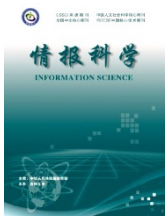

情报科学
Information Science
ISSN 1007-7634,CN 22-1264/G2

《情报科学》网络首发论文

题目：融合弹幕内容特征与行为特征的用户画像研究——以 B 站教学类视频为例
作者：杨阳，余维杰
收稿日期：2022-01-04
网络首发日期：2022-09-28
引用格式：杨阳，余维杰．融合弹幕内容特征与行为特征的用户画像研究——以 B 站教学类视频为例[J/OL]．情报科学.
https://kns.cnki.net/kcms/detail/22.1264.G2.20220926.1738.017.html

网络首发时间：2022-09-28 14:16:46
网络首发地址：https://kns.cnki.net/kcms/detail/22.1264.G2.20220926.1738.017.html

融合弹幕内容特征与行为特征的用户画像研究
——以B站教学类视频为例

杨　阳，余维杰

（中山大学 信息管理学院，广东 广州 510006）

摘　要：【目的/意义】基于弹幕的用户画像描述了用户的行为模式，有助于弹幕视频平台理解用户的需求与偏好，针对性地改善平台的内容与服务，以增强用户参与度与忠诚度。【方法/过程】以B站教学类视频为例，采集弹幕与视频数据，计算用户发送弹幕的内容特征与行为特征，根据特征对用户聚类得到用户画像，并探讨各个特征的内在关系。【结果/结论】研究结果表明：用户从行为特征角度可以分为交互性优先学习者、一般学习者、质量优先学习者与完整性优先学习者；从内容特征角度可以分为签到、表达学习感受、表达情感与讨论课程知识四种内容偏好。用户发送弹幕的频率越高，真实花费在观看视频上的时间就越少。完整性优先取向的用户更倾向于讨论课程知识，而不倾向于表达学习感受。【创新/局限】创新性地融合了用户的弹幕文本内容与时间信息，实现了更为全面的用户画像。后续研究可继续探讨不同内容主题的视频中用户画像的区别与联系。

关键词：用户画像；弹幕分析；在线学习；信息行为；主题模型

1　引　言

近年来，在线教育事业与产业蓬勃发展，在线学习已经成为了网民学习生活的重要组成部分。截止2021年6月，我国在线教育用户规模达到3.25亿人，占网民整体的32.1%[1]。在线教育以不同的形式开展，其中在线视频以其强大的表现力、直观的讲解成为大众最喜闻乐见的形式。专注于在线教学的MOOC（大规模在线开放课程平台）如学堂在线、网易云课堂以及综合性在线视频网站如YouTube、Bilibili（后简称为B站）为网络用户提供了大量的教育视频资源。

为了更好地理解用户需求，实现个性化、精准化的信息服务，用户画像在包含在线教育场景在内的众多领域中得到了应用。用户画像是对用户的统计学特征、社交关系、行为模式等特征进行描述而形成的用户模型[2]。用户基本信息、用户网络日志、用户发表的言论与评论以及用户之间的社会网络关系[3]等数据都被用作用户画像的数据源，除此之外用户观看视频所发送的弹幕评论数据也引起学者与从业者的关注。

弹幕是在线视频中以滚动形式出现的评论性字幕。区别于传统的在视频讨论区发表、回复评论的交互方式，弹幕更加彻底地打破了时空限制，使得多个用户可以方便地在不同的时间对于某个较短的视频片段表达观点与情感，以其独特的评论展现方式来营造一种“共同观影”的氛围。主流的在线视频网站皆设有弹幕评论功能，单个热门视频往往能积累成千上万条弹幕数据。数量庞大又承载了鲜明用户个人特征的弹幕数据是进行用户画像的良好数据源。

弹幕数据因其优良性质，已经有学者尝试使用弹幕数据进行用户画像，其中在线教育场景备受关注。本文总结前人的研究成果，改进了基于弹幕的用户画像模型。并以在线教育场景为例，将B站一个系列共53个分集的教学类视频作为实证分析样本，从中挖掘用户发送弹幕的内容特征与行为特征，以此为用户做画像，以期理解不同类型用户的行为模式以及各个用户特征的内在联系。在线教育场景下基于弹幕的用户画像有利于理解用户的学习特点与风格，从而为在线学习平台的建设提供参考，提高用户的使用体验与学习效果。基于弹幕的用户画像模型也可以被推广到其他场景下，通过用户画像理解用户的需求与偏好，针对性地改善平台的内容与服务，以增强用户的参与度与忠诚度。

2　相关研究

2.1 基于弹幕的用户画像模型

用户画像（User Portraits）最早由 Alan Copper[4]提出，尽管不同时期、不同学科的学者对用户画像的理解有所不

收稿日期：2022-01-04
基金项目：中山大学中央高校基本科研业务费专项资金资助（19wkpy149）。
作者简介：杨阳（1997-），男，山西晋中人，硕士研究生，主要从事信息分析与信息行为研究；余维杰（1986-），男，广东汕头人，博士，副教授，主要从事智能信息处理研究。

同[5]，但是其基本内涵一致。本文使用宋美琦等[6]从用户画像构建流程角度提出的用户画像的操作性定义，即用户画像由包含用户个人特质的原始数据，经过用户属性、用户特征、用户标签三个阶段精馏而得到。其中，用户属性指的是用户的姓名、性别与职业等基本信息构成的静态属性，以及用户的浏览频次、时长、言论与评论等行为信息组成的动态属性。用户特征是从用户属性中选择或计算出来的最能描述用户特质的属性。用户标签是将用户特征进一步提炼得到的标签化的文本，可以精炼准确地表达用户特征且易于理解与应用。根据用户画像的操作性定义，用户画像的构建流程可以分为：采集用户属性数据、提取用户特征、表示用户画像三个步骤。

