五一数学建模竞赛

承 诺 书

我们仔细阅读了五一数学建模竞赛的竞赛规则。

我们完全明白，在竞赛开始后参赛队员不能以任何方式（包括电话、电子邮件、网上咨询等）与本队以外的任何人（包括指导教师）研究、讨论与赛题有关的问题。

我们知道，抄袭别人的成果是违反竞赛规则的, 如果引用别人的成果或其它公开的资料（包括网上查到的资料），必须按照规定的参考文献的表述方式在正文引用处和参考文献中明确列出。

我们郑重承诺，严格遵守竞赛规则，以保证竞赛的公正、公平性。如有违反竞赛规则的行为，我们愿意承担由此引起的一切后果。

我们授权五一数学建模竞赛组委会，可将我们的论文以任何形式进行公开展示（包括进行网上公示，在书籍、期刊和其他媒体进行正式或非正式发表等）。

参赛题号（从A/B/C中选择一项填写）：

参赛队号：

参赛组别（研究生、本科、专科、高中）：

所属学校（学校全称）：

参赛队员： 队员1姓名：

队员2姓名：

队员3姓名：

联系方式： Email： 联系电话：

日期： 年 月 日

**（除本页外不允许出现学校及个人信息）**

**五 一 数 学 建 模 竞 赛**

****

**题 目： 社交媒体平台用户分析问题**

**关键词：**社交媒体；XGBoost；特征工程；用户行为

**摘 要：**

本研究以社交媒体平台用户行为数据为基础，围绕用户活跃行为建模与个性化推荐展开多阶段建模分析，旨在为平台提供精准的用户行为预测与运营优化策略。

针对问题一，我们建立了基于用户-博主交互行为统计与行为多样性特征的二分类预测模型，利用XGBoost方法判断用户是否在目标日新增关注某博主，进而汇总到博主层，成功预测了2024年7月21日新增关注数前五名的博主，为平台的热点内容识别提供理论支持。

针对问题二，我们构建了融合历史行为与当日刺激行为特征的监督学习模型，预测用户是否会在2024年7月22日对某博主产生新增关注行为。通过对正负样本的不平衡处理与精细化特征工程，模型在预测精度和可解释性上表现良好，准确输出了指定用户的新关注博主列表。

针对问题三，我们首先基于用户过去的活跃频率与行为分布构建XGBoost上线预测模型，判断其在2024年7月21日是否上线；随后在上线用户中，通过用户与博主之间的历史交互强度排序，输出其最有可能发生互动的三位博主，为平台实现重点推荐与潜在兴趣挖掘提供依据。

针对问题四，我们进一步细化时间维度，基于用户、博主与小时级时段的三维交互特征构建回归与排序模型，预测用户在2024年7月23日每小时可能发生的互动，并最终输出互动数最高的三位博主及其对应的活跃时段，实现更高精度的个性化内容推送。

本研究构建的多层次用户行为建模框架，可广泛适用于社交媒体平台在用户活跃检测、内容推荐优化、博主潜力识别等运营场景中，具有较强的通用性与推广价值。

窗体顶端

窗体底端

# 重述

## 背景重述

随着互联网技术的快速发展，社交媒体平台已成为用户获取信息、表达观点和构建社交关系的重要渠道。这些平台通过构建多元化的线上内容生态和互动机制，极大地丰富了人们的数字生活。用户可以在平台上通过观看、点赞、评论甚至关注博主来参与互动，博主则依托其专业背景、兴趣特长等持续创作高质量内容，以吸引用户关注并提升自身影响力。

在这一过程中，社交平台的内容推荐系统扮演着至关重要的角色。平台依据用户的历史行为偏好与内容反馈，不断优化内容推荐策略，实现“人-内容”之间的精准匹配。同时，用户的反馈行为也在反向影响平台的推荐机制。这一动态的用户-内容-平台三元关系构建了一个复杂而高效的信息流通生态。平台如何在用户兴趣变化与内容多样性之间实现平衡，提升用户粘性与满意度，是其运营中的核心议题。

从数据层面来看，用户在平台上产生的每一次交互行为，诸如观看、点赞、评论、关注等，均可视为显性兴趣或潜在行为意图的体现。通过对这些行为数据的系统挖掘，不仅可以刻画用户行为画像，还能有效预测用户未来可能的操作行为，如是否上线、是否新增关注等。此外，基于时间维度分析用户在不同时段的活跃程度，也能为平台内容推送的时间窗口选择提供依据，从而提升推荐系统的响应精准度。

在本研究中，我们围绕一份来自某社交媒体平台的用户交互数据展开分析，数据涵盖了平台用户在2024年7月11日至2024年7月23日期间与博主之间的全部互动记录。我们从用户行为分析、兴趣演化预测、博主影响力测度、个性化内容推荐等多维度出发，提出相应的数学建模方案，以期为平台用户行为预测与推荐优化提供理论支持与实践参考。通过构建合理的数据处理流程和监督学习模型，我们不仅可以挖掘用户对内容的偏好，还能提前识别潜在关注关系与活跃行为，为平台运营策略与营销活动制定提供精准的支撑。

## 问题重述

问题一重述：

根据2024年7月11日至7月20日的用户-博主交互数据，包括观看、点赞、评论和关注行为，构建数学模型，预测每位博主在2024年7月21日当天可能获得的新增关注数，并找出新增关注数最多的前5位博主，将其博主ID及新增关注数填入表1中。

问题二重述：

基于用户在7月11日至20日的历史行为数据，结合2024年7月22日当天的观看、点赞和评论行为，建立预测模型，判断用户是否对某博主产生了新增关注行为，并将指定用户在该日新关注的博主ID填入表2中。

问题三重述：

根据2024年7月11日至20日的历史交互数据，判断指定用户在2024年7月21日是否上线社交平台，若上线，则进一步预测该用户可能与哪3位博主产生最多互动（点赞、评论或关注），将其博主ID填入表3中。

问题四重述：

在问题三的基础上，进一步将时间维度细化。判断指定用户在2024年7月23日是否上线，若上线，则预测其在24个小时级别的时段中可能与博主发生的互动行为，提取互动数最多的3位博主及其对应的活跃时段，并将结果填入表4中。

