引用格式:徐翔,杨心茹. 网络舆论主题的周期长度对主题热度的影响及预测[J]. 中国传媒大学学报(自然科学版),2024,31(03):01-09.
文章编号:1673-4793(2024)03-0001-09

网络舆论主题的周期长度对主题热度的影响及预测

徐翔*,杨心茹
(同济大学艺术与传媒学院大数据与计算传播研究中心,上海201804)

摘要:网络平台中舆论主题演化、起伏的周期与节律是否以及如何影响到主题的热度,是具有困惑度和缺乏研究的问题,探索该问题对于研究主题周期功能、主题演化机制等具有重要意义。本文基于今日头条平台帖子样本,探索主题的周期长度对主题热度的作用及可预测模型。采取功率谱分析计算主题周期长度,使用逻辑回归与决策树检验主题周期长度对主题热度的预测效果。结果表明:1)多数主题具有1周、½周、⅓周这三种周期,可称为平台三大"主频","主频"对主题热度起到抑制作用;2)对主题热度起到抑制作用的还有⅓周左右的"波长"集群、1周左右的"波长"集群;3)[2.6天,3天]区段(α区段)、[5.2天,6天]区段(β区段)内的周期长度,对主题热度起正向作用,它们主要是处于6天长度内的"快频率"和短波长,且在三大"主频"的抑制区波长以外;4)α区段约处于⅓周和½周的这2个主频中间,β区段约处于½周和1周这2个主频中间,且α区段与β区段之间有近似的2倍关系。上述周期长度的特征,可对主题热度起到一定的预测作用。

关键词:网络平台;周期长度;传播热度;功率谱
中图分类号:G206 **文献标识码**:A

Influence and prediction of the cycle length of the internet public opinion topics on the popularity of the topics

XU Xiang*, YANG Xinru
(College of Arts & Media, Tongji University, Shanghai 201804, China)

Abstract: It is a problem with confusion and lack of research whether and how the evolution, fluctuation cycle and rhythm of public opinion in the network platform affect the popularity of the topics. Exploring this problem is of great significance for studying the function of topics cycle and the mechanism of topics evolution.Based on post samples of Toutiao platform, exploring the effect of the period length of the topic on the popularity of the topic and its predictable model in this paper. Power spectrum analysis was used to calculate topic cycle length, and logistic regression and decision tree were used to test the prediction effect of topic cycle length on topic popularity.The results show that: 1) Most themes have a cycle of 1 week, ½ week, ⅓ week, which are called the prominent "main frequency" of Toutiao platform, and these three cycle lengths have a negative effect on the popularity of the theme; 2) The "wavelength" cluster of about ⅓ week and 1 week also have a negative effect on the popularity of the theme; 3) The period length in the [2.6 day, 3 day] segment (α segment) and [5.2 day, 6 day] segment (β segment) have a positive effect on the subject heat, which are mainly in the "fast frequency" and short wavelength within the 6-day length, and outside the wavelength of the above three categories of suppressed regions; 4) The α segment has a positive influence on the frequency of ⅓ and ½ weeks, and the β segment is in the middle of the 2 frequencies of ½ weeks and 1 weeks, and there is an approximate 2x relationship between the α segment and the β segment. Through the characteristics of the above cycle length,

基金项目:国家自然科学基金项目(71804126);上海市"科技创新行动计划"软科学研究项目(23692110600)
作者简介(*为通讯作者):徐翔(1983-),男,博士,教授、副院长,主要从事网络传播研究。Email:xuxiang210089@163.com;杨心茹(2000-),女,硕士研究生,主要从事智能媒体与网络研究。Email:3046560397@qq.com

the theme popularity can be predicted to a certain extent.

Keywords: network platform; cycle length; propagation popularity; power spectrum

1 引言

随着技术发展,算法推荐机制与信息平台紧密结合,为公众推送海量内容,已有研究围绕互联网信息或主题的传播规律、热度等展开了诸多讨论。本文从主题的周期性与主题热度之间的关系出发,探索主题周期的功能与作用。许多研究已发现网络话题的演化具有生命周期规律,分析生命周期各个阶段特征[1-2],从而探索高热度事件、典型事件的舆情演化规律[3],有研究从突发事件的单一周期中提取关键节点、测算热度走向[4]。然而,对热点话题某种生命周期的划分、热度起伏的描述研究仍有不足:首先,未探索周期对舆论的具体作用和影响;其次,缺乏探讨众多非高热主题、普遍事件的周期规律,导致研究所得的周期性不具备共通性;最后,未继续探究周期如何作用于热度,以及是否可以预测热度。总之,已有研究仍未解读网络平台中话题周期性变化影响话题热度涨落的作用机制,使得舆论调控缺乏具体可操作的路径与手段。本文基于时间序列,对推荐算法平台帖子进行综合分析,探讨了主题周期的多种类型、周期长度以及关键周期,构建了主题周期对主题热度的预测模型。

2 研究缘起与文献回顾

网络主题具有周期性,且主题热度的峰值、谷值均与周期阶段存在关联,这为本文研究提供重要依据。Qiao等[5]指出网络舆情的产生和发展必然在一定的时空环境中出现、发展和实现,即从兴起到衰落的生命周期。匡文波[6]指出新媒体舆论都经历了从有到无的过程,其演变发展符合舆论酝酿期、爆发期、消解期这种周期规律,由此建立了新媒体舆论的议题出现、存活、整合、消散模型。昂娟[7]则提出政府应当从“潜伏期”、“爆发期”、“持续期”和“衰落期”四个舆论生命周期阶段引导网络舆论。陈福集等[8]通过E-Divisive算法将网络舆情演化过程划分为波动、高峰、衰退三个阶段。但这方面的已有研究更注重时域分析而缺乏频域分析。

