计算机科学
COMPUTER SCIENCE

基于改进主题模型方法的三级短视频用户画像的研究

黄玉民, 赵婵婵

引用本文

黄玉民, 赵婵婵. 基于改进主题模型方法的三级短视频用户画像的研究[J]. 计算机科学, 2024, 51(6A): 230800093-7.

HUANG Yumin, ZHAO Chanchan. Study on Three-level Short Video User Portrait Based on Improved Topic Model Method [J]. Computer Science, 2024, 51(6A): 230800093-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230800093

基于改进主题模型方法的三级短视频用户画像的研究

黄玉民　赵婵婵
内蒙古工业大学信息工程学院　呼和浩特 010051
(hym_15927182110@163.com)

摘　要　针对如何从海量短视频数据、用户数据、交互数据中快速抽象出精准的用户兴趣的问题,提出了基于主题模型的三级标签用户画像构建方法。基于主题构建方法,将融合的 LDA 和 GSDMM 主题模型所获取的视频主题词作为用户兴趣表达向量。首先,搭建了 LDA 过滤器,通过比对阈值剔除与主题无关的文本信息,缩小文本规模,降低非主要语料对于兴趣表达向量生成的影响。然后,提出结合语义信息和语境信息的特征词权重矩阵的构建方法,使用 Bi-GRU 神经网络计算词向量的上下文特征,并将其作为语境特征,使用 TF-IDF 算法计算出的词频权重作为语义特征,结合语境和语义特征扩充特征词含义。最后使用带有兴趣权重分配的 GSDMM 模型学习特征向量权重矩阵,实现用户兴趣标签生成和用户不同喜好程度影响下的兴趣权重修正。实验结果表明,该方法能够比较完备准确地表征用户画像,优于单一的主题构建方法,并且在聚类效果上表现出色。通过构建完备的用户画像,能够精准把握用户痛点,为后续个性化推荐提供服务。

关键词:短视频;用户画像;主题分析模型;语义权重;语境权重
中图分类号　TP391

Study on Three-level Short Video User Portrait Based on Improved Topic Model Method

HUANG Yumin and ZHAO Chanchan
College of Information Engineering,Inner Mongolia University of Technology,Huhhot 010051,China

Abstract　Aiming at the problem of how to quickly extract accurate user interests from massive short video data,user data and interactive data,a three-level label user portrait construction method based on topic model is proposed. Based onthe topic construction method,the video topic words obtained by the fused LDA and GSDMM topic models are used as user interest expression vectors. Firstly,an LDA filter is built to eliminate the topic-independent text information by comparing the threshold,so as to reduce the scale of the text and reduce the influence of non-main corpus on the generation of interest expression vector. Then,the construction method of the feature word weight matrix combining semantic information and context information is proposed. The Bi-GRU neural network is used to calculate the context feature of the word vector as the context feature,and the word frequency weight calculated by the TF-IDF algorithm is used as the semantic feature. Combining context and semantic features to expand the meaning of feature words. Finally,the GSDMM model with interest weight distribution is used to learn the feature vector weight matrix,and the user interest tag generation and the interest weight correction under the influence of different user preferences are realized. Experiments show that this method can represent user portraits more completely and accurately,which is better than single topic construction method,and performs well in clustering effect. By constructing a complete user portrait,the user's pain points could be accurately grasp,so as to provide services for subsequent personalized recommendation.

Keywords　Short video,User portraits,Topic analysis model,Semantic weight,Context weight

1　引言

企业利用用户画像能够精准地把握用户的兴趣爱好,甚至能够挖掘到用户自己都没有意识到的隐藏兴趣。当今用户画像广泛应用于生活的方方面面,构建用户的虚拟形象能够很好地对已有内容进行评估。在短视频平台下,通过用户画像技术可以将用户的各种信息进行标签化,为用户添加各种tag 标签,将用户形象抽象为标签合集。根据用户画像对用户进行分类汇总,以实现精准营销。而目前短视频领域中,并没有比较明确合理的算法流程用于构建用户画像。因此本文主要采用两种主流的主题融合模型,通过对用户行为所涉及的短文本进行建模,获得用户的兴趣标签,形成完备的用户画像。其目的在于拓展用户画像的构建方法,深入理解短视频平台下用户的兴趣偏向,从而提高短视频平台的服务水平,把