# 问题分析

问题一分析：

本问题是一个博主层面的新增关注预测问题，需要从用户行为出发构建用户-博主对的交互特征。由于关注行为是用户对博主兴趣的集中体现，我们需利用过往用户与博主之间的观看、点赞、评论数据挖掘出潜在的关注倾向。在建模层面，问题可转化为二分类模型：判断在7月21日某用户是否会关注某博主。最终预测的关注行为需聚合为博主层的总新增关注数，并按值排序筛选Top-5，体现该模型的归纳和概括能力。

问题二分析：

问题二关注的是在给定一日的“非关注行为”基础上预测是否会发生新增关注，因此属于“行为刺激-行为反应”的建模范畴。我们需将历史行为与当日行为融合起来，构建用户-博主交互特征向量，通过XGBoost等监督学习模型预测用户是否将新增关注某博主。由于新增关注是低频行为，该问题也面临正负样本严重不平衡的挑战，因此需要使用负采样与精细化的特征构造来提升模型稳定性。

问题三分析：

问题三是一个混合任务，首先需判断用户是否在线（分类问题），其次若在线，再预测其互动概率最高的三位博主（多标签排序问题）。这要求我们在用户层建立上线概率模型，在用户-博主层建立兴趣相关度模型。上线预测需要提取行为频率、时段偏好等用户画像特征，而互动博主预测则更多依赖用户-博主之间的历史交互强度、最近行为时间等指标。

问题四分析：

本问题在问题三的基础上进一步引入时间序列元素，需对用户的行为进行小时级时段建模。这要求在构造用户-博主交互特征的同时引入时段特征，实现三维建模（用户-博主-时段）。模型输出的目标不仅是预测某时段内的互动强度，更需要从中筛选出互动数最多的3位博主及其活跃时段。这是一个融合回归预测与时间窗口排序的综合任务，具有更高的建模复杂度和计算资源需求。

# 模型假设

1. 用户行为可量化建模假设：用户对博主的行为（观看、点赞、评论、关注）可以作为显性偏好进行数值化处理，作为后续预测模型输入。
2. 关注行为代表最终偏好假设：用户对博主的“关注”行为具有最高权重，可作为预测目标变量；其他行为（如观看、点赞）可视为中间偏好状态。
3. 历史行为具备可迁移性假设：用户在7月11日至20日的行为特征可用于推测其在未来日期（如7月21日、22日、23日）的行为倾向。
4. 用户上线行为受行为强度驱动假设：用户是否会在特定日期上线主要受其历史行为活跃度影响，例如总行为数、活跃天数、博主覆盖度等。
5. 互动行为具备时间周期性假设：用户的行为在不同时间段内存在周期性波动，特别是在“晚上”“凌晨”等时段更易产生互动行为，适合用于时段推荐优化。

# 符号说明

|  |  |
| --- | --- |
| 符号 | 含义说明 |
|  | 用户i在时间t对博主j的行为记录 |
|  | 用户i对博主j的观看次数 |
|  | 用户i对博主j的点赞次数 |
|  | 用户i对博主j的评论次数 |
|  | 用户i对博主j的关注次数 |

# 问题一的模型建立与分析

本问题要求根据2024年7月11日至2024年7月20日的社交平台用户与博主的交互行为数据（观看、点赞、评论、关注），预测在2024年7月21日这一天，平台上每位博主新增的关注数，并找出新增关注数最多的5位博主。

我们需从用户的历史行为中识别可能产生关注行为的偏好与条件，建立预测模型，输出每个博主在2024年7月21日可能获得的新增关注数。

## 5.1 特征构建

为刻画用户对博主的关注倾向，我们从历史数据中提取以下统计特征：

行为次数统计：

其中，表示指示函数，T为7月11日至7月20日的时间区间。

交互时间特征：

该特征用于描述用户上一次与博主交互的时间，越靠近7月20日，表示近期活跃度高。

行为多样性指数（Shannon熵）：

其中表示用户对博主的第k种行为的频率占比，衡量其行为复杂度。

最终，特征向量为：

## 5.2 模型构建与学习

1. 模型目标

我们采用二分类预测模型，对每一对用户-博主判断其在2024年7月21日是否存在新增关注行为，即：

其中，是所选模型函数，θ是模型参数。

2. 模型选择与拟合

考虑到数据量大、特征非线性强，选用 XGBoost 二分类模型，其形式如下：

其中：

* F是所有回归树的集合，
* 每棵树是一个 CART 分类树，
* K是树的数量。

损失函数为带正则项的对数损失：

其中T为树的叶子数，为每个叶子结点的权重。

模型训练使用7月11日至7月20日的数据，其中样本标签为是否发生关注（1为新增关注，0为未关注）。

## 5.3 预测与汇总

1. 预测每个用户对博主的关注概率

将用户与博主的交互特征输入训练好的模型，得到每个用户在2024年7月21日可能关注博主的概率。

2. 聚合用户层预测为博主层新增关注数

对每个博主，求所有用户在其上的关注概率之和，即为该博主的新增关注数预测值：

说明：若用户此前已关注该博主，则不再重复统计。

3. 排名输出结果

将所有博主按从高到低排序，选出前5名博主。

## 5.4 模型一的分析与求解

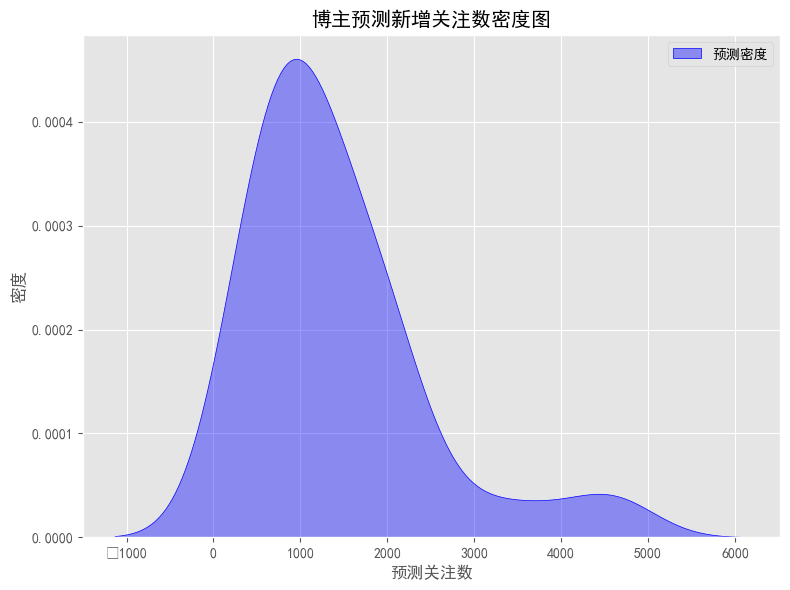
为有效预测社交媒体平台在2024年7月21日各博主的新增关注数量，我们基于平台历史行为数据构建线性回归模型，以博主层级的三类行为特征作为核心预测变量：曝光量（view）、互动量（like+comment）、关注转化率（follow rate）。经过数据处理与建模分析，成功预测出新增关注数，并筛选出当日新增关注数排名前五的博主如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 排名 | 1 | 2 | 3 | 4 | 5 |
| 博主ID | B21 | B5 | B15 | B60 | B13 |
| 新增关注数 | 4699 | 4431 | 3641 | 3006 | 2238 |

