基于 ARIMA 的短视频数据分析方法

左源[1] 王一可*[2] 尚芳[2] 张凌云[2] 王仕超[2]

1 沈阳鑫印象科技有限公司，辽宁沈阳，110031；

2 沈阳城市学院，辽宁沈阳，110112；

摘要：移动互联网发展迅猛，短视频平台深度融入公众生活，成为关键信息传播与社交互动媒介。以抖音为代表，借创新算法与精准推荐拓展全球用户，构建活跃生态。其功能多元，为用户提供展示空间，助力个体价值传播；也是商业推广、文化交流的核心途径。基于此，本研究聚焦大流量主播数据，经深度剖析，能明晰平台影响力分布，为创作者、营销者、政策制定者提供依据，推动短视频行业健康发展。

关键词：数据分析；数据挖掘；人工智能

DOI：10.69979/3041-0673.25.07.007

1 研究现状

1958 年，IBM 的 Hans Peter Luhn 首提商业智能概念，助力企业决策科学化。1970 年，Edgar F. Codd 提出关系数据库模型，大幅优化数据管理。1998 年，谷歌搜索亮相与数字存储成本降低，推动互联网数据量激增，为数据分析技术发展赋能。2001 年，Doug Laney 给出大数据 “3V” 定义，促进相关技术研究[1,2]。2005 年，Doug Cutting 和 Mike Cafarella 开发 Hadoop，推动大数据技术商业应用。2012 年，社交媒体数据分析技术成熟，企业借此开展市场与用户研究。2016 年，物联网、VR 和 AR 技术与数据分析融合，发挥不同效用[5]。2021 年，人工智能、机器学习在数据分析中应用加深，数据湖等技术提供灵活存储分析方式[6]。近年，人工智能与数据分析深度融合，如短视频平台借主播数据分析优化推荐算法，推动线上经济发展。

2 具体算法

在当下用户行为分析这一兼具高度复杂性与重大研究价值的学术领域，致力于解析用户行为模式及内在规律的相关算法呈现出丰富多元的格局。从宏观视角审视，可清晰归纳为两类主要范畴。

第一类是以用户对内容类别的偏好倾向为基础构建的分析体系。在此体系中，研究人员运用系统且全面的数据采集策略，广泛收集用户在不同内容类别层面所产生的关注行为数据，诸如用户对特定内容类别的主动订阅、收藏等操作记录；浏览行为数据，涵盖在各类内容页面的停留时长、页面跳转轨迹等信息；以及互动行为数据，包括点赞、评论、分享等操作详情。随后，借助数据整理技术，如采用数据清洗手段以去除异常值和重复数据，运用聚类分析等归纳方法将具有相似行为模式的用户归为一类。通过对这些数据的深度处理与挖掘，得以深入洞察用户在内容选择过程中的兴趣偏好。

第二类则是依托用户内容消费深度学习算法开展数据挖掘与建模工作。该算法凭借深度学习技术强大的数据表征学习能力，对用户在内容消费全流程中产生的多元异构数据进行深度解析。这些数据不仅包含能够直观反映用户对内容专注程度的浏览时长指标，还涵盖点击顺序数据，从中可挖掘用户的浏览逻辑及兴趣转移路径，以及搜索记录数据，通过对搜索关键词的语义分析，能够精准洞察用户的潜在需求与兴趣方向。

而本文通过对用户活跃时间度的深入分析来精准推断用户行为。具体而言，所运用的核心算法为 ARIMA（自回归积分滑动平均）算法。该算法在时间序列分析领域拥有坚实的理论基础与广泛的应用实践经验。其通过自回归（AR）部分，构建当前时间点数据与过去若干时间点数据之间的线性关联，以此捕捉时间序列数据中的趋势性特征；借助积分（I）操作，将非平稳时间序列转化为平稳序列，有效应对用户活跃时间数据可能存在的趋势性与季节性波动；利用滑动平均（MA）部分，对时间序列中的随机噪声进行建模与平滑处理。通过这一综合性的算法机制，ARIMA 算法能够高效捕捉并精确分析用户活跃时间数据中的趋势性、季节性及周期性等关键特征。基于用户活跃时间的历史数据，运用该算法能够精准预测用户未来在不同时间维度下的行为模式与活跃规律，为深入理解用户行为的时间动态特性提供了全新视角与有力工具，对于相关领域的理论研究深化

与实际应用拓展均具有显著的推动作用。

ARIMA（Autoregressive Integrated Moving Average）算法作为一种在时间序列分析和预测领域具有广泛影响力的经典统计模型，在众多学科与实际应用场景中发挥着关键作用。其基本思想蕴含着对时间序列数据内在规律的深刻洞察，即充分挖掘数据本身所携带的历史信息，以此作为预测未来趋势的有力依据。