热点主题存在特殊的周期特征,并非任意周期长度特征都能关联和影响到主题热度。He等[9]发现热点话题具有周期短、增长快、衰退快、成熟期短等特点,且绝大多数热点话题只有一个峰值,一般不超过两个峰值。毛太田等[10]则以“上海警察绊摔小孩事件”这一热点事件为例,指出新媒体时代下的网络热点事件传播速度更快、信息更新周期更短,整个生命周期并未经历一个明显的“成长期”。Zhang等[11]分析了突发事件期间网络舆情的特征和演化机制,他们以飓风“艾尔玛”事件为例,研究了推文情绪与舆论生命周期之间的关系及其对推文量的影响。

公众对事件或主题的关注具有周期性,这意味着主题热度随着“注意力周期”的变化而变化。Downs构建了“议题注意力周期”的五个阶段[12],即前问题阶段、问题惊现与热情高涨阶段、困难与成本认知阶段、热情逐渐消退阶段、舆论消退阶段[13]。李永宁等[14]指出,公众对多条相关信息的关注时长构成了公众的议题注意力周期,而2010年至2016年,公众对公共议题的关注周期显著缩短。Shih等[15]指出媒体对每种流行疾病都有不同的注意力周期模式。在公众注意力周期的不同阶段中,主题因不同的关注度而呈现出不同的热度水平。

已有研究基于主题数量时间序列对网络主题的热度展开一定程度的预测,说明本文预测热度存在可行性。Saleiro、Soares[16]通过新闻周期来预测新闻中个人、组织、公司或地理位置等实体在社交媒体上的受欢迎程度。Su等[17]指出网络舆情趋势呈现出非线性和季节性波动的特征,并采用季节性修正指数灰色伯努利模型与ARIMA模型,对“林生斌”和“唐山打人”网络事件的热度进行了较高精度的预测。张虹等[18-19]则基于小波多尺度分析、神经网络方法对网络论坛话题热度趋势进行了精度良好的预测。

此外,关于社交媒体信息的周期性,部分研究指出社交媒体信息具有半衰期、老化期,并进行测量,为网络主题的周期测量及其生成机理分析提供了参考。例如梁芷茗[20]基于信息生命周期理论基础,描绘新浪微博“热点话题”的生命周期曲线,测出新浪微博这一网络结构单元的半衰期为8天。江燕青等[21]采用内容分析法和SPSS统计检验方法分析了微博在转发半衰期和评论半衰期上的信息老化差异及其影响因素。马费成等[22]通过实证分析得出某网站上的书签信息半衰期为557天。但这些研究还是未能进一步充分回答周期能否预测信息热度、舆论热度这一问题。

3 研究设计与实证检验

本文选取今日头条平台帖子作为研究样本,采用八

爪鱼采集器编写网络爬虫进行数据采集，从今日头条首页自动推荐的18个内容版块（财经、科技、热点、国际、军事、体育、数码、娱乐、历史、问答、美食、游戏、旅游、育儿、养生、时尚、视频、同城）中，每天早中晚各随机抓取一次帖子，且凭借今日头条自身分区设置，使得帖子样本均匀而广泛地分布在每个内容版块内，从而确保每个内容版块内随机抓取的帖子用户数量基本相等。对于每个版块的种子用户，按照历史发帖情况，研究初始抓取了6453万条帖子样本，选择其中2020年7月1日到2023年6月30日共49024037条帖子样本，再按照18个内容版块中每个版块内10109个用户的数量进行筛选，共得到所有版块用户所发布的28473604条帖子样本。本次抓取样本量以及分布情况令研究样本具有代表性和普遍性，符合本文对于广泛主题周期性的探索需求。

3.1 帖子向量化与聚类

本文依据时间序列进行分析，所抓取的每日帖子数量都大于或等于4955条，因此从中每天随机抽取4955条，保障每天分析的帖子数量相等。采取Word2Vec模型[23]对帖子文本内容进行量化，由此将帖子文本内容转换为向量，从而计算帖子之间的余弦相似度。利用GenSim[24-25]将分词后的每一个词转换成一个300维的Word2Vec词向量，对这些词的词向量平均池化后得到该帖子的语句向量[26-27]。Shen等[28]的研究对简单词向量模型，即词向量进行等权求平均向量（Simple Word-Embedding Model，SWEM）的方法，与循环和卷积神经网络进行比较，多数情况下SWEM表现出高性能。训练Word2Vec所使用的语料采用自行抓取的26G的中文语料库，来源包括媒体新闻、网络帖子、经典名著等。随后，采取K-means聚类的方式，采取簇内误差平方和反映聚类误差，聚类误差的变化见图1。结合“肘拐点”方法将帖子聚类为800类，其中聚类数目超过800类后的聚差误差减少较为缓慢，再增加聚类数目对类间区分度的帮助不大。

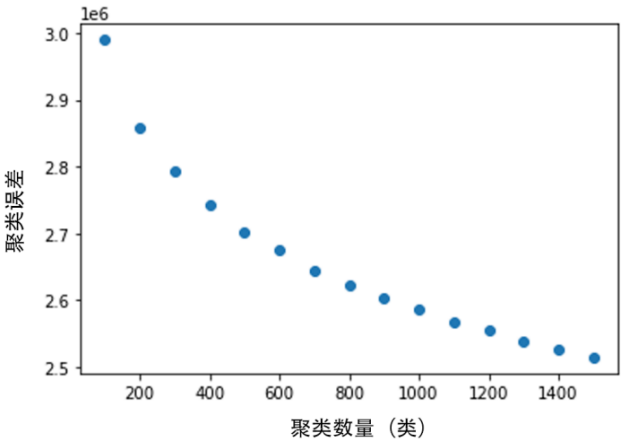

图1 K-means聚类误差变化

提取这800个主题中TF-IDF（一种统计方法，评估字词在一个文本集中的重要程度，TF-IDF越高，字词重要程度越高）最高的10个词，节选序号最前的10个主题示例如表1。可见聚类得到的每个主题的特征词具有良好的区分度。