基金项目:内蒙古自治区直属高校基本科研业务费项目(ZTY2023022,JY20230082);内蒙古自治区硕士研究生科研创新项目(S20231129Z);内蒙古自治区自然科学基金项目(2023LHMS06016)
This work was supported by the Basic Scientific Research Business Fee Project of Colleges and Universities Directly under the Inner Mongolia Autonomous Region(ZTY2023022,JY20230082),Inner Mongolia Autonomous Region Postgraduate Research Innovation Project(S20231129Z) and Inner Mongolia Autonomous Region Natural Science Foundation Project(2023LHMS06016).
通信作者:赵婵婵(cczhao@imut.edu.cn)

握用户动向，为用户提供更加精准的短视频推荐服务。

以用户画像为基础，本文提出的用户画像构建方法具有很大的技术价值和应用价值。技术价值方面，与传统基于统计学的用户画像的构建方式不同，本文基于神经网络，融合主流的文本主题分析算法，对于用户在短视频平台中所产出的信息进行分析学习，扩展了用户画像的构建方式，丰富了用户画像的相关研究。本文研究了用户画像的发展背景以及国内外现状，探索了目前用户画像发展存在的优缺点，拓展了 5G 时代下大数据簇拥的用户画像构建模型。应用价值方面，用户画像可以协助企业及时准确把握用户的最新动向和痛点。目前互联网企业竞争的核心因素还是在于用户，用户画像的构建能够使企业对用户的描述更加具体可感，便于企业及时了解用户兴趣和动向，把握住用户资源。用户画像被广泛地应用于推荐系统中，使推荐结果更加精准。用户画像作为推荐系统重要的一环，可以有效地改善推荐过程，使短视频平台为用户提供更加优质的推荐服务，用户在刷视频时更加方便。

2 相关研究现状评述

2.1 用户画像构建技术

目前针对用户画像的构建方法有多种，本文从不同的主体出发，采用不同的方式，根据现有的国内文献内容，将用户画像的构建方法大致分为 6 种，分别为基于本体、基于规则、基于贝叶斯网络、基于统计分析、基于聚类算法以及基于主题模型[1]的方法。Shan 等[2]从基于本体的角度出发，以酒店的评论信息为基础，以本体为核心确定酒店本体与用户属性之间的关联关系，从而构建出用户使用酒店的画像特征；Wang 等[3]基于规则的定义定量描绘用户的参与度、资历等，以问答的方式描绘出用户的属性特征；从贝叶斯网络的角度出发 Wang 等[4]提出将用户行为影印到贝叶斯网络中，实现用户特征节点的权重转换，从而提取了用户标签；Zhang[5]等基于统计学的方法对用户的个人属性和行为特征进行统计，结合数据标签体系快速归纳提取有效数据，并将用户画像应用在协同过滤推荐过程中，挖掘出隐藏的用户组关系；从基于聚类的构建方法出发，Wang 等[6]利用 K-means 聚类将用户分为多个聚类簇，分析不同簇下的用户特征；从基于主题的角度出发，Zhang 等[7]利用 LDA 模型分析微博评论数据获得了用户在不同微博主题下的兴趣特征。用户画像的构建方式具体如图 1 所示。

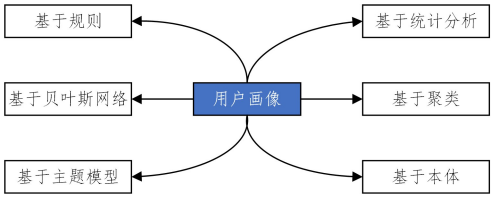

图 1 用户画像构建方法

Fig. 1 Construction method of user portrait

用户画像概念的产生最早来源于国外，最初是由交互设计之父 Cooper[8]提出，用作一种可交换的设计工具。用户画像因此又称为 User Personas 或 User Profiles[1,9]。国外的相关文献同样从不同角度构建用户画像，主要包括基于目标、基于具体场景、基于本体和基于概念[9]的方法。从基于目标的导向出发，Nielsen 等[10]在 Cooper 研究的基础上，以更加细致的粒度提出构建用户画像的步骤，假定用户是注重细节的人物，从细节谈论产品；Bylthe 等[11]则将场景情况作为输入加入用户画像的构建中，以具体场景为基础，在具体的环境搭建典型的虚构用户，从而测试用户可能会出现的本能反应；Middleton 等[12]将用户定义为基于主题的本体，使用用户在浏览器中的浏览和反馈记录去构建用户画像，但是面临着冷启动的问题；Leung 等[13]从概念角度出发进一步获得用户的行为数据，利用用户的搜索引擎日志和点击记录构建基于概念的用户画像。总之，到目前为止，用户画像的利用还在不断的探索改进中。