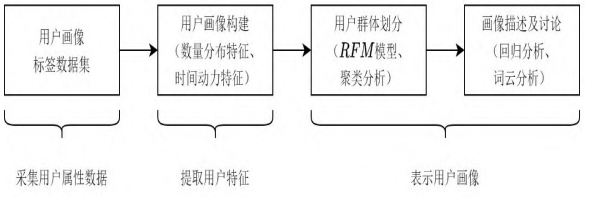

图1 基于弹幕的用户画像的早期模型

Figure 1 Early model of user portrait based on danmaku

构建用户画像的流程可以被称为用户画像模型，用户画像模型是用户画像的核心。早期基于弹幕的用户画像模型的明确描述由朱钰涵等[7]提出，该模型包含了构建通用用户画像模型的三个步骤，并根据弹幕数据的特质设计了各个步骤中的具体处理与分析方法，如图1所示。虽有珠玉在前，但是相关研究并不热衷于使用或改进基于弹幕的用户画像模型，而是更加关注于具体的用户属性的选择、用户特征的计算以及用户画像的表示。

2.2 弹幕用户属性的选择

在用户画像研究中，用户属性数据可以通过访谈、观察、调研等社会调查方法获取[8]，或从平台数据库中直接导出结构化的用户数据[9]，也可以使用网络爬虫获取用户的公开数据[10]。弹幕用户的属性数据来源于弹幕视频网站，现有研究大多编写网络爬虫获取用户属性数据。弹幕用户属性分为用户基本属性、视频属性以及弹幕属性。弹幕用户的基本属性可从弹幕视频网站的个人主页获取。朱钰涵等[7]、张璐等[11]以及严炜炜等[12]在其用户画像模型中使用了性别、官方号身份认证、粉丝数量、关注数量等用户的基本属性。视频属性来自于视频播放页面，多在计算行为特征时参与计算[13]，因此直接使用视频属性的研究较少。弹幕属性来自于视频播放窗口中的弹幕，包括内容属性与行为属性。弹幕的内容属性来源于弹幕文本内容。朱钰涵等[7]、张璐等[11]都选择了弹幕的平均长度（字数）作为内容属性。弹幕的行为属性包含弹幕的样式、发送弹幕时的视频时间与物理时间。陈忆金等、张婧婧等、杨贺晴等探讨了用户发送弹幕的数量在同一系列视频中不同分集的分布、在同一个视频分集中不同时间段的分布、在一天24小时内的分布[13-16]。陈忆金等[13]额外关注了弹幕的样式，发现弹幕的字体、大小、位置等样式与视频的题材或者场景具有相关性。

2.3 弹幕用户特征的计算

在用户画像研究中，用户特征可以通过研究者的知识与经验人工提取[17]，也可以使用决策树[18]、逻辑回归[19]、支持向量机[20]以及LDA主题模型[21]等机器学习算法计算得到。在基于弹幕的用户画像领域，现有研究多从弹幕的内容属性与行为属性中计算弹幕的内容特征与用户的行为特征。从弹幕文本中提取用户的情感特征是处理弹幕内容属性的常用方法。严炜炜等[12]、张婧婧等[14]以及李稚等[15]使用词典法或深度学习方法对弹幕文本做情感分析，发现在线教育场景下的弹幕以积极情绪为主。一些研究从用户发送弹幕的物理时间与视频时间属性中挖掘了用户的行为特征。朱钰涵等[7]将首次发送弹幕距离视频发布的时间（近度）作为其用户画像模型的用户行为分量。张婧婧等[14]定义了用户在某一系列的学习类视频下发送的第一条弹幕和最后一条弹幕的时间差为“跨度”，用于计量用户学习的持久性。严炜炜等[12]通过计算视频更新周期来量化内容创作者的影响力。

2.4 弹幕用户画像的表示

用户特征多为连续的数值数据，必要时常使用聚类方法将用户特征离散化以方便用户画像的表示。张璐等[11]、杨贺晴等[16]根据弹幕的文本内容对弹幕用户聚类。朱钰涵等[7]则利用多项逻辑回归方法针对用户特征进行群体画像，得出三类用户分别为“理性探讨型”“弹幕引领型”和“大众笼统型”。用户画像的表示多使用标签云[9]的形式，也可以借助于各种统计图表[22]表达。

综合已有的研究可以发现，基于弹幕的用户画像模型已被提出，但少有研究在此框架下进一步开展工作。学者对于单一弹幕视频网站的数据源的潜力已经充分开发，所使用的属性已经基本涵盖了B站所能提供的所有公开数据字段。对基本属性深入挖掘，计算得到的新的特征如视频更新周期、弹幕时间近度、跨度等都别具匠心，反映了用户的特征，具有较好的区分度。但是现有研究对原始属性的挖掘不够深入全面，没能探讨各个属性的内在联系。本文改进了基于弹幕的用户画像模型，计算了观看记录数量、平均弹幕发送周期、平均相对专注时间等行为特征，并使用K-Means算法对用户行为特征聚类求得用户标签；使用LDA主题模型方法对弹幕文本进行主题分析，将其作为用户的兴趣取向对用户聚类。最后根据弹幕内容特征与弹幕用户行为特征两种聚类结果创建列联表，分析用户发送弹幕的行为与弹幕内容文本之间的关系。本文以在线教育场景为例，应用基于弹幕的用户画像模型的构建流程，证明基于弹幕的用户画像模型的有效性；并根据实验结果对在线教学情境下的平台与用户提供建议，期望提高用户的使用体验与学习效果，促进在线虚拟社区的健康发展。

表1 用户属性

Table 1 User attribute

属性分类	用户属性	计算属性(用户特征)
基本属性	性别、等级、官方号身份认证、粉丝数量、关注数量、获赞数量、播放数量、阅读数量	
视频属性	视频标题、视频时长、上传日期与时间、视频清晰度、播放数量、弹幕数量、评论数量、收藏数量、投币数量、分享数量、点赞数量	视频原创性、视频更新周期
弹幕内容属性	弹幕文本内容	弹幕情感特征
弹幕行为属性	弹幕样式 弹幕物理时间 弹幕视频时间 弹幕数量	弹幕字数 弹幕中包含用户的数量 弹幕中包含视频的数量 弹幕时间近度 弹幕时间跨度