表1 800个主题中TF-IDF最高的10个关键词（节选）

主题	Top 10关键词									
主题0	国产	德国	骑士	美国	美军	中国	导弹	二战	装备	坦克
主题1	家庭	自己	丈夫	妻子	父亲	妈妈	母亲	父母	孩子	儿子
主题2	开箱	体验	系列	手机	vivo	曝光	评测	oppo	华为	三星
主题3	人口	全球	30	国家	日本	世界	我国	20	10	美国
主题4	日常	情侣	恋爱	婚后	爱情	盘点	之间	感情	关系	婚姻
主题5	服务	第一	专家	医疗	健康	开展	中心	北京	患者	人民
主题6	力量	梦想	自由	温暖	精彩	美好	时光	岁月	人心	生命
主题7	夏天	夏季	夏日	外卖	外国	外国人	外婆	外媒	复仇	龙凤胎
主题8	运动	继续	工作	自己	学习	快乐	加油	坚持	每天	生活
主题9	球队	表现	胜利	挑战	湖人	时刻	关键	优势	对手	成功

3.2 帖子周期与频率分布

通过计算这800个主题在每天所有帖子文本中的占比数值，得出这些主题按照天数所存在的不同的演化数值，也即得到800个主题各自在1095天的分布比例，从而得到1095行×800列的面板数据。对这800列数据进行平稳性检验，由于存在一部分序列并不符合平稳性要求，因此将它们全部进行一阶差分处理，一阶差分处理后经过单位根检验（Augmented Dickey-Fuller Test，ADF），全部符合平稳性要求。进而对一阶差分后时间序列上的800列数据进行周期的计算与检验。

关于周期计算，有研究采取小波分析法来计算分析我国森林草原火灾、台风风暴潮直接经济损失、微博中的社会情绪等自然现象或社会事件的周期特征[29-31]。而本文则采取功率谱分析法，利用功率谱能够得到时间序列中的能量在不同频率上的分布情况，从而能够分析主要周期[32]。功率谱是一种以傅里叶变换为基础的频域分析方法，本文采取的功率谱计算方式如下[33]：

对于一个样本量为n的离散时间序列$x_1, x_2, \cdots, x_n$，根据谱密度与自相关函数互为傅里叶变换的重要性质，通过自相关函数间接做出连续功率谱估计。对一时间序列x_t，最大滞后时间长度为m的自相关系数$r(j)$（j=0,1,2,…,m）为：

$$r(j)=\frac{1}{n-j}\sum_{t=1}^{n-j}\left(\frac{x_t-\bar{x}}{s}\right)\left(\frac{x_{t+j}-\bar{x}}{s}\right) \quad (1)$$

式中,$\bar{x}$为序列的均值,s为序列的标准差。

由式(2)得到不同波数k的粗谱估计值:

$$\hat{s}_k=\frac{1}{m}\left[r(0)+2\sum_{j=1}^{m-1}r(j)\cos\frac{k\pi j}{m}+r(m)\cos k\pi\right],k=0,1,\cdots,m。 \quad (2)$$

式中,$r(j)$表示第j个时间间隔上的相关函数。在实际计算中考虑端点特性,常用式(3):

$$\begin{cases}\hat{s}_0=\frac{1}{2m}\left[r(0)+r(m)\right]+\frac{1}{m}\sum_{j=1}^{m-1}r(j)\\ \hat{s}_k=\frac{1}{m}\left[r(0)+2\sum_{j=1}^{m-1}r(j)\cos\frac{k\pi j}{m}+r(m)\cos k\pi\right]\\ \hat{s}_m=\frac{1}{2m}\left[r(0)+(-1)^m r(m)\right]+\frac{1}{m}\sum_{j=1}^{m-1}(-1)^j r(j)\end{cases} \quad (3)$$

式(3)中,m是给定的,在已知序列样本量为n的情况下,功率谱估计随m的不同而变化。当m取太小值时,谱估计过于光滑,不容易出现峰值,难以确定主要周期,本文的m取为$\frac{n}{2}$。

然而,并非所有谱值都具有显著性,为进一步有效检验主题周期的显著性,本文引入红噪声进行筛选。红噪声作为检验波谱中谱值显著与否的重要指标,常被用于各类气候、水文、地质等主题的振荡周期研究中,例如运用于极端气候、洪水变化、地下水位与降水的周期性研究、火山活动的周期性研究中[34]。在计算中,画出所有主题在整个时间序列上的频谱后,将大于红噪声标准谱的频谱定义为显著周期,否则该周期不显著。红噪声具体计算过程如下。

红噪声标准谱的计算公式为式(4)所描述:

$$S_{0k}=\bar{S}\left[\frac{1-r(1)^2}{1+r(1)^2+2r(1)\cos\frac{\pi k}{m}}\right] \quad (4)$$

式中,$\bar{S}$为m+1个谱估计值的均值,即如式(5)所示:

$$\bar{S}=\frac{1}{2m}(S_0+S_m)+\frac{1}{m}\sum_{k=1}^{m-1}S_k \quad (5)$$

主题的功率谱计算及红噪声检验的结果,是将红噪声标准谱作为“基准值”,某频率上的功率谱值大于红噪声标准谱取值时才认定该频率或其对应的周期长度具有显著性。分别抽选第200个、400个、600个、800个主题,其功率谱和红噪声标准谱的比较图依次如图2、图3、图4、图5所示,其中蓝色实线是功率谱的取值,红色虚线是红噪声标准谱的取值。