从国内外的相关文献中可以发现，目前针对用户画像的构建方法比较丰富且已经完备，对于数据模型下的画像构建基本上都经历了数据采集到标签挖掘的过程。基于此，本文以短视频平台所产生的用户信息为基础，利用用户画像技术构建用户兴趣模型，在技术上具有可行性。而本文以基于主题的构建方式为基础构建用户画像，不同于传统的单一主题模型构建画像，本文融合了目前常用的两种主题分析模型——隐含狄利克雷分布（Latent Dirichlet Allocation，LDA）算法和基于狄利克雷多项式混合模型的收缩型吉布斯采样算法（Gibbs Sampling for the Dirichlet Multinomial Mixture，GSDMM），并引入特征词的上下文权重（或语境权重）和词频权重（或语义权重），从而获得贴合用户特征的标签词集。除此之外，由于短视频能提供的信息较少，导致提取的用户特征比较稀疏，因此采用词共现模型扩充用户画像标签。

2.2 自然语言处理技术

自然语言处理（NLP）技术主要研究人类语言与计算机交融下的发展，模拟人脑思考方式下的人类语言计算机化处理过程。NLP 技术包含主题词提取、特征分析、情感分类等多种分支。主题分析方面，文献[14]针对短视频弹幕进行 LDA 主题建模，获取高光时刻下视频弹幕的主题词，以此将短视频进行主题分类；文献[15]针对 Web 服务描述，提出了一种改进的 GSDMM 主题分析模型用于快速有效地从服务描述信息文本中提取出表达该条 Web 服务的关键特征向量。特征分析方面，文献[16]结合卷积神经网络和词性词频特征获取文档的上下文语言信息，提出了一种基于全局和局部特征表示模型的关键词抽取方法，该模型在公开数据集测试后，关键词提取效果和准确率大幅度提升；文献[17]引入半监督学习概念，取代以往需要进行大量人工标注工作的监督学习方面，并结合无监督学习的自扩展迭代过程形成了一种无需手工标注的半监督学习关键词提取算法，实验结果表明，该方法既解决了监督学习下的繁杂的标注工作和资源环境较差的运行速度缓慢的问题，也解决了无监督学习准确率低下问题。情感分类方面，文献[18]针对目前评论短文本数据稀疏、特征模糊、并且缺乏上下文语境等问题，采用主题模型进行文本扩充，输入 Word2vec 训练获得词向量，将词向量注入双向长短期网络挖掘文本上下文特征并使用注意力机制分配权重，使用 softmax 获得情感分类，实验结果表明该模型在情感分析时的分类效果有明显提升；文献[19]在对微博评论进行二类情感分析时，以短文本的字为基本处理单位，通过 FastText 算法生成字向量和词向量，对比两者在 Bi-GRU 的训练效果预测微博评论的情感分类，实验结果表明，以字向量作为训练输入可以降低模型过拟合的风险，并且在准确率、F1 值等分

数评价上均达到 0.92 以上。

短视频的标题、简介等文本内容能够准确反映该视频的核心内容，用户往往会根据这些文本信息决定是否观看视频，是否对短视频本身感兴趣。因此本文针对标题、简介等短视频下的短文本进行处理，利用 NLP 技术对短文本进行主题建模和特征提取。将文献中处理短文本的方法进行归纳和综合，并将其迁移到短视频文本的处理过程中形成一套针对短视频的用户画像构建流程，并进行实验证明模型的有效性和准确性。