3 基于弹幕的用户画像模型

3.1 用户画像模型

本文参考通用用户画像模型改进了现有的基于弹幕的用户画像模型,如图2所示。通用用户画像模型包含了用户数据、用户属性、用户特征、用户标签四个状态以及采集用户属性数据、提取用户特征、表示用户画像三个步骤。基于弹幕的用户画像模型使用数据层、属性层(属性大类层)、特征层(原始属性层与计算属性层)、标签层与通用模型的四个状态相对应,体现了用户画像的构建过程。

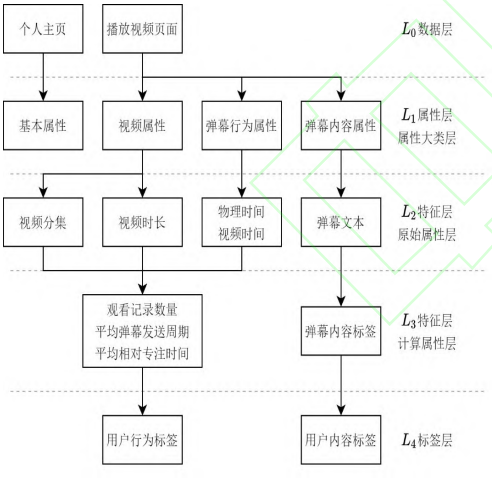

图2 基于弹幕的用户画像模型

Figure 2 User portrait model based on danmaku

数据层描述了基于弹幕的用户画像的数据来源。在不考虑多源数据融合的用户画像的情况下,用户在弹幕视频网站的个人主页,以及所关心的场景下的播放视频页面已经包含了属性层所需要的所有数据。属性大类层是对可以从数据层获得的用户属性的简单分类,其分类体系从基于弹幕的用户画像的相关研究中总结而来。原始属性层是求得计算属性层的中间层。在本文中,观看记录数量、平均弹幕发送周期、平均相对专注时间等特征需要通过视频分集、视频时长、发送弹幕的物理时间与视频时间计算得到;弹幕的内容标签通过对弹幕文本进行主题分析得到。标签层是用户画像的最终结果,本文为每个用户提供了行为标签与内容标签两种分类方法。

本文的模型具有良好的通用性,其应用可以不局限于在线教育这一场景。数据层与属性层的内容相对固定,不同场景下的基于弹幕的用户画像研究可以共享数据来源与属性分类体系而不对或少对模型进行修改。特征层与标签层中特征的选择与计算,标签的生成方式则与所研究的场景紧密相关,可以需要根据实际情况设计个性化的方案。

3.2 用户属性分类体系

用户画像由用户属性精馏而来,从原始数据源中采集的用户属性的丰富全面程度决定了用户特征的筛选与计算方法,也决定了用户画像模型的质量。本文总结基于弹幕的用户画像现有研究中使用到的用户属性,以及由用户属性计算得到的用户特征(或计算属性),如表1所示。

表1属性分类中的基本属性、视频属性、弹幕属性可以通过弹幕视频网站提供的公开数据获取,计算属性则可以基于用户属性通过特定的算法求得。对于某个基于弹幕的用户画像研究,表1中的属性并非都有价值,因此根据研究设计的侧重点与实验条件的限制选择某些属性是合理的选择。

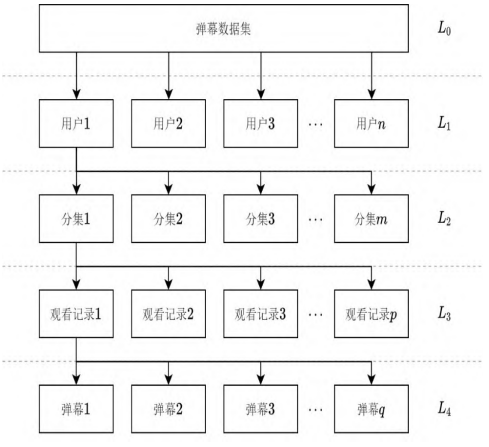

图3 弹幕数据集的逻辑结构

Figure 3 The logical structure of the danmaku data set

3.3 用户特征的算法

通用用户画像从用户属性中提取用户特征已经积累了

很多方法路径与最佳实践,基于弹幕的用户画像由于相关研究较少、弹幕相关的用户属性庞杂,用户画像刻画难度较大,如何深入挖掘用户属性中的用户特征成为了在此领域有所突破的决定性因素。本文梳理了弹幕数据集的逻辑结构如图3所示,并在此基础上计算得到观看记录数量、平均弹幕发送周期、平均相对专注时间等特征;使用主题模型方法分析弹幕文本,得到弹幕内容特征。

3.3.1 弹幕用户行为特征

采集某个系列视频下所有视频分集的弹幕,构成弹幕数据集。对弹幕数据集按照用户ID分类,可以得到每个用户所发的所有弹幕。针对单个用户,可以按照分集ID分类,得到该用户在每个视频分集下所发的弹幕列表。对单个视频分集,理想情况是这个分集下的弹幕的物理时间应该足够接近,否则说明用户可能在不同时间多次观看了这一分集。在实验中应该设置某个阈值,当某个用户的某个分集下存在两条弹幕的物理时间的差超过这个阈值,就认为这两条弹幕分属不同的观看记录。根据上述算法对分集下的弹幕分组,得到多次观看记录。在本文中,观看记录取代分集成为研究用户发送弹幕的行为的最小单位;用户、分集、观看记录、弹幕四个层次组成的弹幕数据集的逻辑结构是后续计算弹幕用户行为特征的基础。