ARIMA 算法的性能优越之处在于其巧妙地结合了自回归（AR）、差分（I）和移动平均（MA）三种主要成分。自回归成分通过构建当前值与过去值之间的线性关系，捕捉时间序列中的长期依赖特征；差分操作则用于将非平稳时间序列转化为平稳序列，这是因为在实际应用中，许多时间序列数据呈现出非平稳特性，如趋势性、季节性等，直接对其进行建模会导致模型参数估计不准确以及预测效果不佳，而差分能够有效消除这些非平稳因素的影响；移动平均成分则侧重于对时间序列中的随机波动进行建模，通过考虑过去若干期的误差项来平滑数据，提高预测的稳定性和准确性。

ARIMA 算法通常表示为 ARIMA(p, d, q)，其中：

p：自回归（AR）项的阶数。

d：差分（I）的阶数，用于使时间序列平稳。

q：移动平均（MA）项的阶数。

自回归部分基于时间序列自身的滞后值（即过去的值）来预测当前值。AR 模型的阶数为 p，表示当前值依赖于过去的 p 个值。其数学表达式为：

$$Yt=c+\phi 1Yt-1+\phi 2Yt-2+\cdots+\phi pYt-p+\epsilon t$$

（公式 1）

其中：

Yt 是当前时间点的值。c 是常数项。$\phi 1,\phi 2,\cdots,\phi p$ 是模型参数。ϵt 是误差项，假设为白噪声。

差分操作用于将非平稳时间序列转换为平稳时间序列。平稳时间序列的均值、方差和协方差不随时间变化，更适合进行统计分析。差分的阶数为 d，表示对时间序列进行 d 次差分。例如：

一阶差分：

$$\Delta Yt=Yt-Yt-1$$ （公式 2）

二阶差分：

$$\Delta 2Yt=\Delta Yt-\Delta Yt-1$$ （公式 3）

通过选择合适的 d，可以使时间序列达到平稳状态。

移动平均部分基于过去的误差项来预测当前值。MA 模型的阶数为 q，表示当前值依赖于过去的 q 个误差项。其数学表达式为：

$$Yt=\mu+\epsilon t+\theta 1\epsilon t-1+\theta 2\epsilon t-2+\cdots+\theta q\epsilon t-q$$

（公式 4）

其中：

μ 是序列的均值。$\theta 1,\theta 2,\cdots,\theta q$ 是模型参数。ϵt 是误差项。

ARIMA 模型通过结合 AR、I 和 MA 三个部分，对时间序列进行建模和预测。其完整的表达式为：

$$\Delta dYt=c+\phi 1\Delta dYt-1+\cdots+\phi p\Delta dYt-p+\epsilon t+\theta 1\epsilon t-1+\cdots+\theta q\epsilon t-q$$ （公式 5）

其中，ΔdYt 表示经过 d 次差分后的时间序列。

3 实验结果

本研究深度聚焦于视频应用数据的全方位、精细化剖析，致力于达成对抖音主播多层面特征的深刻洞察，其中涵盖用户特征、内容影响力以及地理分布格局等关键领域。研究的核心宗旨在于深度探究软件平台所架构的社交影响力结构体系，细致剖析主播在内容创作以及与用户展开互动的过程中，是如何凭借一系列策略与行为，逐步构建起自身在平台生态中的影响力，并实现影响力的持续拓展与深化。

在对视频平台中主播数据进行系统性、多维度分析的进程中，除了能够清晰勾勒出主播在性别、影响力层级、类型划分以及地理位置等方面的分布特征外，还将引入科学、严谨的评估体系对其内容质量及所产生的社会影响力展开深入评估。对于内容质量，将从内容创新性、专业性、情感共鸣度等多个维度进行考量，运用文本分析、图像识别、语义理解等技术对视频内容进行拆解与评估。在社会影响力评估方面，不仅关注主播对用户消费行为、观念认知等方面的影响，还将探究其在社会文化传播、舆论引导等宏观层面所发挥的作用。

从研究成果的应用价值视角审视，对于内容创作者而言，本研究详尽的分析结果能够为其提供优化内容创作策略的具体方向。例如，依据不同类型内容的点赞、评论等反馈数据，创作者可以精准把握用户兴趣点，调整内容选题与表现形式，进而提升自身在平台中的影响力。对于品牌营销人员，本研究能够为其筛选契合品牌定位、具备高营销价值的合作主播提供坚实的数据支撑。通过对主播影响力层级、粉丝画像以及内容质量等多维度数据的综合考量，品牌营销人员可以制定精准有效的营销策略，提高营销资源的投入产出比。对于政策制定

者，本研究有助于其深入了解社交媒体的影响力机制，明晰不同类型主播在信息传播、文化引领等方面的作用与影响。基于这些研究成果，政策制定者能够制定更为科学、合理的政策法规，规范社交媒体平台的运营秩序，推动社交媒体行业朝着健康、有序的方向蓬勃发展。