3 基于改进主题模型方法的三级短视频用户画像的构建

本文提出的用户画像构建模型根据短视频平台的特点形成三级标签，贴合短视频平台的普遍特点，形成的画像完整准确，能为后续的个性化推荐提供服务。

按照用户画像构建流程，以短视频平台的数据为原始数据来源，总体流程如图 2 所示。根据研究内容流程，对主要研究过程的具体说明如下：

从各个短视频平台下获取相当数量的短视频信息，例如视频简介、视频标题、视频热度等，作为处理的初始语料。数据获取方面，bilibili 平台作为当今最受青少年喜爱的短视频平台之一，其每日的数据流量也达到了亿兆。bilibili 短视频平台（下面简称 b 站）是最早使用视频弹幕功能的平台，并且为用户提供给视频点赞、投币、收藏、转发等操作，目前与抖音、快手占据着国内的短视频平台领域。本次数据来源主要依靠爬虫技术，从网页版 b 站获取数据信息，并根据 b 站的数据特征形成三级标签形式。一级标签（也称基础标签）为网页所能提供的用户的昵称、id 号、用户所在地、生日、注册日期、活跃等级、用户头像等信息，这些信息为用户的基本属性信息，并不需要进行处理，获得数据后经过数据清洗，再利用可视化技术展现。二级标签为用户收藏夹中的视频、最近点赞的视频、最近投币的视频以及用户自己所发布的视频的信息，包括视频的 BV 号、视频标题、视频简介、视频播放量、投币量、点赞量、评论量、弹幕信息、评论信息等，爬取后存入本地数据库，作为初始语料。三级标签为用户关注列表中的用户所发布的视频信息，由于三级标签是由关注列表反推获得，因此三级标签又被定义为隐含标签，在一定程度上表示用户的隐含兴趣，同样通过爬虫技术下载到本地作为初始数据。

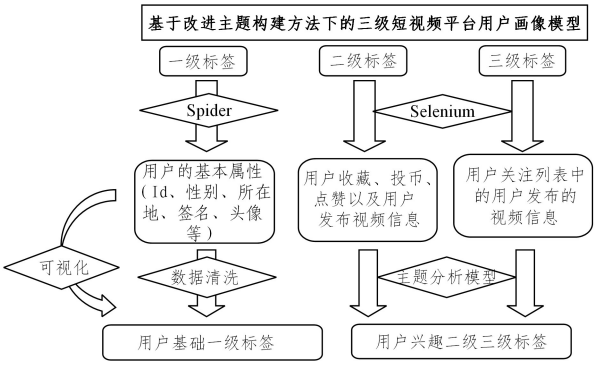

图 2 用户画像构建流程图

Fig. 2 Flow chart of user portrait construction

一级标题为短视频用户的基本信息，获取较为容易。而二级、三级标签则从视频文本中抽象而来，并且处理方式相似，因此主要介绍二级标签的获取。

二级标签处理方面，对于获取的视频语料，考虑到视频容量，并且短视频的标题能够直接反映该视频的核心内容，视频简介则是由视频发布者对于短视频内容做出的简短介绍，两者相结合能够完备地反映出短视频的内容。因此将视频的标题与简介作为分析的主要语料，对其进行相关的数据预处理，从中抽取出视频主题集合，该集合即为用户的偏好特征。

短视频的标题和简介均属于短文本，在主题模型中 GSDMM 模型比较适用于短文本的主题提取，但是在处理大批量短文本时运行速度慢，运行时间长。而主流的 LDA 模型的运行效率远高于其他主题分析模型并且输出结果明晰、可视化能力较强，但其输出主题的连贯性和一致性远低于 GSDMM 模型，并且需要人为确定主题数目。因此本课题融合两种主题模型，利用 LDA 模型过滤初始预料，人为设计阈值以剔除主题无关短文本，减少初始语料的规模，形成预备输入文档集合输入到 GSDMM 模型中计算主题标签，获取高质量用户兴趣标签输出。

输入到主题模型的词料库需要根据词频转化为可计算的数值信息，即文本数字化。但通常采用的 TF-IDF 算法计算的词频权重所生成的主题特征向量稀疏松散，质量较差，忽略了特征词本身在上下文中的重要程度。因此本课题引入循环神经网络获取特征向量的上下文语境信息，计算语境权重，并融合特征词的语义权重，形成特征向量重要程度。

一般而言，由用户发布的视频为用户最喜爱和感兴趣的内容，而 b 站用户进行投币的同时，默认为其进行点赞。因此在预处理阶段需要过滤掉最近点赞中已经包含在最近投币内容中的视频，并且因为用户的重视程度会形成不同的视频权重：用户发布的视频＞收藏视频＞投币视频＞点赞视频。为了在词频矩阵中显示出各个视频的重要比重关系，在进行主题分析建模时，根据重要程度赋予不同的偏好权重，引入兴趣权重修正因子改进主题分析模型，使不同重要程度的视频获得不同的兴趣偏重。