(1)观看记录数量

基于弹幕数据集的逻辑结构,容易计算得到每个用户的观看记录数量。观看记录数量体现了用户跟随视频学习的持续性。某个用户观看记录数量多,说明这位用户相比其他用户看了更多的视频分集或者对某些视频分集的内容进行了复习。观看记录数量也是对用户观看视频分集数量的优化,有助于提高平均弹幕发送周期与平均相对专注时间算法的准确性。

(2)平均弹幕发送周期

取某用户一次观看记录所属的视频分集的时间长度,除以这个用户在这次观看记录期间发送弹幕的数量,可以求得这个用户在这一观看记录中平均多长时间发送一条弹幕。再对这个用户的所有观看记录的弹幕发送周期取平均值,可以获得平均弹幕发送周期。平均弹幕发送周期较为客观地描述了用户使用弹幕形式参与互动的意向。某个用户平均弹幕发送周期越短,那么这个用户越热衷于使用弹幕的形式参与交互。

(3)平均相对专注时间

在理想情况下,一次观看记录下任意两条弹幕的物理时间的差值与视频时间的差值应该是相等的,但是现实情况下这两个差值却较少严格相等。以图4为例,现有一个用户的一次观看记录,假设该用户在视频进度10分钟处第一次发送弹幕且在视频进度20分钟处第二次发送弹幕,那么视频时间的差值为10分钟。此时用户发送两条弹幕的物理时间的差值与视频时间的差值有三种关系。当用户发送两条弹幕的时间间隔内没有快进、暂停行为的时候,就会出现(b)这种理想情况,例如用户在物理时间8点10分与8点20分发送两条弹幕,物理时间的差值为10分钟,与视频时间的差值相等。假设用户在8点10分快进了5分钟,用户若要在视频进度第20分钟处发送弹幕,那么用户在这个时刻发送弹幕的物理时间只能是8点15分,此时物理时间的差值为5分钟,小于视频时间的差值,对应情况(a)。同理,假设用户在8点10分暂停了5分钟,那么用户在视频进度第20分钟处发送弹幕的物理时间只能是8点25分,此时物理时间的差值为15分钟,大于视频时间的差值,对应情况(c)。

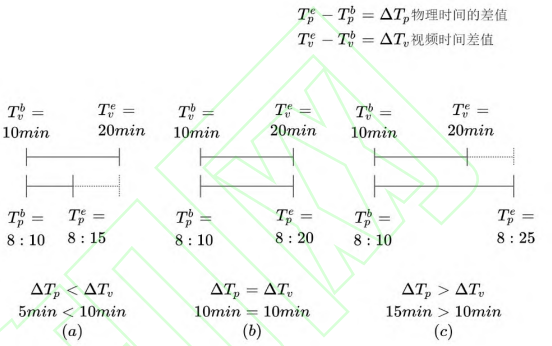

图4 平均相对专注时间原理

Figure 4 The principle of average relative focus time

根据一次观看记录下任意两条弹幕的物理时间的差值与视频时间的差值不相等的现象可以侦测用户的暂停、快进等行为。情况(a)发生时,存在两条弹幕之间的视频时间的差值大于物理时间的差值,说明用户在物理世界花费了更少的时间看完了两条弹幕之间的视频内容,即用户在这两条弹幕之间有快进或者加倍速度观看的行为。同理情况(c)发生时,存在用户在物理世界花费了更多的时间看完了两条弹幕之间的视频内容,即用户在这两条弹幕之间有暂停的行为。物理时间的差值与视频时间差值的比值反映了用户为单位视频时间投入的物理时间的多寡,这个比值可以被定义为平均相对专注时间。

平均相对专注时间可以用来描述用户观看视频的风格是走马观花还是细嚼慢咽。平均相对专注时间越大,用户实际花费更多的时间在单次观看记录上。值得注意的是,用户的平均相对专注时间属性不能作为衡量用户学习效果的指标,用户的学习效果如何仍然需要设计严谨的教育学实验来验证[23]。容易发现,平均相对专注时间的算法要求同一观看记录至少存在两条弹幕数据参与计算。

3.3.2 弹幕内容特征

用户画像研究通常使用LDA(latent Dirichlet allocation)主题模型对用户产生的文本做内容挖掘。主题模型包含了词、主题、文档三层结构,其基本思想是文档中的词有概率属于某个主题,这个主题也有概率通过某个词来表达。通过主题模型可以较好地解释文档、主题与词之间的关系,能够给出某个主题可以通过哪些词来表达以及某个文章是属于什么主题。

LDA模型采用Dirichlet分布作为概率主题模型多项分布的先验分布,其贝叶斯网络图如图5所示。其中K为文档

的主题总数，M为文档总数，N为每篇文档中包含的词汇总数，隐变量Z表示某个主题，W则是文本中的单词。α与β分别是“文档-主题”概率分布θ与“主题-词汇”概率分布φ的先验分布超参数，W是唯一可观测的变量。在LDA模型中，最重要的参数是主题的数目K，调参优化模型的过程就是设定不同的主题数目，并根据聚类效果选择最优的主题数目K的过程。研究中通常采用pyLDAvis工具将聚类结果可视化，并人工检查聚类得到的各个主题包含的词语的“主题内相似，主题间相别”的特点是否足够强，综合考虑聚类结果的优劣，取最优聚类结果对应的K即确定了主题数量。

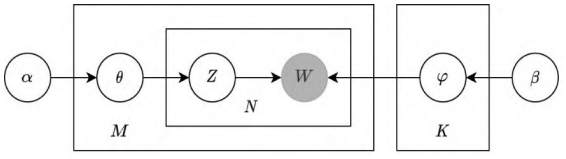

图5 LDA模型的贝叶斯网络图

Figure 5 Bayesian network diagram of the LAD model

使用LDA主题模型分析弹幕文本，能够给出指定弹幕文本所属的主题，从而获得弹幕的内容特征以及弹幕用户的兴趣取向。由于LDA方法本身适用于长文本的主题分析，而弹幕的文本长度较短，为了提高实验效果，本文将同一个用户的所有弹幕短文本合并为一个长文本，以发送弹幕文本的用户为单位进行主题分析，最终可以获得每个用户发送不同主题类型弹幕的比例。