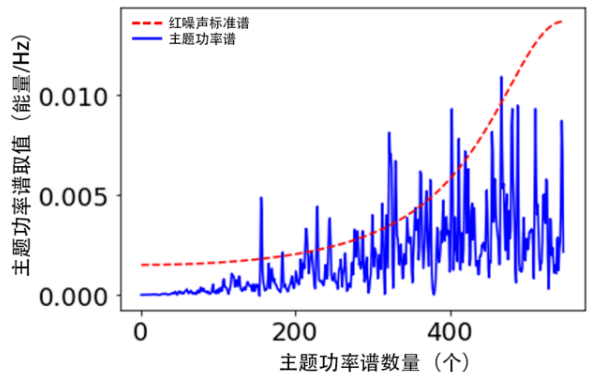

图2 第200个主题的功率谱和红噪声标准谱对比

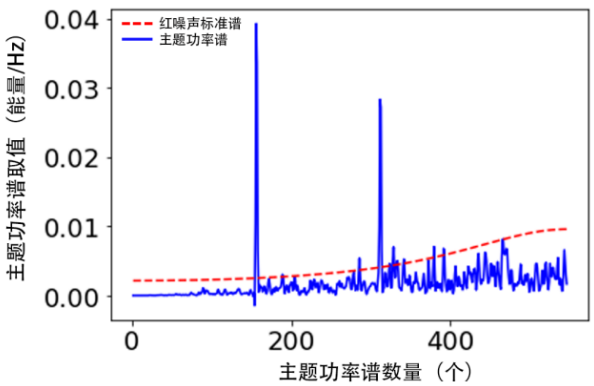

图3 第400个主题的功率谱和红噪声标准谱对比

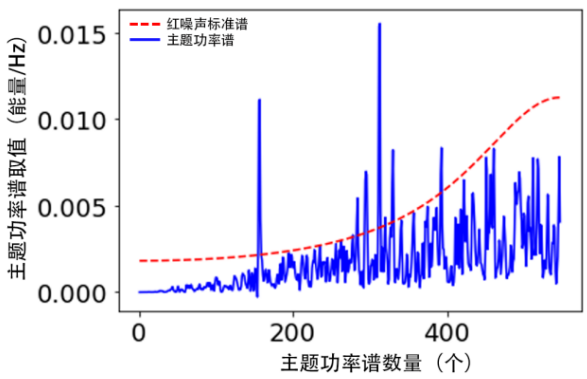

图4 第600个主题的功率谱和红噪声标准谱对比

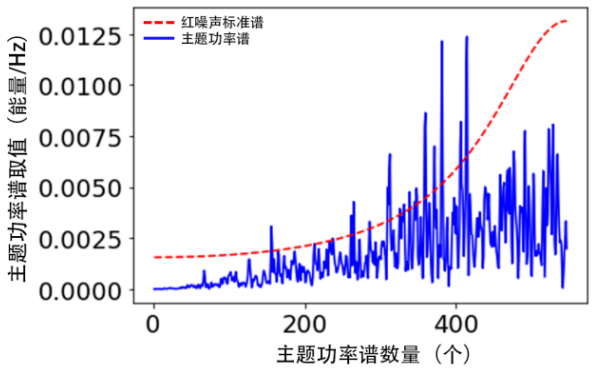

图5 第800个主题的功率谱和红噪声标准谱对比

800个主题及其周期/频率分布的独热码图如图6所示，横轴表示频率（即时间序列内振荡的次数，每个频率对应于一个独特的周期长度，类似于“波长”），纵轴表示主题有或无该频率，色点表明任意主题在横轴频率上经红噪声检验后具有显著性，亮点表明该行对应的主题具备该频率，蓝色点表明该行对应的主题不具备该频率。

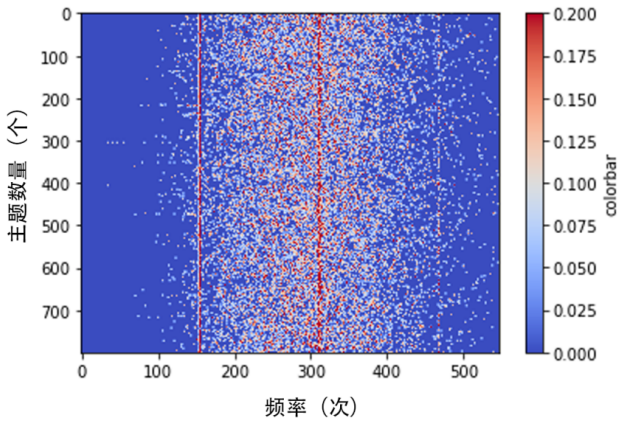

图6 800行×547列的显著周期独热码图示

如上图所示，从帖子样本中提取的800类主题在三年里的演化具备一定的周期特征，独热码图显示这些主题的频率较多地集中分布于7.013天、3.495天、2.333天（分别处于图6从左、中、右依次的三根线所处的横坐标），它们恰好分别对应1周、½周、⅓周，这三种周期长度构成今日头条平台的“主频”。

800个主题在上图所示各频率及其对应的波长上的总体分布，也可看到800个主题的振荡上的共性，以及在7.013天、3.495天、2.333天三个尖峰处所示的“平台主频”，如图7。