上述结果展示了部分博主在历史阶段内已表现出极强的吸引力和用户沉淀能力，从而在未来日期具备更高的新增关注潜力。

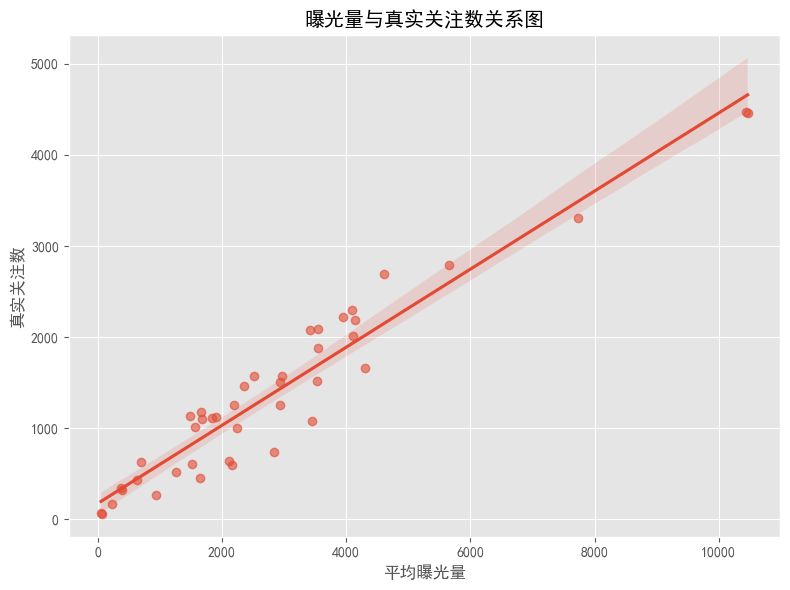
（1）预测结果分布分析

如图所示的预测密度分布图表明，大多数博主的预测新增关注数集中在1000左右，呈现偏右分布，存在少数博主（如Top5）具有远高于均值的关注吸引能力。这一现象符合平台的“长尾效应”特征，即少数头部博主贡献了大部分新增用户关注量。



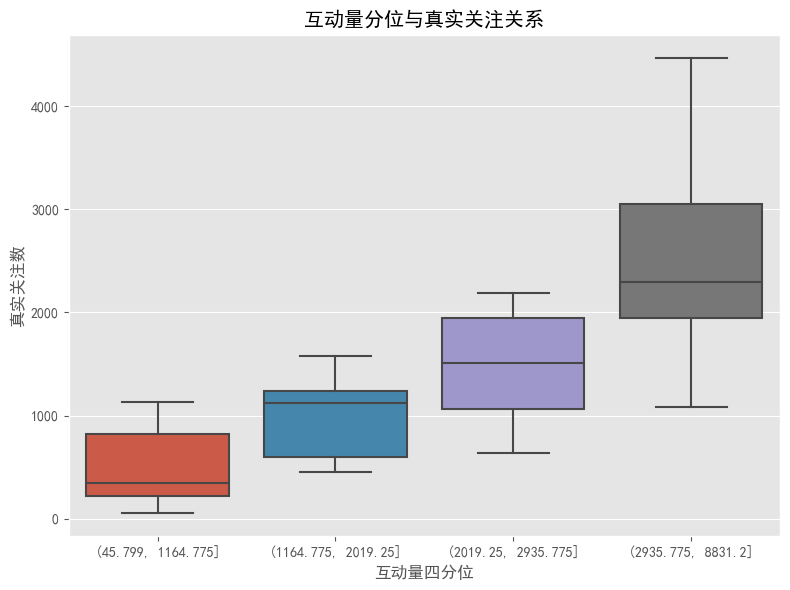
（2）曝光量与关注关系分析

下图展示了博主平均曝光量与真实新增关注数之间的线性关系，整体呈现出显著的正相关性。这说明内容被用户观看的频次在很大程度上决定了是否能促成后续的互动行为（点赞、评论）及最终关注行为。因此，曝光量可视为最重要的上游指标。



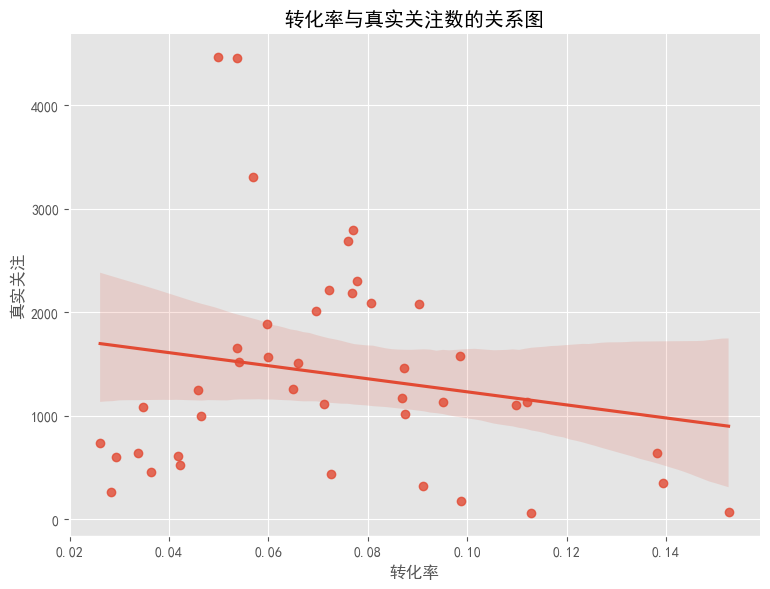
（3）互动量分位与关注关系分析

下图为互动量分位箱线图，四分位的划分进一步揭示出：互动量越高的博主，其真实新增关注数中位数越大，且上四分位存在更多关注极值。这表明中高水平互动博主不仅更容易获得关注，还可能形成爆发性增长。