3.4 用户标签的生成

基于用户的行为与内容特征可以进一步对用户进行聚类，从而为每个用户赋予行为与内容标签。用户的内容特征为该用户发送不同内容主题的弹幕所占的比例，本研究取比例最大值对应的内容主题作为用户的内容标签。

用户的行为特征是三个连续的数值型变量，存在多种聚类算法适用于基于用户行为特征的聚类。K-Means算法是一种经典的聚类算法，其基本思想是从包含N个对象的数据集中选择K个点作为初始中心，随后其余各点选择距离自己最近的初始中心点作为自己所属的类别，得到每个类别的所有点后计算所包含对象的各个特征的均值作为新的中心点，之后重复这一过程直到各个类别的中心点不再变化。

K-Means算法的核心参数是类别数量K。K的值既可以通过肘部法则确定，也可以将数据降维到三维甚至二维后绘制散点图，根据不同的K值对散点图的分类与上色效果确定。K-Means算法实现简单、运行快速且解释性强，是对弹幕用户行为特征进行聚类的首要选择，根据聚类结果分别给各个类别赋予标签，就得到了弹幕用户的行为标签。

4 基于弹幕的用户画像模型的实现

4.1 数据采集与特征计算

本文于2021年3月19日采集了B站学习类系列视频《小甲鱼零基础入门学习Python》[24]下的弹幕与视频数据。该系列视频共有97个分集，其中前53个分集属于对Python程序设计语言的系统教学，后44个分集则提供了具体的实战案例。本研究使用前53个分集的视频与弹幕数据，共采集得到156667条弹幕。获得的视频与弹幕的属性如表2所示，它们组成了本研究所使用的原始数据集。

表2 采集得到的视频与弹幕属性

Table 2 Captured video and danmaku attributes

数据源	属性项
视频	视频分集、视频时长
弹幕	弹幕发送者ID、弹幕物理时间、弹幕视频时间、弹幕文本内容

由于原始数据集较为庞大，本研究在数据清洗时只要一条弹幕中存在缺失值，就剔除这条弹幕，清洗后得到156066条弹幕，保留了约99.6%的数据。按照的弹幕数据集逻辑结构，将弹幕数据集按照用户分类，对于每个用户，以1个小时为阈值（按照经验确定）切分观看记录。按照计算平均相对专注时间的算法，剔除所有观看记录中都只有一条弹幕的用户，最终得到13593位用户，以及这些用户的所有分集与分集下的每次观看记录。

4.2 基于弹幕行为特征的用户聚类

计算每个用户的观看记录数量、平均弹幕发送周期、平均相对专注时间三个特征，绘制三个特征的直方图，如图6所示。用户的观看记录数量基本服从幂律分布，绝大多数用户的观看记录数量少于10次，部分用户观看记录数量可以达到50次，说明存在一些用户基本上看完了整个系列视频。用户的平均弹幕发送周期基本服从正态分布，大多数用户平均大约750秒（12.5分钟）发送一条弹幕。用户的平均相对专注时间服从右偏分布，大多数用户的平均相对专注时间为1，即实际花费的时间与视频时长相等，平均相对专注时间大于1的用户数量比小于1的更多，说明也有一些用户注重学习质量，投入更多的时间到学习视频上去。

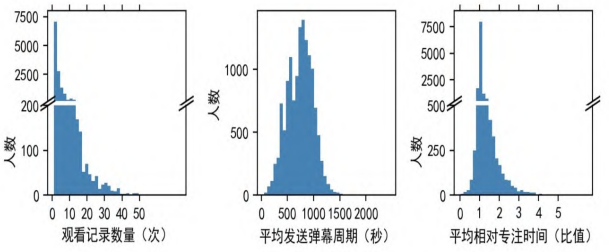

图6 行为特征直方图

Figure 6 Behavior characteristics histogram

本研究使用scikit-learn机器学习工具包中的K-Means实现对用户行为特征进行聚类。实验过程中分别设目标聚类数量参数K=2, K=3, K=4, …, K=8。实验结果表明，目标聚类数量为4时，聚类取得的效果最好，聚类效果如图7所示。因此按照用户的行为特征，可以将用户分为四种学习者。设四类用户的标签分别为a、b、c、d，分别将四类用户各自的三

个计算属性绘制箱线图，并增加所有用户数据（标签为o）的箱线图作为参照，如图8所示。

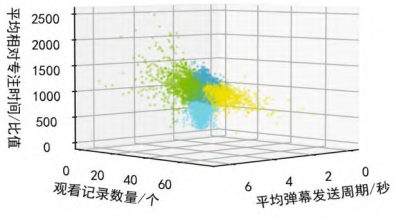

图7 用户行为特征聚类散点图

Figure 7 User behavior characteristics clustering scatter plot

从图8可见，不同类型的用户在三个计算属性上的数据特征具有明显的差异，总结这些差异可以得到表3。表3中展示了聚类的结果即四种风格的学习者。交互性优先的学习者（a）的观看记录数量少，且平均弹幕发送周期短于其他类型，说明这类用户发送弹幕的频率较高，热衷于参与以弹幕为形式的教学互动，但是其学习的持久性欠佳。一般学习者（b）的行为特征与所有用户数据（o）相比无明显差异，具有观看记录数量少，平均弹幕发送周期中等、平均相对专注时间小的特点。质量优先的学习者（c）的平均相对专注时间与平均弹幕发送周期均大于其他类型，说明这类用户发送弹幕频率较低（周期是频率的倒数），其在每个视频上花费的时间与精力更多。完整性优先学习者（d）的观看记录数量的中位数显著大于其他几种类型的学习者，说明这类用户更加关注尽可能地看完整个系列视频，构建相对完整的知识体系。部分完整性优先的学习者的平均相对专注时间小于1，可以猜测存在一些完整性优先的学习者具有较好的学习基础，更倾向于跟随视频查漏补缺，而非面面俱到地学习。

