 i 时间步的参与用户影响力特征表示向量.

如图 3 所示，θ_u 与 θ_g 按照时间步进行对齐和拼接，先通过 LSTM 层再接入预测模型，LSTM 层的参数在预测中进行训练. 将通过 LSTM 层后的向量记作 $\boldsymbol{\Theta}_c$，代表累积性因素分组表示向量，其维数由 LSTM 层神经元数量决定.

3）事件话题语义因素表示

通过将事件话题语义因素转化固定长度的表示向量，可以将其与其它因素相结合，应用到热点事件流行度预测中去. 为简便起见，时间话题语义因素直接采用话题字符串. 对于在线社交网络平台，如微博、Twitter 等，均内置话题功能，通过形如“#”的井号加上一个词组或者一段十分简短的句子来表示话题，并且提供以话题为中心的讨论、搜索等功能. 话题的关键字是人为选定的，一般足以涵盖话题中的主要语义因素. 例如，新浪微博上的话题“#世卫官员说基因序列证据表明病毒来源自然界#”为一个短句，美国新冠病毒在 Twitter 上的话题为“#Coronavirus USA”，由两个词组成.

考虑到这些话题字符串均为短文本，为避免稀疏性问题，采用外部辅助语料库来得到话题字符串的表示[11]. 具体地，采用维基百科文本语料库作为辅助语料库，通过 SkipGram 模型训练得到 l 维的预训练词向量集合[12]. 给定一个话题字符串，利用预训练词向量集合得到 $l\times n$ 的表示矩阵为

$$\boldsymbol{M}=[\mathcal{V}_1,\mathcal{V}_2,\cdots,\mathcal{V}_n] \tag{2}$$

式中：$\mathcal{V}_i(i=1,2,\cdots,n)$ 表示话题字符串对应的词向量；n 为话题中词的个数，这里根据取词向量维数 $l=300$. 由于话题中词的个数不确定，为了得到统一长度的表示向量，设计如图 4 所示的网络结构，图中的全连接层输出即为事件话题语义因素的表示向量. 在这里，采用话题分类任务进行预训练，在全连接层后接入分类所需的全连接 + Softmax 层即可得到与训练所需分类用网络，分类标注信息则利用了社交网络事件库中所存储的类别信息，训练时采用交叉熵作为损失函数.

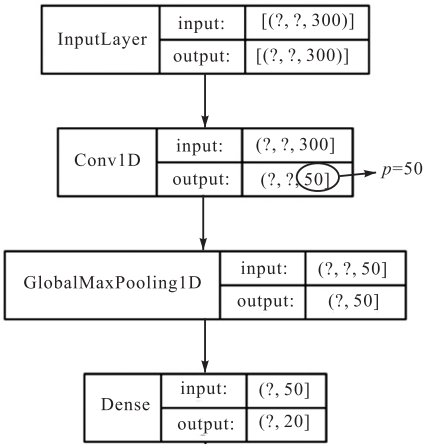

图 4 用于获取话题语义因素嵌入表示的网络结构（设 $p=50$、输出矩阵维数为 20 的情况）

Fig.4 Network structure used to obtain the embedded representation of the topic semantic factors（if $p=50$ and the dimension of output matrix is 20）

给定话题字符串的预训练词向量表示矩阵 $\boldsymbol{M}$，首先使其通过如图 5 所示的卷积层，其中采用 ReLU 函数作为激活函数，针对短文本的特点取步长为 1、卷积核大小为 3. 将卷积核数量记作 p，则卷积层输出为

$$\boldsymbol{S}=[s_1,s_2,\cdots,s_p]^{\mathrm{T}} \tag{3}$$

式中 s 为各个卷积核的输出. 之后利用一维最大池化操作得到固定长度为 p 的矩阵

$$\boldsymbol{S}_{\max}=[\max(s_1),\max(s_2),\cdots,\max(s_p)]^{\mathrm{T}} \tag{4}$$

最后通过全连接层，得到向量 $\boldsymbol{\theta}_s$，其维数由全连接层神经元数量决定.