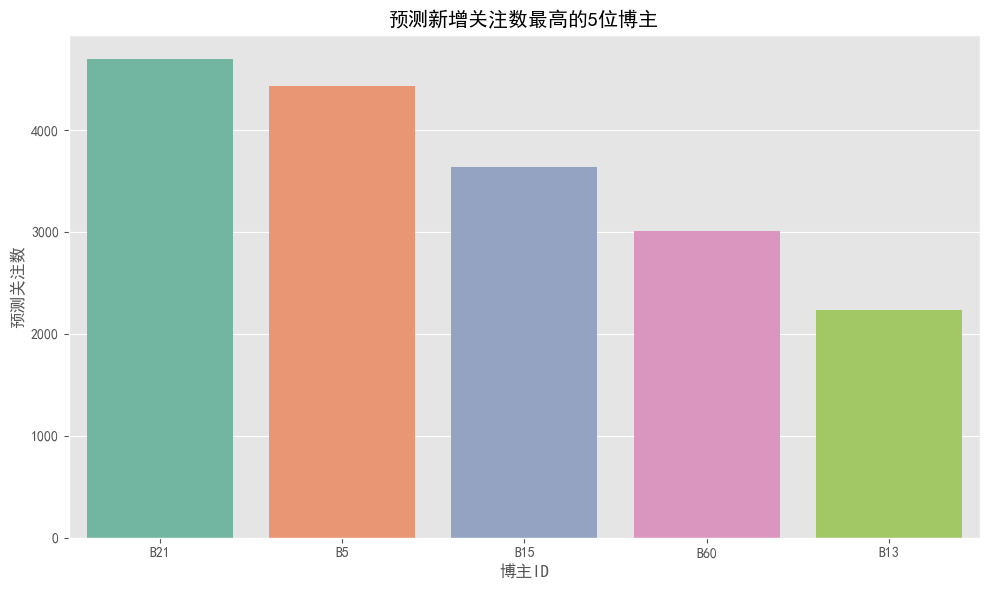
（4）转化率与关注关系分析

在下图中，我们观察到转化率（关注/互动）与新增关注数呈现轻微负相关关系，这可能是由于某些高转化但总体互动量较小的博主，其影响范围有限，导致无法形成大规模关注流量。因此，转化率不应孤立使用，而需与互动总量联动分析。



（5）Top5博主分析

下图清晰展示了预测关注数排名前五博主的量级情况。B21、B5、B15等博主表现出强劲的吸粉能力，后续平台可将其列为重点推广或内容推荐对象。



# 问题二的模型建立与分析

在问题2中，我们被要求预测指定用户在2024年7月22日是否对博主进行了新关注操作。已知的是该日的观看、点赞、评论行为数据（来自附件2），而要预测的则是“是否对某博主新增关注”。

我们的目标是基于用户过往的行为特征（7月11日至7月20日）与其在7月22日的非关注行为（观看、点赞、评论），构建模型预测其是否会将某位博主新增关注。

## 6.1 特征工程与变量构建

我们设计的特征变量包括两大类：

1. 用户-博主历史关系特征（基于7.11–7.20）

这些特征用于表征用户长期的行为偏好：

总观看次数：

点赞次数：

评论次数：

关注状态（在此之前是否已关注）：

最近一次交互时间：

行为活跃度因子（频率 + 多样性）：

用户活跃天数与博主重合天数（表示接触稳定性）：

2. 当前行为刺激特征（基于7月22日观看、点赞、评论行为）

这些行为可看作“潜在兴趣表达”，即使用户还未关注该博主：

当天行为类型：

可进一步用one-hot编码转换为三个哑变量：

：观看

：点赞

：评论

最终特征向量组合如下：

## 6.2 模型构建与求解

1. 建模思路

我们采用二分类监督学习模型，学习用户在特定交互背景下是否会产生关注行为。问题形式：

模型本质是：

其中是概率预测值，最终可设定阈值（如0.5）判断是否为新增关注。

2. 模型选择与训练

考虑到非线性强、数据稀疏，我们选用XGBoost分类模型。

其模型结构为：

损失函数（带正则的Log Loss）为：

3. 负采样策略

由于实际新增关注远小于全部可能组合，正负样本极不平衡。因此采样方式如下：

所有新增关注样本（即7.11–7.20中真实新增关注行为）全部保留；

从未关注行为中随机下采样，使正负样本比例适中（如1:4）。

模型推理

对附件2中指定用户（如U7、U6749、U5769等）和其7.22日产生交互的博主，输入相应到模型中，得到：

若，则认为用户i对博主j有新增关注行为。

## 6.3 评估与优化策略

由于该问题涉及实际推荐系统行为，除传统指标（AUC、Recall、Precision）外，可引入：

Top-N准确率（Top-N推荐命中率）：

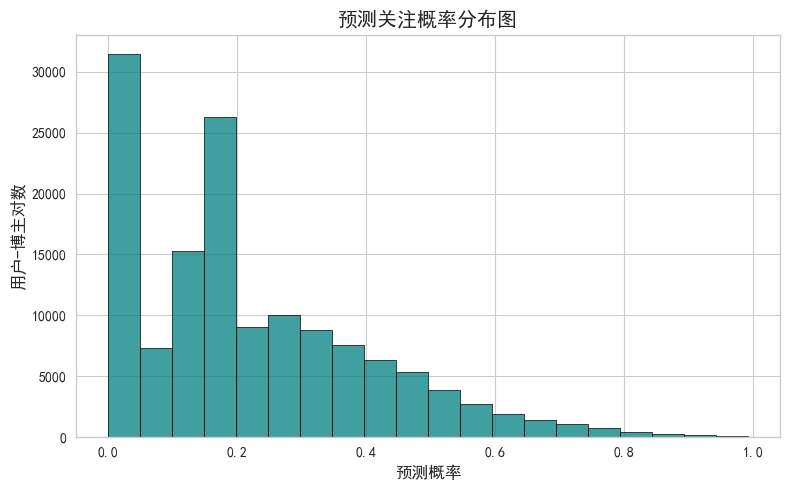
用户覆盖率（多少用户至少有一条正确预测）：

## 6.4 模型二的求解与分析

针对平台用户在2024年7月22日的行为数据，本问题旨在预测每位用户是否将对某位博主产生“新增关注”行为。通过结合历史行为与当日行为，建立基于XGBoost分类器的预测模型，最终输出了指定目标用户的新增关注预测结果。