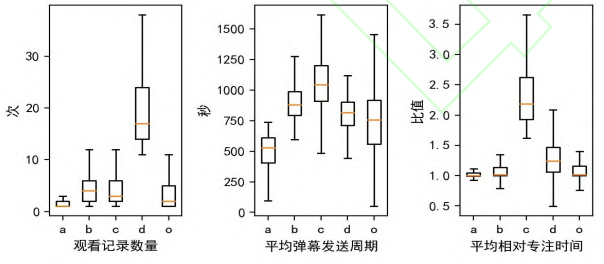

图8 不同类型用户行为特征对比箱线图

Figure 8 Box plots comparing different types of user behavior characteristics

表3 用户行为特征分类表

Table 3 User behavior characteristic classification table

分类	观看记录数量	平均弹幕发送周期	平均相对专注时间
交互性优先学习者(a)	数量少	周期短	比值小
一般学习者(b)	数量少	周期中等	比值小
质量优先学习者(c)	数量少	周期长	比值大
完整性优先学习者(d)	数量多	周期中等	比值中等

四种不同学习风格的用户的数量分布如图9所示。一般学习者占据了最大比例。交互性优先的学习者次之，说明B站这一系列视频下的学习者大多是弹幕功能的忠实用户。数量最少的是完整性优先与质量优先学习者。弹幕视频网站并非专门的在线学习平台，用户观看弹幕视频并非只为满足获取知识的需求，因此交互性优先学习者的比例大于完整性优先与质量优先的学习者也符合常理。另一种可能是学习者类型的数量分布与视频平台或者视频的类型与主题相关，这一猜想还有待进一步验证。

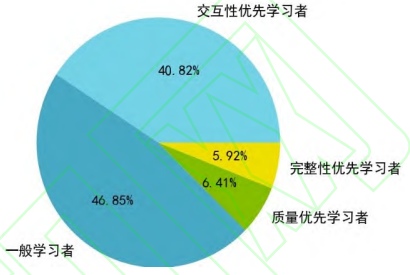

图9 用户行为特征分类饼状图

Figure 9 User behavior characteristics classification pie chart

为了探讨用户行为特征的内在关系，本研究计算了各个行为特征两两之间的皮尔逊相关系数，发现平均弹幕发送周期与平均相对专注时间两个变量低度正相关（$r=0.37$，$p=0.00<0.05$），这两个变量的散点图如图10所示。即平均弹幕发送周期越长，平均相对专注时间就越长；也就是说用户越是热衷于教学交互，花费在单位视频上的时间就越少。

现有的研究不能区分各类学习者的知识与技能基础、学习目的，因此不能断言平均相对专注时间越高越好，也不能简单否定交互性优先学习者的学习风格。较为妥善的方式是帮助用户分析其学习风格，并根据这一学习风格的用户容易出现的共性问题为其提供建议，如平均弹幕发送周期数值与平均相对专注时间数值都过低的用户应该及时确认自己的学习效果，避免投入过多时间在教学互动上而影响到知识摄取的现象。

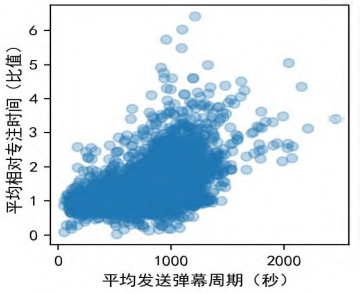

图10 平均弹幕发送周期与平均相对专注时间的关系

Figure 10 Relationship between the average danmaku sending period and the average relative focus time

4.3 基于弹幕内容特征的用户聚类

通用用户画像中，主题模型方法用于对用户产生的文本

表4 用户内容特征分类表

Table 4 User content characteristics classification table

分类	主题词举例	用户数量	弹幕举例
签到	小伙伴、10个人、加油、你们好、2020	3489	打卡签到 互谅的广大朋友大家好 2017.5.18
表达学习感受	听不懂、课后、0.5倍速、简单	2570	还好吧，简单的一匹 还是得回头复习一下 迭代好难
表达情感	哈哈哈、牛、路过、高能、嘎嘎、笑、666	4009	不知道多少遍了(□_□) 妙呀 从零开始的那个笑死我了
讨论课程知识	对象、函数、方法、类、变量、调用、属性	3371	用正则表达式 open()不用输入完整路径??? 学过数据库的想起了 insert into

内容进行主题分析，在给出文本的分类的同时给出这个类型的代表性词汇[25]。在基于弹幕的用户画像中，对用户发送的弹幕文本的多使用情感分析，主题分析则较少，本研究使用LDA对弹幕文本做主题分析。弹幕的语句较为简单，且拥有很多网络用语，因此直接使用现有的停用词表去除停用词会造成误杀现象，从而影响后续分析的结果，因此本研究在进行主题分析之前的预处理过程中使用jieba分词工具包分词，并根据弹幕用语的特点自行编制停用词表并筛除停用词。

本研究使用Gensim主题模型工具库中的LDA实现，分别设置主题数量K=2, K=3, K=4, …, K=8。实验结果表明，主题数设置为4时（超参数$\alpha = 0.25, \beta = 0.25$），聚类结果较好。如表4所示，各个主题之间有足够的区分度且主题词可以充分体现主题的语义。