综上所述，二级标签构建算法流程如图 3 所示。

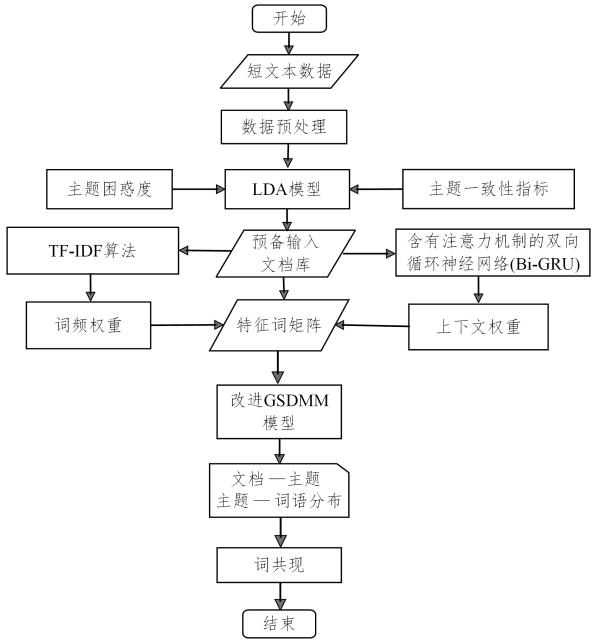

图 3 二级标签构建流程

Fig. 3 Secondary label construction process

3.1 带有阈值的 LDA 过滤器

初始语料库经过数据预处理后获得粗糙语料库，在 LDA 过滤器中进行非特征词滤筛。对粗糙语料库分词后统计词频，形成词频矩阵，使用 TF-IDF 算法计算词频权重。将权重矩阵作为 LDA 模型的输入参数。本文使用的 LDA 模型是基于传统经典的 GIBBS 采样过程估计参数。并且，使用 LDA 模型时需要自行输入主题数量，这里引入主题困惑度(Perplexity)和主题一致性指标(Coherence)两个参考量。对主题困惑度进行评分，困惑度得分越低，主题数量确定得越多；主题一致性指标的值越大确定的主题数量越多。根据经验事先人为确定阈值，对于划分好的主题下小于阈值的词汇进行记录形成滤筛词集，通过滤筛词库过滤粗糙语料库获得预备输入文档库。LDA 过滤器处理筛选文档的流程如图 4 所示。

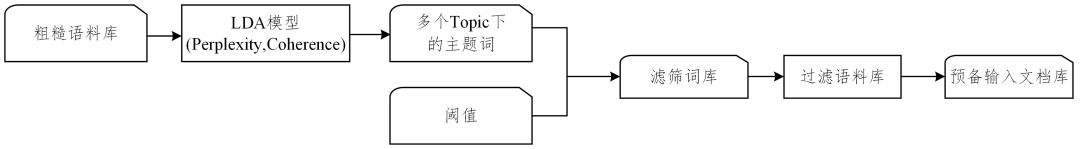

图 4 LDA 过滤器处理流程

Fig. 4 LDA filter processing flow

LDA 算法流程如表 1 所列，LDA 模型主要以两个 Dirichlet-Multinomial 共轭为基础，在概率选择不确定但是服从狄利克雷分布的前提条件下来确定文档-主题，主题-词之间的多项分布关系。LDA 模型也经历了一系列的发展，最初 Unigram model 模型假定文本中的词服从多项概率分布，并且只有单一的狄利克雷分布去推出多项分布；之后到 Mixture of unigrams model 模型，其推断某一文档的一个主题；再到与 LDA 模型最接近的概率潜在语义分析(Probabilistic Latent Semantic Analysis，PLSA)模型，它以确定的概率选取主题和词，使用期望最大(ExpectionMaximization，EM)算法估计推断结果。而 LDA 模型在 PLSA 模型的基础上加入贝叶斯框架，将 PLSA 前提的确定概率条件改为两个服从狄利克雷分布的概念参数。

主题困惑度 Perplexity 的计算式如下：

$$Perplexity(D)=\exp\left\{\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d=1}^{M}N_d}\right\} \tag{1}$$