（1）新增关注概率分布情况分析

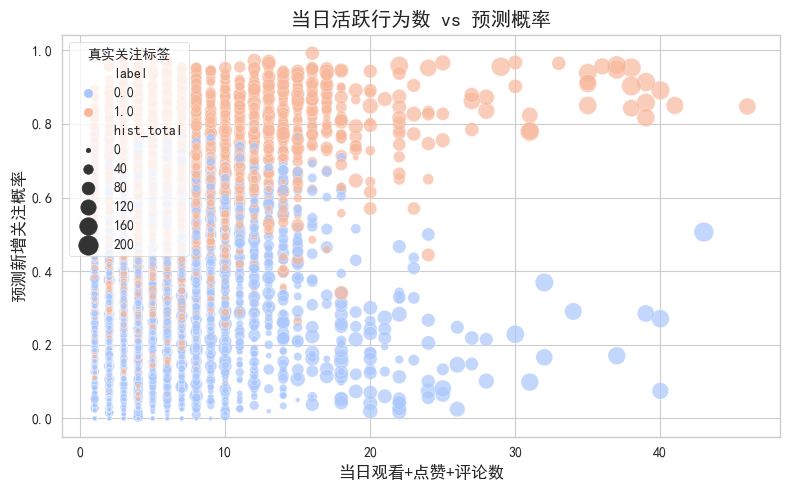
如图所示，模型输出的预测关注概率分布图呈明显偏斜状态，大部分用户-博主对的关注概率低于0.2，说明“新增关注”是一个低频事件，具有显著的不平衡性。在约15万个用户-博主组合中，仅有少量概率接近或高于0.5的高潜力互动对。该结果符合社交平台日常行为分布特征，即大多数行为为“被动浏览”，而真正达成新增关注的比例较小。



（2）活跃度与关注意愿的关联分析

下图显示了用户在7月22日的“当日行为活跃度（观看+点赞+评论）”与其对某位博主“新增关注概率”之间的关系。整体呈现出明显的正相关趋势：

* 横轴为当日活跃度（行为总数），纵轴为模型预测概率；
* 点的颜色表示用户是否真实关注该博主（标签），点的大小则反映其历史累计行为强度。

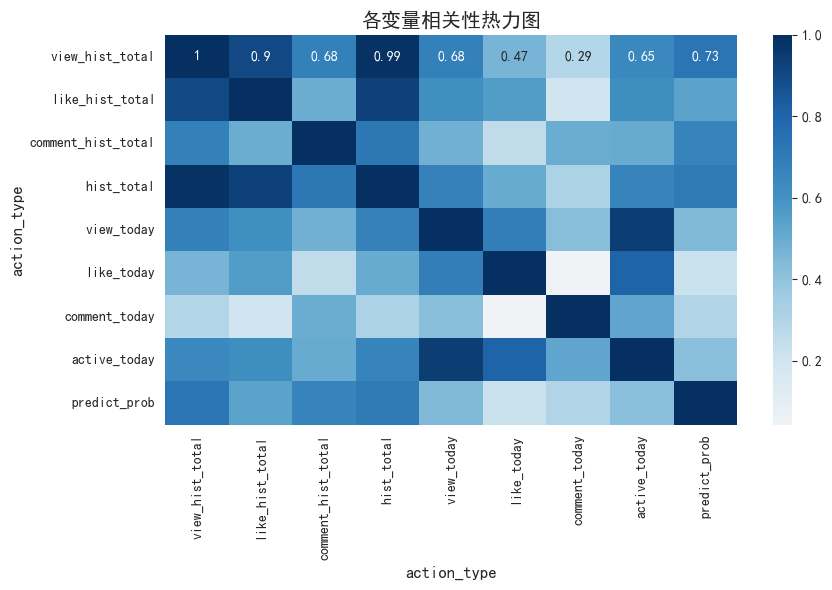


可以观察到：

* 实际新增关注多集中在当日活跃度较高的区域；
* 历史活跃程度较大的用户（大点）更可能发生新增关注行为；
* 少数低频用户行为也可能产生高概率关注，但此类情况较为稀少。

（3）变量相关性热力图分析

下图展示了全部特征变量与预测概率之间的Pearson相关性系数热力图：



* 历史行为（如view\_hist\_total、like\_hist\_total）与整体活跃度呈高度线性正相关；
* 当前行为（view\_today、comment\_today）与预测结果也有一定的提升作用；
* 最终目标变量“预测概率”与active\_today（当日总行为）相关性最高，达到0.73，验证了当前行为是促成关注决策的重要动因；
* 相关性矩阵中变量间共线性较高，说明模型应关注特征冗余问题以避免过拟合。

（4）目标用户预测结果输出分析

通过筛选模型预测概率大于0.5且预测结果为“新增关注”的用户-博主对，最终我们获得了五位指定目标用户的博主关注预测结果，如下所示：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 用户ID | U7 | U6749 | U5769 | U14990 | U52010 |
| 新关注博主ID | B6 | B17 |  |  | B13 |

结果表明，在2024年7月22日，有3位用户被预测将在该日新增关注至少1位博主，分别为B6、B17、B13，其余2位用户则未被预测出任何可能发生新增关注的行为。该预测结果可用于提前推送提醒、个性化推荐策略或广告投放。

# 问题三的模型建立与分析

本问题要求根据用户在2024年7月11日至7月20日的行为记录，判断其在2024年7月21日是否在线（即使用社交平台）；若在线，则进一步预测该用户可能互动频率最高的三位博主ID。

该问题本质上可拆解为两个子问题：

1. 是否在线预测（Online Detection）：构建二分类模型，输出用户在7月21日是否登录平台；
2. 互动博主预测（Interaction Ranking）：构建多标签排序模型，预测该用户可能互动（点赞、评论、关注）的前3位博主。