实验首先记录用户每次打开应用、点赞、评论、分享等操作的时间戳。并将一天分为若干个时间区间，如，可以将一天者更粗粒度地分为早（6-12 点）、中（12-18 点）、晚（18-24 点）、夜（0-6 点）四个区间。接着在每个时间区间内，统计用户的操作次数。最后找到活跃指数最高的时间区间，即为用户的活跃高峰。图 3.1-图 3.2 为本文的可视化结果展示

图 3.1 短视频平台主播各类型点赞汇总图

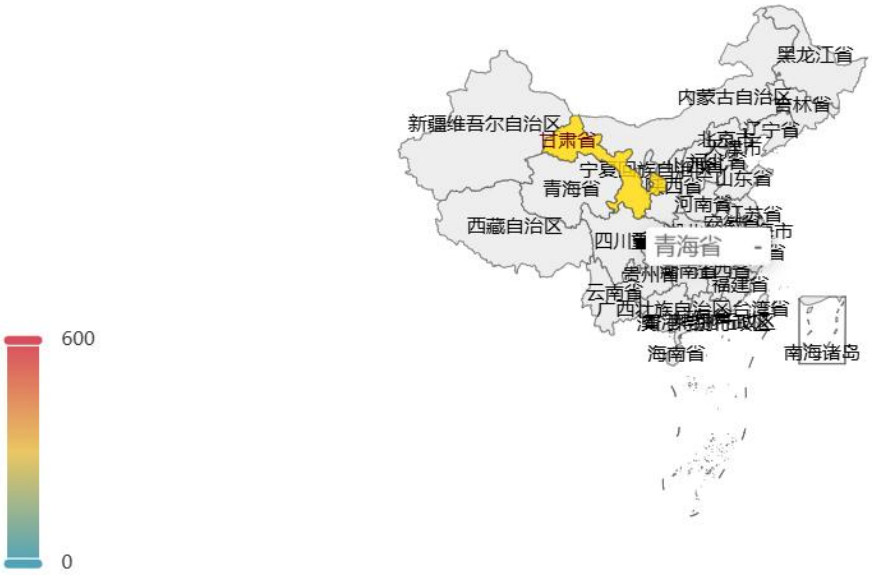

图 3.2 短视频平台主播省份分布

4 结论

后续研究将围绕关键环节推进推荐系统构建。在数据收集与处理及特征工程方面，通过多元渠道采集用户行为与内容数据，运用关联规则挖掘等技术构建精准用户画像；利用 NLP、计算机视觉等技术综合提取内容、协同过滤、上下文特征，为推荐精准度奠基。 推荐算法选型上，深入剖析协同过滤、基于内容、混合推荐及深度学习模型等主流算法，明确其在不同场景的适用性。模型训练、评估与优化时，按分层抽样等原则划分数据集，用随机梯度下降等算法训练模型，借准确率等指标评估，引入在线学习等优化策略。 系统部署与监控及用户体验优化环节，遵循容器化等理念将模型部署至生产环境，借日志分析等技术监控性能；设计个性化推荐界面，依据用户反馈形成闭环优化，提升用户体验。

参考文献

[1]王茜.发展与挑战:大数据时代的新闻传播研究方法[C]//第二届上海交通大学——ICA 国际新媒体论坛.2015.
[2]杨晓晗,程国振,刘文彦,等. 基于深度学习的拟态裁决方法研究[J]. 通信学报,2024,45(2):79-89.
[3]张良. 面向新浪微博的水军识别技术的研究与实现[D]. 湖南:国防科学技术大学,2015.
[4]黎玲萍,毛克彪,付秀丽,等. 国内外农业大数据应用研究分析[J]. 高技术通讯,2016,26(4):414-422.
[5]任瑞龙. 大数据在生活中的应用[J]. 中国科技信息,2018(3):142.
[6]周林.大数据技术在物流管理中的应用研究[J].产品可靠性报告,2023(1):71-73
[7]刘世哲.湖仓一体大数据 驱动业务发展新格局[J].中国农村金融,2023(14):94-95
[8]吕本富,陈健.大数据预测研究及相关问题[J].科技促进发展,2014,10(1):60-65
[9]潘可. 面向多源社交网络的用户兴趣爱好特征分析与推荐技术研究[D]. 杭州电子科技大学,2018.
[10]刘丽峰. 基于大数据的网络信息挖掘与用户行为分析[J]. 信息记录材料,2024,25(8):162-164.

作者简介：左源（1995 年 12 月），男，满族，辽宁丹东人，软件工程师，本科学士，沈阳鑫印象科技有限公司，研究方向：数据分析，数据挖掘
通讯作者：王一可（1996 年 2 月），女，汉族，辽宁沈阳人，人工智能专业副主任/讲师，硕士研究生，沈阳城市学院，深度学习与图像处理。