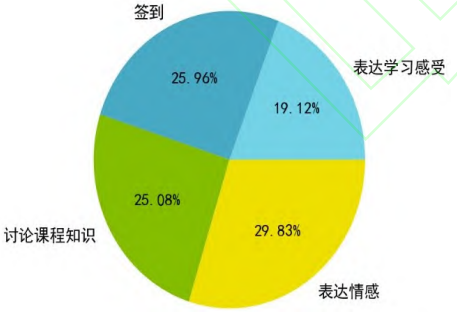

图11 用户内容特征分类饼状图

Figure 11 User content characteristics classification pie chart

LDA主题分析方法将学习者分为了四个类型。签到类型的用户主要发布当前的日期如“2020”，也会记录当前的观看人数如“10个人”，这类用户还不忘向同在观看的用户问好与激励如“你们好”“加油”。表达学习感受类型的用户通过弹幕及时记录了自己学习时的感受，如“听不懂”“好难”。区别于表达学习感受类型的用户，表达情感类型的用户对课程讲师用于调节课堂气氛的话语以及其他学习者的互动弹幕表达情感。用户表达的情绪多为正面，这与已有的研究结论一致[14]。讨论课程知识类型的用户在讨论如“对象”“函数”等面向对象程序设计语言的核心概念。

四种不同内容偏好的学习者的数量分布如图11所示。总体来看，四种内容偏好的学习者的数量占全体用户的比例差距不大。表达情感类型的学习者稍多，表达学习感受类型的学习者略少。

4.4 弹幕内容特征与行为特征的关系

融合弹幕内容特征与行为特征的用户画像有利于理解用户发送弹幕的内容与行为的内在关系，从而相对准确地理解用户需求。但是由于现有研究在从弹幕属性中挖掘内容特征与行为特征的探索较少，可以获得的计算属性较为浅显，且这些计算属性之间的关系较为简单明了，因此已有研究中很少探索用户特征之间的关系。本研究增加了多个计算属性作为用户行为特征，且对弹幕内容进行了主题分析，得到了弹幕的内容特征，为探索弹幕内容特征与行为特征之间的关系即融合弹幕内容特征与行为特征的用户画像创造了条件。

本研究将13593名用户按照弹幕内容特征与弹幕行为特征两个变量进行交叉分类，得到的列联表如表5所示（表5仅按列计算百分比）。对列联表进行卡方检验，得出用户的弹幕行为属性与内容属性中存在属性值对有显著差异（χ^2= 190.34, p=0.00<0.05）。观察表5，发现讨论课程知识类型的用户在完整性优先类型的用户中占比41.24%，远大于其他类型；表达学习感受类型的用户在完整性优先类型的用户中占比10.06%，显著低于其他类型。这说明完整性优先类型的用户更倾向于讨论课程知识，而不倾向于表达学习感受。总体来看，内容特征的四种类型在行为特征的四种用户中都有分布。这说明除了讨论课程知识，签到、表达学习感受、表达自己的情感都是用户选择发送弹幕的动机，也是支撑用户继续观看学习类视频的动力。为了适配不同类型与学习风格的学习者，区分并根据各自的特征开发功能或提供服务与建议，才符合“因材施教”的教育思想，才能满足不同类型学习者的需求，提高其使用体验。

表5 内容特征与行为特征列联表

Table 5 Contingency table of content characteristics and behavior characteristics

	交互性优先	一般学习者	质量优先	完整性优先
签到	1466,27.00%	1651,26.06%	214,24.65%	158,19.63%
表达学习感受	1144,21.07%	1203,18.99%	142,16.36%	81,10.06%
表达情感	1583,29.15%	1964,31.00%	228,26.27%	234,29.07%
讨论课程知识	1237,22.78%	1518,23.96%	284,26.27%	332,41.24%
合计	5430,100%	6336,100%	868,100%	805,100%

5 结 语

5.1 研究结论

本研究参照通用用户画像构建流程,改进了基于弹幕的用户画像模型。从已有研究中梳理了基于弹幕的用户画像的数据源与用户属性,新增了观看记录数量、平均弹幕发送周期、平均相对专注时间等计算属性作为用户的弹幕行为特征,并使用K-Means算法对弹幕行为特征聚类得到四种类型的取向,分别为交互性优先学习者、一般学习者、质量优先学习者与完整性优先学习者,其中一般学习者与交互性优先学习者在本文选取的实证案例中占比较大;使用LDA主题模型方法对弹幕文本进行主题分析,将其作为用户的兴趣取向对用户聚类,得到了签到、表达学习感受、表达情感与讨论课程知识四种取向。

通过研究弹幕行为特征的内在关系发现,平均弹幕发送周期与平均相对专注时间之间具有低度的正相关关系,即用户发送弹幕的间隔时间越长,付出的时间与精力就越多。研究弹幕行为特征与内容特征之间的关系发现完整性优先取向的用户更倾向于讨论课程知识,而不倾向于表达学习感受。

实验证明,本文改进的基于弹幕的用户画像模型在在线教育场景下可以较好地描述学习者的特征,从而区分不同的学习者类型,作为"因材施教"的基础。在其他场景下,也可以应用本文的模型,通过用户画像理解用户的需求与偏好,针对性地改善平台的内容与服务,以增强用户参与度与忠诚度。从用户画像角度看,本文的模型丰富了用户画像模型的数据源,为多源数据融合的用户画像打下基础。从弹幕分析技术角度看,本研究移植了用户画像的方法与技术,拓展了弹幕分析技术的深度与应用空间。

5.2 给弹幕视频平台的建议

基于本文的研究结论,提出以下三点建议:

(1)平台将用户的内容特征与行为特征提供给对应的视频内容创作者,创作者可以根据用户的行为模式与内容偏好调整自己的创作,从而为平台和用户提供更好的作品,提高用户观感与对平台的粘性。

(2)用户发送的弹幕中有大量的讨论课程知识的内容,这些内容体现了用户的交流过程,且对后续的读者具有重要意义。其余种类的弹幕也在营造学习的氛围感与趣味性上起着重要的作用,平台可以根据用户的学习风格或个人意愿,提供个性化的弹幕显示,比如为完整性优先的学习者提供更多讨论课程知识类型的弹幕,为交互性优先的学习者提供类型分布较为均衡的弹幕,从而提高用户的学习体验与学习效果。

(3)完整性优先的用户发送了更多的讨论课程知识的内容,为其他用户与平台做了贡献,平台可以提供精神或物质上的奖励给这些用户,鼓励用户参与视频中的讨论,从而营造良好的学习氛围,促进在线虚拟社区的长期健康发展。

参考文献

1 CNNIC.第48次中国互联网络发展状况统计报告[EB/OL].[2021-09-15].http://www.cnnic.net.cn/hlwfzyj/hlwxzbg/hlwtjbg/202109/P020210915523670981527.pdf.