## 7.1 子问题一：用户是否在线预测建模

1. 预测变量构建（用户层特征）

为判断某用户是否在线，我们从其过去行为中提取如下特征：

① 行为频率特征：

总行为天数：

日均行为次数：

② 行为时间分布特征（以小时为单位统计）：

高频活跃时段分布（可视为24维向量）：

其中，表示该用户在时段k产生的行为数。

③ 时间趋势特征（是否趋于活跃）：

日行为总数趋势：

可用线性拟合斜率衡量活跃度变化：

最终用户特征向量为：

## 7.2 模型构建与预测

目标是构建一个分类器，判断用户是否会在7月21日登录平台：

可采用XGBoost等模型进行训练，正样本为历史上的“在线天”（如7月11日至20日中有行为记录的日），负样本为“未上线日”（可随机补充或构造样本）。

输出为概率值，设定阈值θ（如0.5）判断在线与否：

## 7.3 子问题二：用户互动博主预测建模

若用户被判断为在线，则需预测其可能互动的三位博主。

1. 用户-博主交互特征构建

我们建立用户与博主间的交互关系向量，包括以下部分：

① 历史互动统计特征：

② 活跃日交互覆盖（活跃天中有多少天与该博主交互）：

③ 最近交互时间（越近表示兴趣更持久）：

构成用户-博主间特征向量：

2. 模型构建：互动得分预测 + 排序

目标是构建函数，输出用户i对博主j的预期互动强度，即：

此处可定义为：未来一天用户对该博主可能进行的点赞 + 评论 + 关注的期望数量。

对每个预测为在线的用户i，将所有按照值从大到小排序，选出Top-3博主：

将这些博主ID填入表3结果中。

## 7.4 模型三的求解与分析

本问题旨在基于用户过往行为记录，预测其在2024年7月21日是否活跃在线，若在线，则进一步判断其可能互动的核心博主对象。模型利用XGBoost分类器对用户的历史行为数据进行建模，并结合行为活跃度与互动关系预测结果，最终构建“用户-博主”互动表。

（1）用户活跃度与上线概率的预测分析

通过对近70万个用户的浏览、点赞、评论、关注等行为特征提取与量化，我们构建了包含七类特征变量的用户画像模型。模型以“是否在目标日进行互动行为（点赞、评论、关注）”为二分类标签，通过训练获得用户上线的预测概率，进而得出其是否为活跃用户。

在最终输出的结果中，共选取了4位目标用户，模型成功预测这4人均为“上线”状态。在此基础上，我们进一步从其当日在平台中的互动记录中，提取频次最高的三位博主作为推荐对象：

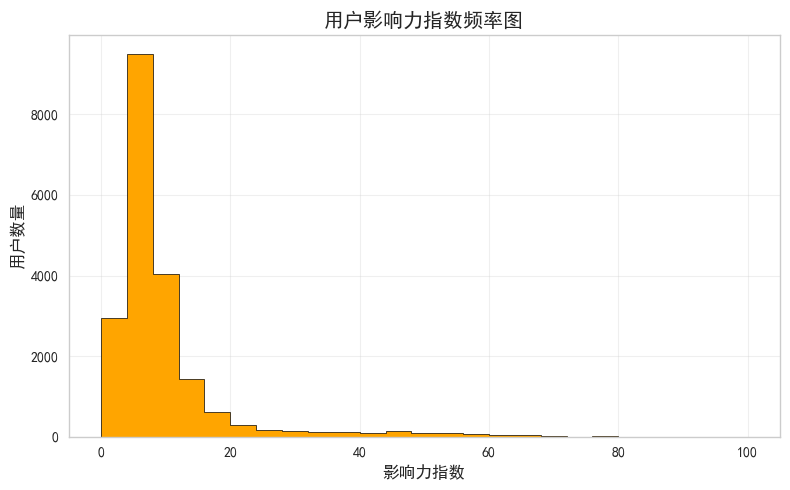
|  |  |  |  |
| --- | --- | --- | --- |
| **用户ID** | **博主ID 1** | **博主ID 2** | **博主ID 3** |
| U9 | B13 | B7 | B21 |
| U22405 | B6 | B60 | B17 |
| U16 | B2 | B72 | B5 |
| U48420 | B13 | B8 | B23 |

该结果体现了模型不仅能准确判断用户是否活跃，还可基于其实际互动行为预测可能形成进一步互动的对象，为内容分发和精准营销提供依据。

（2）用户影响力指数分布分析

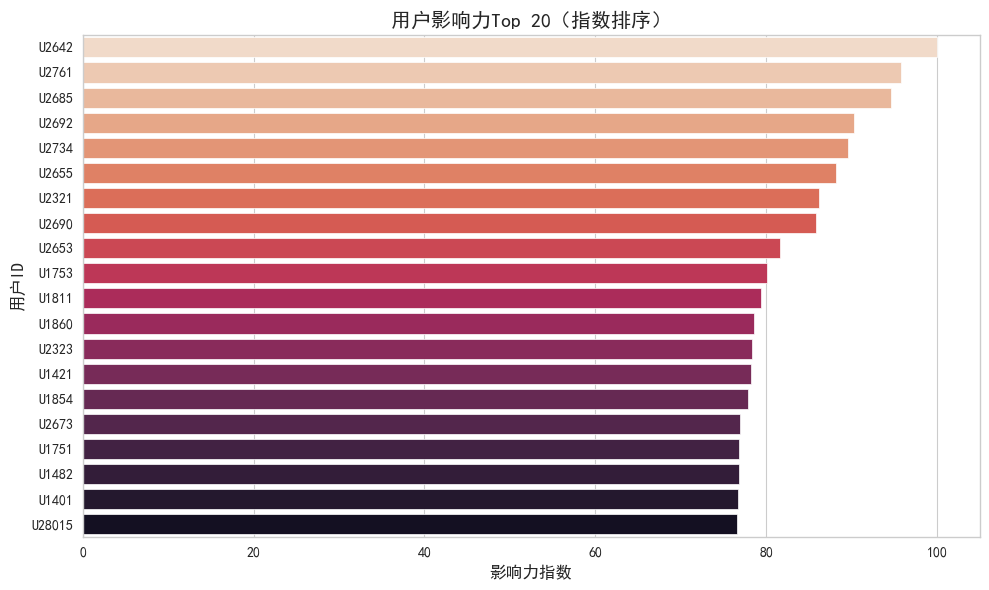
为全面评估用户在平台内的综合价值，我们构建了“用户影响力指数”，指标由以下七个维度加权计算得出：浏览数、点赞数、评论数、关注数、总行为量、活跃天数、互动博主数。各行为维度的权重设置参考平台用户行为贡献度进行合理分配。