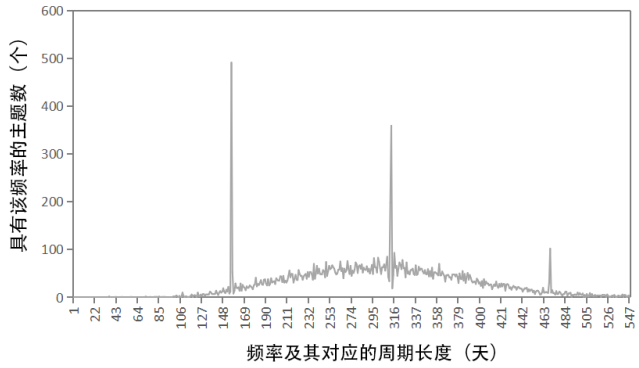

图7 800个主题在各频率上的总体分布

在800个主题中，主题的“最长周期”指标的最小值为2.96天，最大值为30.39天，均值为7.30天；主题的“最短周期”指标的最小值为2.00天，最大值为4.24天，均值为2.39天。主题的周期长度和周期种数描述详见表2所示。

表2 主题的周期长度和周期种数特性描述统计

主题的周期特征	范围	最小值	最大值	均值	标准偏差	方差
周期数	27.0000	2.0000	29.0000	18.3213	4.1445	17.1770
最长周期天数	27.4241	2.9648	30.3889	7.2956	1.7081	2.9180
第2长周期天数	23.1380	2.9096	26.0476	6.2116	1.2122	1.4690
第3长周期天数	22.7917	0.0000	22.7917	5.4671	1.0852	1.1780
最短周期天数	2.2366	2.0037	4.2403	2.3904	0.2364	0.0560
各周期平均长度	9.2857	2.8165	12.1022	4.0376	0.4530	0.2050
各周期长度中位数	0.0000	99999.0000	99999.0000	99999.0000	0.0000	0.0000
各周期长度标准差	8.9949	0.0916	9.0865	1.3184	0.4908	0.2410

3.3 帖子周期与热度计算

3.3.1 数据准备和预处理

这800类的帖子所属的发布者，其用户总数111397个（因为18个版块发帖用户中有重复用户），其中2747个用户缺失粉丝数等数据，剩下的用户的粉丝数x经由：$x_{new}=\log2(x+1)$处理后，分布情况如图8所示。

每条帖子具有点击数（分别为阅读数或播放数），对点击数原始值x进行$\log2(x+1)$转换，然后针对帖子中包括的阅读数或播放数两类，各自通过min-max归

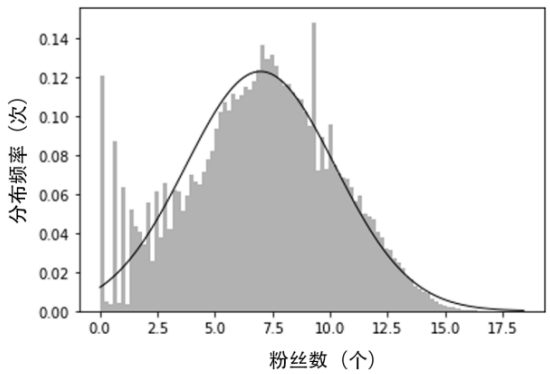

图8 样本用户粉丝数分布直方图

一化处理为[0,1]区间内的值,28473604条样本帖的热度值分布情况如图9所示。

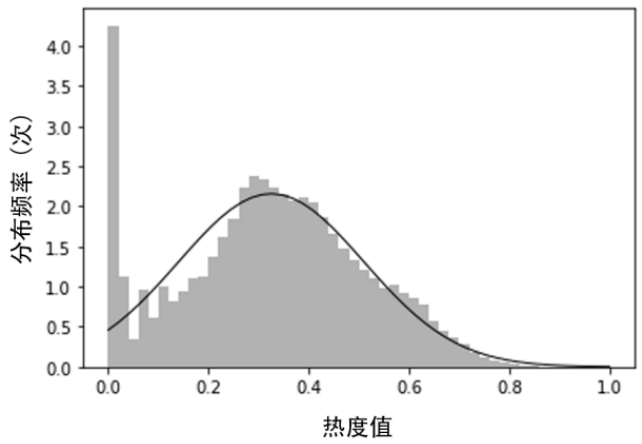

图9 样本帖数据热度值分布直方图

每类主题的热度为该主题中的帖子的热度均值,为了统一口径,每个主题计算热度时一律随机抽取主题中的相同数量的帖子。本文选1408条作为统一数量,因为规模最小的主题内最多包含该数量帖子。800类帖子的热度均值为0.3171,极小值0.01,极大值0.55,标准差0.0883,其总体分布如图10所示。这800类帖子的热度,正是全文分析的因变量。

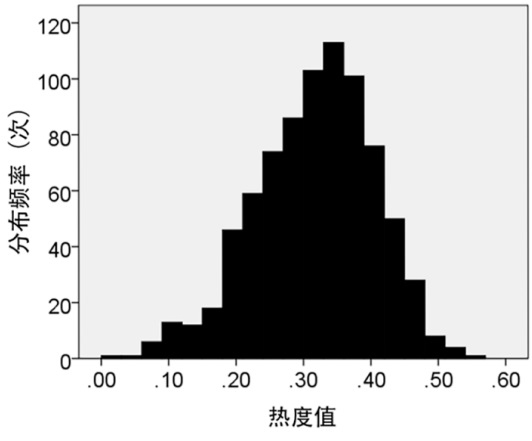

图10 800个主题的热度分布直方图

3.3.2 有效周期长度的特征提取

针对800类主题的547种周期长度进行独立样本t检验,筛选出对热度具有统计上的有效影响的"波长"。对于第n个周期长度,以图6中第n列的数据为待检验因子,也即包含800个0或1的独热码向量;以800类主题的热度值为因变量,通过独立样本t检验考察对n列向量对应的因变量是否存在显著差异。其后从547种周期长度中,筛选出了37种特征频率及其对应的"波长",也就是具有显著性的37种周期长度。经过此特征提取步骤后,用于后续的关键波长的分析以及预测模型的分析。