主题一致性指标 Coherence 的计算式如下：

$$C_v=\mu(\{s_{\cos}(\vec{u},\vec{w})\mid\vec{u},\vec{w}\in W\}) \tag{2}$$

LDA 算法流程如算法 1 所示。

算法 1 LDA 算法

1. 从狄利克雷分布 α 中取样生成文档 i 的主题分布 θ_i
2. 从主题的多项式分布 θ_i 中取样生成文档 i 第 j 个词的主题 $z_{i,j}$
3. 从狄利克雷分布 β 中取样生成主题 $z_{i,j}$ 对应的词语分布 $\varphi_{z_{i,j}}$
4. 从词语的多项式分布 $\varphi_{z_{i,j}}$ 中采样最终生成词语 $w_{i,j}$

3.2 融合语境权重和语义权重的特征矩阵构成

TF-IDF 算法可以计算词在文本中的词频权重，该权重表征某个词在整个文档中出现的词频，所占词频越高说明该词在整个文档中的重要性越高，本文将这种词频权重称为语义权重。仅仅考虑词条出现的频次而忽视该词条在整个文档中的语境意义，会导致在选择表征特征向量的词集时部分无意义的词被选入而影响特征生成质量。因此本文在考虑语境权重的基础上加入上下文权重，也即语境权重，该权重用于衡量词条在整个文档语境中的影响力占比。

循环神经网络在训练时可以有效地保留词向量的上下文特征，因此本文选取 Bi-GRU 双向循环神经网络获取上下文语意信息，在单向 GRU 网络的基础上附加反向 GRU，能够有效的捕捉 t 时刻前后的历史信息，从而可以整体获取词向量的上下文特征。同时在网络中引入注意力机制加快训练速度，快速获取特征向量的语境权重。

引入加权系数 α，融合特征词 w 的语义权重 $Word\_Semanteme(w)$ 和语境权重 $Word\_Context(w)$ 形成词权重 $Word(w)$，计算公式如下：

$$Word(w)=\alpha*Word\_Semanteme(w)+(1-\alpha)*Word\_Context(w) \tag{3}$$

双向门控循环神经网络层模型如图 5 所示。

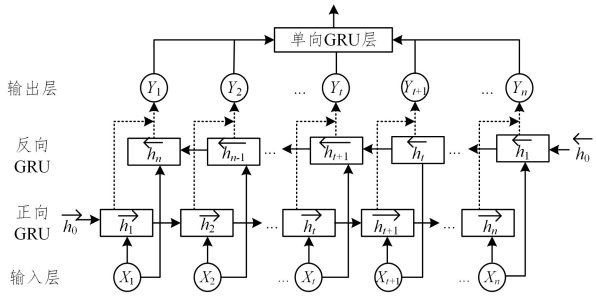

图 5 Bi-GRU 网络单元结构

Fig. 5 Bi-GRU network unit structure

3.3 引入兴趣权重修正因子的 GSDMM 模型

GSDMM 模型能够有效地提取短文本的主题，不同于 LDA 模型，GSDMM 模型假设文档只有一个主题，只需要一个最大主题数量作为上限，并以该上限开始从数据中推断主题的数量。该模型使用狄利克雷混合模型生成文档并采用吉布斯采样求取近似解，获取主题-词语的概率分布。

根据用户对于视频的不同行为划分为不同偏好等级，其中偏好大小为发布行为＞收藏行为＞投币行为＞点赞行为。因此在使用 GSDMM 模型训练过程中加入兴趣权重修正因子，增加高兴趣行为影响下的特征向量权重。

GSDMM 模型的算法流程如算法 2 所示。

算法 2 GSDMM 算法

1. 指定最大主题数量为 K，K 值作为主题分类上线
2. 对于每一篇文档 d，进行分类的概率服从多项式分布。将 d 分类到标签为 z 的族，更新该族的文档数、字数和每个字的出现次数的统计结果。在原来的基础上，文档数＋1，字数加上文档 d 的字数，该类每个字的统计结果加上 d 对应字的统计信息。
3. 分类完成后，对下面操作进行迭代：对于每篇文档 d，记录它所分类的标签 z1，在该类中剔除文档 d，更新 z1 的相关参数。重新为 d 指定一个类，此时分类的概率服从以标签 z1 剔除 d 和 d 为先验条件的条件概率分布。重新指定类的标签 z2，更新相关的参数。