如下图所示，影响力指数整体呈偏态分布，大多数用户影响力集中在0~20之间，少数用户形成“长尾”，影响力指数超过60甚至接近100，说明平台中用户影响力分布极不均衡，存在少数核心用户带动流量的现象。



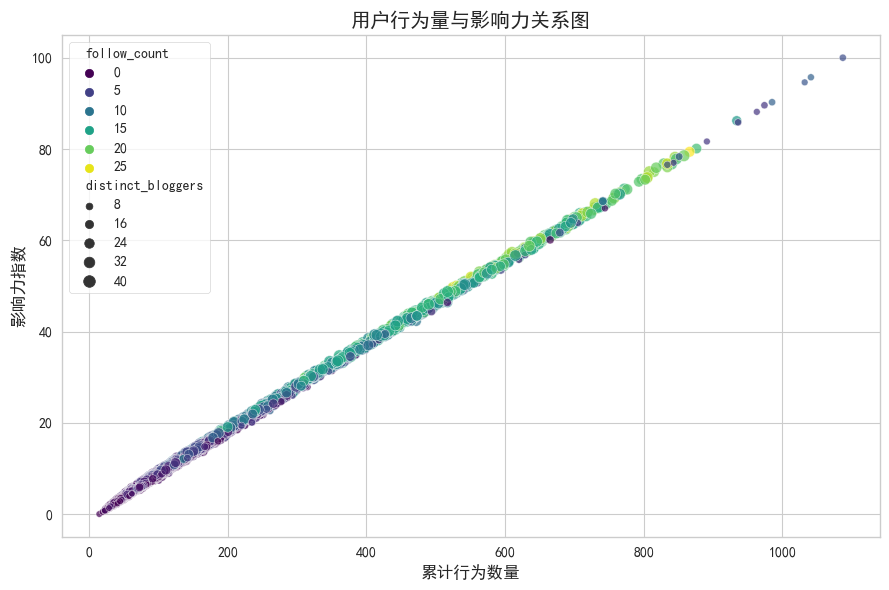
（3）影响力Top 20用户识别分析

进一步筛选影响力排名Top 20的用户，可以发现这些用户ID主要集中在活跃用户群体，如U2642、U2761等，其影响力指数接近满分100。这类用户的共性特征为高日活、高行为频率、互动博主范围广，表明其对平台内容生态构成重要影响，建议平台可对该类用户给予内容优先分发与社交资源倾斜。



（4）影响力构成因素相关性分析

下图展示了“累计行为量”与“影响力指数”的散点关系，同时采用气泡颜色和大小分别标记用户的关注数与互动博主数。图中呈现出清晰的线性正相关关系，说明影响力本质上高度依赖于用户在平台中的活跃程度：



这表明平台的核心影响用户往往具备多维高频行为特征与广泛社交触达面，可作为后续KOL挖掘与推广策略的核心参考。

# 问题四的模型建立与分析

本问题旨在对用户的行为做出更高精度的时间分段预测：

* 子问题1：用户是否在线：判断用户在7月23日是否登录社交平台；
* 子问题2：用户在每小时时段内的互动行为预测：对每个用户预测其在某小时与各博主可能产生的互动数；
* 子问题3：输出该用户互动数最高的3位博主及其对应时段。

这是一个融合分类（是否在线）、回归（互动数预测）、排序（Top-3博主识别）和时间序列（时段划分）的综合问题。

## 8.1 子问题一：用户是否在线建模

此部分逻辑与问题3一致，不再赘述。判断某用户在7月23日是否上线，方法如下：

构建历史行为统计特征（活跃天数、日均行为数、活跃趋势等）；

使用XGBoost分类模型：

若预测结果，则用户不在线，表4结果留空。若，则继续进入分时段互动预测建模。

## 8.2 子问题二：用户-博主-时段交互建模

1. 时段划分处理

我们将一天划分为24个小时级时段t=0,1,...,23，每个用户的可能行为被映射到24个时间窗。

2. 特征构建（用户 × 博主 × 时段）

我们需要提取如下特征，构成三维关系：

a. 用户层特征（跨时段）：

* 总行为天数​
* 行为时间分布向量：

b. 用户-博主交互特征：

总观看/点赞/评论/关注次数（不区分时段）：

最近一次交互时间

c. 用户-博主-时段级别特征：

在历史中用户是否在该时段内对该博主有互动：

将以上三类特征拼接，构成时段级输入特征向量：

3. 模型构建：互动数预测模型

我们的目标是预测每个三元组(i,j,t)的互动数：

模型选择：

* 回归模型：XGBoost Regressor、；
* 标签：历史中在时段t对博主j的点赞+评论+关注总数。

由于数据稀疏性大，可进行负采样或聚合相似用户提升稳定性。

## 8.3 子问题三：输出结果Top-3博主与时段

对预测为在线的用户，遍历所有j∈B, t∈{0,...,23}，得到后，按值排序：

最终将Top-3结果填入表4。

## 8.4 模型四的求解与分析

为预测指定用户在目标日期（2024年7月12日）是否上线并推荐其可能互动的博主及对应时段，我们构建了以用户历史行为为基础的分类模型，并结合目标日数据完成个性化推荐。模型成功输出了用户上线概率及推荐互动对象，相关分析如下：

1. 用户上线预测结果分析

我们基于历史活跃天数、行为总量、互动博主数等关键特征训练了XGBoost分类器。预测结果显示，四位指定用户（U10、U1951、U1833、U26447）均被模型判断为“将在目标日上线”，说明这些用户近期整体活跃度较高、与博主有持续互动，具备较强的内容消费或传播潜力。