3.3.3 逻辑回归与对热度有正向影响的周期长度

根据800类主题的热度值高低进行排序,并将这些热度值分为两类,即热度高的百分之五十的类标为2,热度低的百分之五十的类标为1,由此分为高热度与低热度各400类的标签值,并对这两类热度做逻辑回归分析。采取Python中sklearn.linear_model库的LogisticRegression()函数,随机打乱数据,学习所有数据中的80%的类建立逻辑回归模型,并用以预测剩下20%的类。求解优化问题的算法选择newton-cg,各个类别的权重设置为balanced,运算结果的准确率、精确率、召回率、F1分别为:0.6938、0.6928、0.6973、0.6917。该结果也表示出,周期长度作为自变量影响着主题热度。

逻辑回归中,通过回归系数可得到各种周期长度对于主题热度的作用方向(正值表示有正向影响,负值表示有负向影响),如表3所示:

表3 逻辑回归中各周期长度对热度的作用(单位:天)

序号	周期长度(天)	回归系数	序号	周期长度(天)	回归系数
1	10.1296	−1.2972	20	3.4952	−0.4033
2	8.8226	−0.3053	21	3.4295	0.6238
3	7.1503	−0.7565	22	3.4081	−0.1399
4	7.0128	−0.6381	23	3.3975	0.1749
5	6.8375	−0.467	24	3.256	0.3493
6	6.5119	−0.2582	25	3.0644	0.3647
7	6.0442	1.0734	26	3.0389	0.2604
8	5.6684	0.5342	27	2.9019	0.256
9	5.5253	0.4906	28	2.8564	0.483
10	5.47	0.0889	29	2.7557	0.5968
11	5.3627	−0.3592	30	2.7282	1.567
12	5.2344	0.4759	31	2.6235	0.3856
13	5.1362	−0.3925	32	2.4474	−0.928
14	4.4836	0.812	33	2.3326	−0.6645
15	4.3936	0.1612	34	2.3227	−0.3507
16	3.9495	0.0836	35	2.2839	−0.1141
17	3.7595	0.4068	36	2.2792	−0.2334
18	3.6589	0.3406	37	2.1621	−0.3388
19	3.6106	0.7769			

从表3看到,特征提取后对主题热度有正向影响的周期长度有21种,起到抑制作用的有16种。总体而言,能帮助预测主题热度的关键"波长"的周期长度

的分布特征如下：

1）接近½周、1周、3/2周的周期长度或“波长”，对热度形成抑制作用，它们分别对应于或近似于上表中的3.495天、7.013天、10.130天。

2）⅓周（2.3326天）也是抑制热度的重要周期长度之一。

3）各波长表现出抑制热度的2大集中性的“波段”，分别是⅓周左右的波长集群（表3中，2.333天的上下连续几个“波长”）、1周左右的波长集群（表3中，7.013天的上下连续几个“波长”）。

4）抑制主题热度的还有3.4081天和6.8375天2种周期长度，它们之间也有近似的2倍关系。

5）对主题热度具有正向影响的周期长度或“波长”，主要是1周长度以下且分布在除上述抑制区以外的波长范围。其主要特征有：

①是以1周长度内的短周期和“快频率”为主，而不包括1周以上的中、长波段和其对应的“慢频率”；

②是去除⅓周、½周、1周这3种长度所代表的“平台主频”后，剩下的为振荡波长区域。推测其原因，可能是在以1周为基准单位及其½周、⅓周等强关联的平台“主频”下，能符合“主频”的主题较为常见平庸，因而其热度往往不高，而高热度的主题则更需要具有“突破主频”的能力及表现。

③正向影响热度的周期长度形成在[2.6天，3天]短波区段为基准及其对应的[5.2天，6天]区段的倍数关系，例如2.62天与5.23天的波长对、2.73天与5.47天的波长对、2.76天与5.53天的波长对、2.86天与5.67天的波长对、3.04天与6.04天的波长对。这个特征也较为鲜明。但具体原因尚不清楚，不排除是[2.6天，3天]左右的波长区段及其2倍的波长区段在平台舆论中具有某种特殊性。为了后文便于分析方便，本文把[2.6天，3天]的波长区段称为α波长区域，把α波长区域的2倍波长区域所对应的[5.2天，6天]波长区段称为β波长区域。其中α波长区域恰好约处于⅓周和½周的这2个主频中间，β波长区域恰好约处于½周和1周这2个主频中间。后文决策树模型的分析也将会显示，α和β这2个波长区域具有对主题热度预测模型的重要性。

3.3.4 决策树模型与关键周期长度的重要性

采取决策树的方式探索主题周期长度与主题热度间的具体关系，自变量为上述提取后的37种周期长度，因变量为二等分做定序处理后的主题热度。计算工具为sklearn中的DecisionTreeClassifier()函数。衡量分割质量的指标选择“gini”，表示基尼系数。运用网格法确定决策树的最优参数，结果如表4所示。

表4 决策树模型寻优参数示意、范围、最优值

参数名	示意	数值范围	最优值
max_depth	决策树的深度，是控制过度拟合的重要参数	2~10	9
min_samples_split	一个节点进行分枝必须要至少包含的训练样本数	2~10	10
min_samples_leaf	一个叶子节点所需的最小样本数	2~10	9

在最优参数下得到的决策树中，采用DecisionTreeClassifier. feature_importances_得到决策树中各自变量的正态化重要性，其中该指标大于均值的自变量由高到低排序如表5：

表5 自变量重要性

自变量	正态化重要性
7.01282（平台三大主频之一，1周）	0.2829
2.33262（平台三大主频之一，⅓周）	0.1363
2.75567（属于前文分析的α波长区域）	0.1081
5.36275（属于前文分析的β波长区域）	0.0736
2.72818（属于前文分析的α波长区域）	0.0662
3.75945	0.0646
3.49521（平台三大主频之一，½周）	0.0549
5.52525（属于前文分析的β波长区域）	0.0420
3.4081	0.0292