2 徐芳,应洁茹.国内外用户画像研究综述[J].图书馆学研究,2020(12):7-16.

3 陈烨,王乐,陈天雨,郭勇.基于社会网络分析的社会化问答平台用户画像研究[J].情报学报,2021,40(4):414-423.

4 Alan C.交互设计之路[M].北京:电子工业出版社,2006:115-135.

5 Miaskiewicz T,Kozar K A.Personas and user-centered design:How can personas benefit product design processes[J].Design Studies,2011,32(5):417-430.

6 宋美琦,陈烨,张瑞.用户画像研究述评[J].情报科学,2019,37(4):171-177.

7 朱钰涵.在线视频社区中弹幕信息交互群体的用户画像研究[D].南京:南京大学,2019.

8 Tzavela E C, Karakitsou C,Halapi E, et al. Adolescent digital profiles: A process-based typology of highly engaged internet users[J]. Computers in Human Behavior,2017,69(4):246-255.

9 刘速.浅议数字图书馆知识发现系统中的用户画像——以天津图书馆为例[J].图书馆理论与实践,2017(6):103-106.

10 韩梅花,赵景秀.基于"用户画像"的阅读疗法模式研究——以抑郁症为例[J].大学图书馆学报,2017,35(6):105-110.

11 张璐,王若佳.在线教育视频用户评论行为比较研究——

以Bilibili网站视频评论为例[J].现代情报,2020,40(2):62-71.
12 严炜炜,王玲,程子洁.视频分享平台意见领袖特征及其形成路径研究[J].情报科学,2021,39(5):27-33.
13 陈忆金,卓林锴,赵一鸣.学习类视频弹幕用户的交互行为研究[J/OL].图书馆论坛:1-8.[2021-09-13].http://kns.cnki.net/kcms/detail/44.1306.g2.20201123.1350.002.html.
14 张婧婧,杨业宏,安欣.弹幕视频中的学习交互分析[J].中国远程教育,2017(11):22-30,79-80.
15 李稚,朱春红.双模态情感分析的弹幕网络视频平台营销策略[J].心理科学进展,2021,29(9):1561-1575.
16 杨贺晴,潘颖,王巢琛.弹幕行为视角下的高校图书馆自主学习服务优化[J/OL].图书馆论坛:1-10[2021-09-13].http://kns.cnki.net/kcms/detail/44.1306.G2.20201124.0912.006.html.
17 王庆,赵发珍.基于"用户画像"的图书馆资源推荐模式设计与分析[J].现代情报,2018,38(3):105-109,137.
18 Mostafa M M, El-Masry A A. Citizens as consumers: Profiling e-government services' users in Egypt via data mining techniques[J]. International Journal of Information Management,2013,33(4):627-641.
19 De Andres J, Pariente B, Gonzalez-Rodriguez M, et al. Towards an automatic user profiling system for online information sites Identifying demographic determining factors[J]. Online Information Review,2015,39(1): 61-80.
20 李映坤.大数据背景下用户画像的统计方法实践研究[D].北京:首都经济贸易大学,2016.
21 赵辉,化柏林,何鸿魏.科技情报用户画像标签生成与推荐[J].情报学报,2020,39(11):1214-1222.
22 吴丹,李一喆.不同情境下老年人网络健康信息检索行为与认知研究[J].图书馆论坛,2015,(2):38-43.
23 杨九民,吴长城,皮忠玲,等.促进学习还是干扰学习——弹幕对学习影响的元分析[J].电化教育研究,2019(6):84-90.
24 小甲鱼.零基础入门学习Python[EB/OL].[2021-10-24].https://www.bilibili.com/video/BV1xs411Q799?from=search&seid=10672576332774307896.
25 吴剑云,胥明珠.基于用户画像和视频兴趣标签的个性化推荐[J].情报科学,2021,39(1):128-134.

(责任编辑:孙晓明)

User Portraits Research That Integrate the Content Characteristics and Behavior Characteristics of Danmaku Users—A Case Study of Bilibili´s Teaching Video

YANG Yang, YU Wei-jie

(*School of Information Management, Sun-Yat-Sen University, Guangzhou* 510006, *China*)

Abstract: 【Purpose/significance】User portraits based on the danmaku describe the user′s behavior pattern, which helps the video platform to understand the needs and preferences of users, and to improve the content and services of the platform to enhance user participation and loyalty.【Method/process】Take the teaching video of Bilibili as an example, collect danmaku and video data, calculate the content and behavior characteristics of the danmaku sent by users, cluster users according to the features to obtain user portraits, and explore the internal relationship of each feature.【Result/conclusion】From the perspective of behavior characteristics, users can be divided into interactive first learners, general learners, quality first learners and integrity first learners; from the perspective of content characteristics, they can be divided into four types: sign-in, learning feelings expression, emotion expression and course knowledge discussion. The more frequently users send danmaku, the less time they actually spend watching the video. Integrity-first users are more inclined to discuss course knowledge, rather than express learning feelings.【Innovation/limitation】Innovatively integrates the user′s danmaku text content and time information, and realizes a more comprehensive user portrait. Follow-up research can continue to explore the differences and connections of user portraits in videos of different topics.

Keywords: user portrait; danmaku analysis; online learning; information behavior; probability topic model