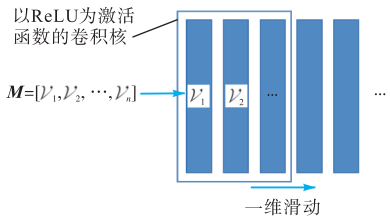

图 5 一维卷积层

Fig.5 One-dimensional convolutional layer

4）事件代表性帖子特征因素表示

对于一个事件话题中的在线讨论，通常会出现大

量帖子. 其中具有代表性的帖子会被新加入话题讨论的用户所阅读，影响着流行度的发展. 因此，对这些帖子的特征进行建模有助于流行度预测. 采用启发式方法对事件代表性帖子的特征进行建模，主要考虑这样一种现象：代表性帖子的吸引人程度影响着用户对事件话题感兴趣的程度，进而影响着流行度的发展趋势. 对于微博、Twitter 来说，图文并茂、内容详细的微博帖子或推文会吸引更多的用户参与讨论、进行转发，从而使得流行度激增. 这里采用统计特征进行建模，主要包括代表性帖子的最大长度，代表性帖子的词典长度以及一个是否包含图片的指示函数，记作

$$\boldsymbol{\theta}_h = [\max_{d\in h}(|d|), \max_{d\in h}(\|d\|), I_h] \tag{5}$$

式中：$|d|$ 表示帖子长度；$\|d\|$ 表示帖子中唯一词数量；I 为指示函数，当代表性帖子中包含图片则取 1，否则取 0.

将 $\boldsymbol{\theta}_s$ 和 $\boldsymbol{\theta}_h$ 进行张量连接得到固有性因素分组表示向量，记作 $\boldsymbol{\Theta}_f$.

2.5　多因素耦合的流行度预测

流行度预测模型的任务是预测未来某一时刻的流行度. 设计基于神经网络的流行度预测模型需解决三方面问题，即如何处理历史流行度，采用怎样的神经网络结构，以及如何利用前文中的各种因素的表示向量. 首先，对于历史流行度，也使其通过待训练的 LSTM 层，其输出记作 $\boldsymbol{\theta}_y$，之后对 $\boldsymbol{\theta}_y$、$\boldsymbol{\Theta}_c$ 与 $\boldsymbol{\Theta}_f$ 进行张量拼接，通过批量归一化（batch normalization）层，之后输入到用于预测的一组以 ReLU 为激活函数的全连接层，最后一层神经元数量为 1 且不设激活函数，以应对回归问题. 在模型的训练中，以均方损失误差（mean squared error，MSE）作为损失函数.

对于除最后一层以外的全连接层，神经元数量可参照经验公式进行设置为

$$n_{\text{layer}} = \text{lb}(|\boldsymbol{\theta}_y|) + \text{lb}(|\boldsymbol{\Theta}_c|) + \text{lb}(|\boldsymbol{\Theta}_f|) + \delta \tag{6}$$

式中：等式右则前 3 项均为向量的维数；δ 为调节因子.

3　实　验

3.1　数据集选取

本研究采用了斯坦福 SNAP 组的 Twitter7 数据集，首先从 6 500 万推文中通过随机采样粗略选取了 330×10^4 个话题标签（Hashtag），其中涉及发布相关推文的 $4\,000\times10^4$ 用户. 之后，分话题标签统计推文数量，对 330×10^4 话题标签的流行度数据进行分析，如图 6 所示. 可以看出，流行度大于等于横坐标的标签数量与流行度之间呈幂律关系，大部分话题标签下的推文数量较少，对于流行度演化分析与预测来说缺少研究价值，因此在进一步筛选中予以排除. 最终，本研究筛选出 3 000 个主题标签将其中包含的推文作为实验对象.

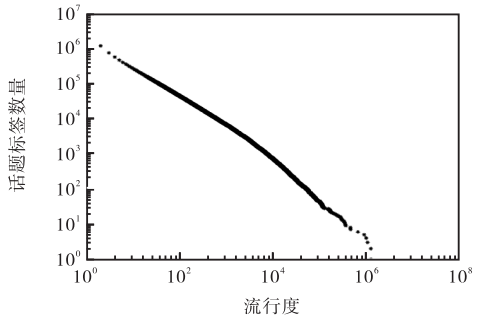

图 6　话题标签流行度分布情况

Fig.6　Popularity distribution of topic hashtags

3.2　实验方案设计

本研究所提出的模型简记作 MF（multiple factors）模型. 实验对比的模型为深层神经网络模型（DNN）、支持向量回归模型（SVR）、SH 流行度预测模型[13].

实验过程中采用 5 折交叉验证. 将数据集随机均分为 5 个子集，利用其中 4 个子集作为训练集，最后 1 个子集作为测试集.

实验首先考察预测模型中全连接层的数量对于预测准确性的影响，然后在相同的数据集上与 DNN、SVR、SH 模型进行对比. 对比以流行度预测准确率为度量标准，主要考察置信度区间内预测流行度与打标流行度的百分比误差在给定范围内的比例，记作

$$A = \text{average}\left[\left(1-\frac{|y_{\text{pred}}-y_{\text{act}}|}{y_{\text{act}}}\right)\times100\%\right] \tag{7}$$

式中：A 为流行度预测平均准确率；y_{pred} 为流行度预测值；y_{act} 为流行度实际值.