关于决策树模型能有效预测主题热度的周期长度，其主要特征如下：1）重要性最高的2类周期为7.01天、2.33天，恰好分别为1周和⅓周。平台三大主频也即1周、½周、⅓周，均在决策树中具有重要性，其中½周的重要性虽稍低但也居于前7，不可忽视。2）有4个周期长度及其对应的波长入围前文分析的α波长区域和β波长区域，说明这2个区域的重要性。3）其中仍然存在着一些近似的倍数关系，例如2.756天和5.525天。总体而言，决策树模型和对各自变量的逻辑回归形成补充印证，而且进一步显示了平台主频、部分关键周期长度的重要性。

4 结论

主题在网络平台上的演化并非杂乱无序，主题不仅具备“生命周期”，也具备多种演化周期和“波长”。

这些丰富的波长特性,反映着主题演化的“振动”节律,也反过来对主题的热度“能量”产生作用,且具备可预测性,具体展现在以下几点。

1)主题的周期长度是关系到主题热度的重要因素,对于主题的热度具备显著的预测作用,且不同周期长度对于主题热度值的影响程度不同。经由逻辑回归分析后,主题周期长度预测主题热度的准确率、F1值指标均接近0.7,表明主题周期能够影响和预测主题热度,这是易被忽视的自变量维度。已有研究多强调舆论或信息具备生命周期这一特征,较为缺乏对该特征的功能的探讨,本文通过计算验证了主题周期特征的“预测”作用,进一步拓宽了周期研究的维度。

2)周期长度对主题热度的影响有正向和负向。任意主题周期若属于平台“主频”所对应的1周、½周、⅓周这三种周期长度及其周边邻近的波长区域,则对主题热度起到抑制作用。若属于避开这3种主频的[2.6天,3天]区域及其双倍的[5.2天,6天]区域,则对主题热度起到正向作用。结合主题显著周期的独热码图,在检验主题周期“预测”功能的基础上,本文进一步发现分别具备抑制与促进作用的特定周期长度,精确计算出对主题热度作用不同的周期天数,说明主题虽然具备周期特性,但这些周期特性对于主题热度、流行度、受欢迎度的作用不一。研究认为若需在舆论调控实践中运用周期,则应当明确按照不同周期长度的不同功能进行调用。

3)从预测主题热度的角度出发,主题的周期长度具备重要性与次重要性。最为重要的预测指标是1周和⅓周等平台“主频”。这些有规律的“波长”,反映主题在获取高热度能量中的特定节律。平台“主频”在预测主题热度方面具有关键重要性,说明“主频”基本主导了所在平台的信息整体周期演化特征,符合“主频”的主题在平台上具有普遍性,在时间序列上的演变也具有固定周期长度,从而保持规律性、高预测精度的热度特征。

4)主题周期存在长周期与短周期之分,但令主题获得高热度的关键周期以短周期为主。本文通过研究分析了800类主题的所有547类周期长度后,最终筛选出很少部分的真正有效的显著周期,而这些具有影响的周期均为一周内或3/2周内的短周期,大量的长周期则并无显著作用。尤其是对于主题热度具有正向作用的周期长度,更是主要集中在6天以内。推荐算法平台每日过滤、推送海量信息,任何信息在平台中均具备“短、频、快”特征,故而短周期能够令主题多次出现,缩短了主题被用户遗忘的周期,从而有效抓住用户注意力,获得高热度与流行度。

本文的意义和主要创新在于以下方面:首先,在理论方面,本文探索周期性规律对于舆论或信息的影响与预测功能,有助于探索主题“振动”波长和主题“能量”之间的关联机理。其次,在实践应用方面,调控舆论可以从控制主题周期长度入手,从而反向应用于有效引导主题热度。以今日头条的舆论调节为例,经由计算与预测后,若需要令某些主题内容具备高热度,则需要避开1周、½周、⅓周这三种平台“主频”,最好需要具备在[2.6天,3天]和[5.2天,6天]波长区域内的重复频次。借用“谎言重复一千遍成为真理”的表述,那么以何种频率来重复是有现实意义而待检验的问题。

研究的不足和展望主要如下:1)本文提取出37种显著周期长度,也是具备高预测能力的关键周期长度。然而,在所计算的周期长度中,为何是这些周期成为关键周期而非其他周期成为关键周期,即不同周期长度为何会呈现显著的预测能力差异,这是尚待进一步探讨分析的,或许其背后隐藏着信息、文化、社会“呼吸”或气候、地理等其他“节律”的隐在影响。2)研究发现一部分的关键周期长度呈现倍数关系,且具备一定的规律性。然而,这样的倍数关系为何存在,尚待深入探究其可能的形成机理。3)本文发现了短周期在预测主题热度中的关键作用,然而,为何是短周期能够影响主题热度,而长周期不具备显著的影响热度的能力,以及长周期对于主题而言究竟具备什么样的功能与意义,这些也是本文还未能继续探讨的部分。长周期相当于“长波”的光,在实际研究中,长波的震荡也是有用的,例如红光的穿透力比绿光、紫光要强,可为后续研究提供跨学科的类比和某种参鉴。本研究猜测:具备长周期(长波)的主题虽然在舆论的短时间尺度中表现欠佳,但可能在长时间尺度中的“延续力”、“穿透力”会有强化的表现,例如在文化中的延续起伏,这为文化周期研究提供了更多可能。

参考文献(References):

[1] 张文杰, 许门友. 主流媒体引导下公共事件社会舆情泛化特征分析[J]. 情报科学, 2022,40(01): 25-30.

[2] 谢科范, 赵湜, 陈刚,等. 网络舆情突发事件的生命周期原理及集群决策研究[J]. 武汉理工大学学报(社会科学版), 2010, 23(04): 482-486.