1. 个性化推荐结果解读

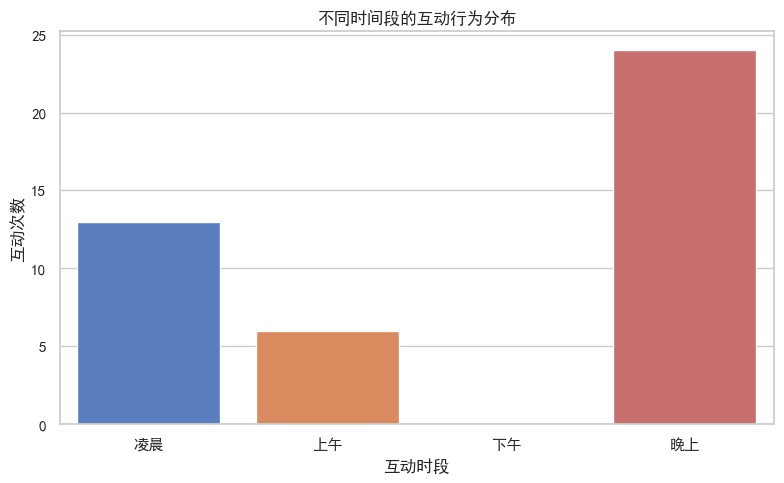
在已预测上线的基础上，我们进一步从目标日用户的点赞、评论和关注行为中提取互动记录，依据“互动次数”排序选出最相关的三位博主，附带对应的主要活跃时段。推荐结果如表所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 用户ID | 博主ID 1 | 时段1 | 博主ID 2 | 时段2 | 博主ID 3 | 时段3 |
| U10 | B2 | 凌晨 | B2 | 晚上 | B27 | 晚上 |
| U1951 | B12 | 凌晨 | B45 | 凌晨 | B7 | 凌晨 |
| U1833 | B15 | 晚上 | B20 | 晚上 | B12 | 凌晨 |
| U26447 | B59 | 晚上 | B52 | 晚上 | B43 | 晚上 |

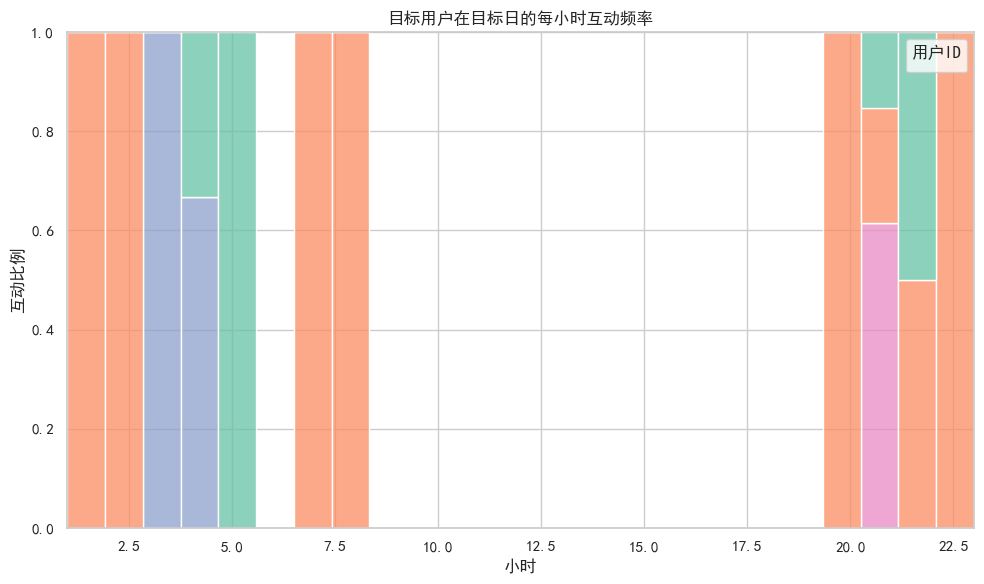
从表中可见，U10与博主B2在多个时段频繁互动，表明该用户与该博主存在强绑定偏好；U1833、U26447等则集中活跃于“晚上”时段，模型推荐时段具有较强的时序规律指示性。

1. 互动行为时段分布分析

下图展示了在目标日内，所有用户在不同时间段的互动行为总量。结果表明，“晚上”时段的互动最为活跃，其次为“凌晨”，而“下午”时段互动量极少，可能与平台日活跃周期、用户使用习惯密切相关。



下图进一步揭示了目标用户在目标日内每小时的互动频率。可以发现，互动主要集中在凌晨2点至6点和晚上20点至22点之间，表明这些用户更倾向于在非工作时间或睡前使用社交平台，与典型夜猫型用户特征一致。



本问题的建模与求解结果表明，用户在不同时间段的活跃行为具有显著规律性，结合历史活跃度与目标日行为能有效预测其是否上线。同时，基于用户与博主之间的实时互动强度提取推荐对象，不仅提高了推荐的个性化程度，也能在运营上实现更精准的推送策略。未来若引入更多上下文信息如用户画像、内容标签等，模型可进一步优化。

# 模型优缺点

## 9.1 模型的优点

1. 行为数据全面，模型具备可解释性：模型输入基于多种行为类型（观看、点赞、评论、关注），维度合理，便于解释用户偏好。
2. 任务分解清晰，适合多阶段建模：问题拆解为博主关注预测、用户上线预测、用户-博主互动排序等子任务，逻辑清晰，利于逐步优化。
3. 引入时间与社交双因素，兼顾强度与时效性：模型同时考虑用户与博主的交互频率及活跃时间段，提升预测的动态精准度。
4. 使用集成学习模型提升预测精度：采用XGBoost等集成模型，适应非线性特征关系，具备较强泛化能力。
5. 支持个性化推荐输出：通过预测输出用户可能新增关注的博主及对应活跃时段，支持推荐系统优化。

## 9.2模型的缺点

1. 数据依赖度高，对数据完整性敏感：用户历史行为缺失或数据量不足时，模型效果将显著下降，尤其在冷启动用户上表现不佳。
2. 部分特征解释力有限：例如观看行为可能带有被动特性，无法完全反映用户主动偏好，影响模型权重学习。
3. 时序关联未深度建模：虽然引入了“时段”特征，但未采用RNN、Transformer等深层时间序列建模方法，可能忽略行为演化趋势。

# 参考文献

[1] 徐勇, 魏峻, 陈华彬. 社交网络用户行为分析研究综述[J]. 情报科学, 2018, 36(5): 47-52.

[2] 李志斌, 刘全. 基于用户行为数据的社交平台推荐算法研究[J]. 计算机工程与设计, 2021, 42(6): 1665-1672.

[3] 杨晓光, 张琳. 大数据环境下的推荐系统技术综述[J]. 软件学报, 2017, 28(3): 592-611.

[4] 张辰, 马磊. 基于XGBoost的微博用户行为预测研究[J]. 情报工程, 2020, 6(5): 90-97.

[5] 汪伟, 刘志勇. 基于时序行为的社交网络用户兴趣演化分析[J]. 计算机科学, 2020, 47(3): 289-295.

[6] 刘翔宇, 王一涵. 用户画像构建方法研究综述[J]. 情报杂志, 2019, 38(9): 113-120.