3.3　实验结果分析

本文将模型全连接层数设置为从 1 层逐渐升至到 9 层. 如图 7 所示，随着层数的增加，预测准确率逐步提高，但是在第 3 层之后，准确率开始下降，模型预测准确率的峰值为第 3 层的 95.79%. 所以，本文设定的全连接层数为 3 层.

在模型对比实验中，本文采用与层数实验相同的数据集. 如表 2 所示，MF 模型在最高准确率、最低准确率和平均准确率 3 项指标上均高于其他 3 类模型. 这主要是由于 MF 模型对网络因素的提取、处理

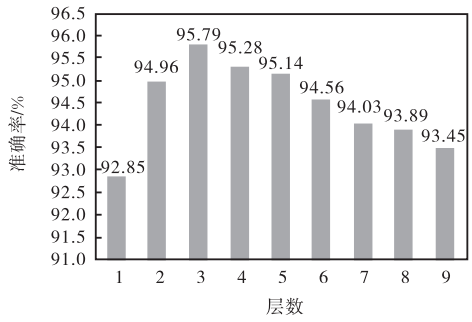

图 7 不同全连接层数的预测准确率实验数据

Fig.7 Experimental data of the prediction accuracy of different numbers of full connected layers

能力更为优异. 由此本文能够看到,基于社交网络事件库多因素耦合的流行度预测方法优于现有的典型流行度预测方法.

表 2 不同流行度预测模型准确率对比

Tab.2 Accuracy comparison of different popularity prediction models %

模型	最高准确率	最低准确率	平均准确率
MF	97.43	93.82	95.63
DNN	96.10	91.11	93.61
SVR	94.33	88.66	91.50
SH	91.56	84.98	82.27

4 结 语

本文研究了在线社交网络事件库多因素耦合的流行度预测. 首先提出了一种基于事件库的多因素指标获取方法,能够利用事件库对于社交网络数据的统一存储,从多源异构数据中提取各因素指标. 其次提出了一种多因素耦合的流行度预测方法,能够通过分组嵌入得到因素指标的可相互结合的低维表示,进而在预测中实现多因素指标的综合利用. 最后基于真实数据集上的实验表明了所提出方法的准确性优于现有方法.

参考文献:

[1] Arnaboldi V, Passarella A, Conti M, et al. Online Social Networks: Human Cognitive Constraints in Facebook and Twitter Personal Graphs[M]. Amsterdam: Elsevier Science Publishers B. V., 2015.

[2] Wang Z, Ye X J, Wang C K, et al. Network embedding with completely-imbalanced labels[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, DOI: 10.1109/TKDE2020.2971490.

[3] Wang Z, Wang C K, Pei J S, et al. Causality based propagation history ranking in social networks[C]// Proceedings of the 25 th International Joint Conference on Artificial Intelligence(IJCAI-16). New York, USA, 2016: 3917-3923.

[4] Zhao J, Wu J, Feng X, et al. Information propagation in online social networks: A tie-strength perspective[J]. Knowledge and Information Systems, 2012, 32(3): 589-608.

[5] Matsubara Y, Sakurai Y, Prakash B A, et al. Rise and fall patterns of information diffusion: Model and implications[C]// Proceedings of the 18 th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Beijing, China, 2012: 6-14.

[6] Bandari R, Asur S, Huberman B A. The pulse of news in social media: Forecasting popularity[C]// The 6th International AAAI Conference on Weblogs and Social Media. Dublin, Iredland, 2012: 26-33.

[7] Szabo G, Huberman B A. Predicting the popularity of online content[J]. Communications of the ACM, 2010, 53(8): 80-88.

[8] Bao P, Shen H W, Huang J, et al. Popularity prediction in microblogging network: A case study on sina weibo[C]// Proceedings of the 22nd International Conference on World Wide Web. Rio de Janeiro, Brazil, 2013: 177-178.

[9] He X, Gao M, Kan M Y, et al. Predicting the popularity of web 2.0 items based on user comments[C]// Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Queensland, Australia, 2014: 233-242.

[10] Kong S, Mei Q, Feng L, et al. Predicting bursts and popularity of hashtags in real-time[C]// Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval. Queensland, Australia, 2014: 927-930.

[11] Li C, Duan Y, Wang H, et al. Enhancing topic modeling for short texts with auxiliary word embeddings[J]. ACM Transactions on Information Systems, 2017, 36(2): 1-30.

[12] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

(责任编辑:孙立华)