[3] Lu X, An J. Evolution analysis of network public opinion

theme based on LDA Model[C]// Proceeding of 2022 4th International Conference on Applied Machine Learning (ICAML), 2022: 396-400.
[4] 王曰芬, 王一山, 杨洁. 基于社区发现和关键节点识别的网络舆情主题发现与实证分析[J]. 图书与情报, 2020(05): 48-58.
[5] Qiao J, Gao Z H, Huang Y R, et al. Analysis on life cycle of network public opinion[C]// Proceeding of 2015 International Conference on Social Science, Education Management and Sports Education, 2015: 725-727.
[6] 匡文波. 论新媒体舆论的生命周期理论模型[J]. 杭州师范大学学报(社会科学版), 2014, 36(02): 112-117.
[7] 昂娟. 突发事件生命周期网络舆论演化及引导研究——以“魏则西事件”和“雷洋事件”为例[J]. 大理大学学报, 2018, 3(05): 56-60.
[8] 陈福集, 张燕. 基于E-Divisive的网络舆情演化分析[J]. 情报杂志, 2016, 35(04): 75-79.
[9] He Y, Li J, Zhu M, et al. Life cycle identification and analysis of microblog hot topics[C]// Proceeding of 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2018, 2: 156-159.
[10] 毛太田,蒋冠文,李勇, 等. 新媒体时代下网络热点事件情感传播特征研究[J]. 情报科学, 2019, 37(04): 29-35+96.
[11] Zhang L, Wei J, Boncella R J. Emotional communication analysis of emergency microblog based on the evolution life cycle of public opinion[J]. Information Discovery and Delivery, 2020, 48(3): 151-163.
[12] Downs A. Political theory and public choice[M]. Cheltenham and Camberley: Edward Elgar Publishing, 1998.
[13] 王积龙, 张妲萍, 李本乾. 微博与报纸议程互设关系的实证研究——以腾格里沙漠污染事件为例[J]. 新闻与传播研究, 2022, 29(10): 80-93+127-128.
[14] 李永宁, 吴晔, 张伦. 2010-2016年公共议题的公众注意力周期变化研究[J]. 国际新闻界, 2019, 41(05): 27-38.
[15] Shih T J, Wijaya R, Brossard D. Media coverage of public health epidemics: linking framing and issue attention cycle toward an integrated theory of print news coverage of epidemics[J]. Mass Communication & Society, 2008, 11(2): 141-160.
[16] Saleiro P, Soares C. Learning from the news: predicting entity popularity on twitter[C]// Advances in Intelligent Data Analysis XV: 15th International Symposium, 2016: 171-182.
[17] Su Q, Yan S, Wu L, et al. Online public opinion prediction based on a novel seasonal grey decomposition and ensemble model[J]. Expert Systems with Applications, 2022, 210: 118341.
[18] 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(31): 159-161+174.
[19] 张虹, 赵兵, 钟华. 基于小波多尺度的网络论坛话题热度趋势预测[J]. 计算机技术与发展, 2009, 19(04): 76-79.
[20] 梁芷铭. 基于新浪微博的网络信息生命周期实证研究[J]. 新闻界, 2014(03): 60-64+69.
[21] 江燕青, 许鑫. 半衰期视角的微博信息老化研究——以高校官方微博为例[J]. 图书情报知识, 2016(02): 92-100.
[22] 马费成, 高静. Web2.0信息半衰期影响因素实证研究——以社会书签网站为例[J]. 情报理论与实践, 2010, 33(11): 1-6.
[23] Mikolov T, Chen K, Corrado G S, et al. Efficient estimation of word representations in vector space[DB/OL]. arXiv: 1301.3781,2020.
[24] 李晓, 解辉, 李立杰. 基于Word2vec的句子语义相似度计算研究[J]. 计算机科学, 2017, 44(09): 256-260.
[25] 唐明, 朱磊, 邹显春. 基于Word2Vec的一种文档向量表示[J]. 计算机科学, 2016, 43(06): 214-217+269.
[26] Xing C, Wang D, Zhang X, et al. Document classification with distributions of word vectors[C]// 2014 Annual Summit and Conference Asia-Pacific Signal and Information Processing Association, 2014: 1-5.
[27] 段旭磊, 张仰森, 孙祎卓. 微博文本的句向量表示及相似度计算方法研究[J]. 计算机工程, 2017, 43(5): 143-148.
[28] Shen D, Wang G, Wang W, et al. Baseline needs more love: on simple word-embedding-based models and associated pooling mechanisms[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018 : 440-450.
[29] 张恒, 乔国伟, 张秋良. 基于小波分析的我国森林草原火灾周期震荡研究[J]. 林业工程学报, 2019, 4(02): 139-145.
[30] 刘旭, 董剑希, 姜珊, 等. 基于小波分析的我国台风风暴潮直接经济损失周期分析及预测[J]. 海洋学报, 2023, 45(07): 137-146.
[31] 徐翔, 杨航宇, 徐舟爽, 等. 社交网络的情绪波动周期性及应对策略——基于新浪微博样本的大数据分析[J]. 新闻与写作, 2021(08): 22-32.
[32] McMillan T C, Rau G C, Timms W A, et al. Utilizing the impact of earth and atmospheric tides on groundwater systems: a review reveals the future potential[J]. Reviews of Geophysics, 2019, 57(2): 281-315.
[33] 魏凤英. 现代气候统计诊断与预测技术(第2版)[M]. 北京: 气象出版社, 2007.
[34] 赵振华, 罗振江, 黄林显, 等. 基于小波分析的济南西郊地下水位对降雨响应机制研究[J]. 中国岩溶, 2023, 42(05): 931-939.

编辑:赵志军