3.4 词共现模型扩充特征标签

短视频标题和简介均属于短文本，通过主题分析模型获得的主题-词语分布矩阵比较稀疏，因此引入词共现模型选择性扩充特征词量。对于特征词汇表中的每个词汇，遍历每篇文档数据。对于与词语 a 共同出现在一篇文档的词语 b，统计其共同出现的次数，并将数字写入共现矩阵 ***Co-Occurrence*** 中。

$$\boldsymbol{Co\text{-}Occurrence}=\begin{pmatrix} CO_{11} & \cdots & CO_{1n} \\ \vdots & \ddots & \vdots \\ CO_{n1} & \cdots & CO_{m} \end{pmatrix} \tag{4}$$

其中，CO_{ij} 表示在共现矩阵中第 i 个词和第 j 个词的共现次数，根据共现词矩阵，选择性地依次扩充主题词内容，并将特征词矩阵中的词权重 $Word(w)$ 作为共现词的主题特征概率。

该步骤属于可选模块，需要根据分析得到的特征词容量选择是否对特征词进行扩展。若获得的用户特征兴趣向量稀疏，则需要采用共现模型平衡。

词共现模型的介绍如表 1 所列。

表 1 词共现模型

Table 1 Word co-occurrence model

概念	基本原理	特点
两个词语出现在同一文档的频率越高，其关联性越大	统计任意两个关键词在同一文档共同出现的次数，以此构建共现矩阵	反应关键词间的关联性强度

三级标签处理方面，三级标签的处理过程基本和二级标签处理过程一致，区别在于，三级标签作为用户的潜在标签不需要对所获得的视频短文本内容赋予权重。三级标签的意义在于挖掘用户可能存在的隐含兴趣。

4 实验分析

4.1 实验数据与环境

本文数据主要来源于 Bilibili 视频平台，并参考了一些快手、抖音、QQ 小世界等目前国内主流的短视频平台，采用爬虫技术将相关的视频短文本数据爬取后保存在本地。利用 Python 的函数库对数据进行预处理，包括数据清洗、格式内容修正、文本逻辑规整、中文分词、去停用词等。并根据短视频介绍文本的特点，去除诸如点赞、转发和由数字和字母组成等对分析视频本身内容无意义的词汇及一些单字词语。总计保留 5346 条短文本数据。

本文的用户画像模型主要采用 python 语言，在 Windows10 操作系统下的 PyCharm 集成开发环境运行，硬件为 CORE i7 处理器，16 GB 运行内存。

4.2 评价标准

主要针对改进下的 GSDMM 算法进行结果评价。一方面，将其与其他主题提取方法进行对比，利用主题连贯性得分判断本文提出的模型流程标签词的好坏；另一方面，从聚类的角度判断用户画像提取的好坏，一般采用轮廓系数(SC)、互信息化程度(AMI)等指标进行衡量。

4.2.1 主题指标衡量

(1)主题连贯性得分

主题连贯性评分是用于评估文本连贯性的一个指标，一般而言相似的词语应该具备相似的上下文。它用于简化文本中各部分之间的逻辑关联程度，即文本是否在主题和观点上保持一致，是否能够不止地传递信息。计算公式为：

$$CS=(C(s_1)+C(s_2)+\cdots+C(s_n))/n \tag{5}$$

其中，将待评估的文本划分为若干片段或句子，记为 s_1，s_2，…，s_n。对于每个片段，使用语言模型来计算该片段的连贯性得分，记为 $C(s_i)$。从而计算整个文本的主题连贯性得分，记为 CS，CS 值越大，主题连贯程度越好，主题建模结果越准确。

4.2.2 聚类效果分析

(1)戴维森丁堡指数(Davies-Bouldin Index，DBI)

DB 系数计算每个簇间的相似程度，求取所有相似度的平均值用于衡量聚类效果的好坏。具体的计算公式为：

$$DBI=\frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{avg(C_i)+avg(C_j)}{d_{\text{cen}}(\mu_i,\mu_j)}\right) \tag{6}$$

其中，μ 代表簇 C 的中心点，$avg(C)$ 为簇 C 内样本的平均距离，$d_{\text{cen}}(\mu_i,\mu_j)$ 对应两个簇中心点间的距离。簇与簇间的距离越远，DBI 的值越小，此时聚类结果越好。

(2)邓恩指数(Dunn Index，DI)

Dunn 指数通过比较豚鼠结果中不同类别之间的最短距离与同一类别内部数据点之间的最大距离来评估恐慌的紧密度和分离度。具体计算公式为：

$$DI=\min_{1\leqslant i\leqslant k}\left\{\min_{j\neq i}\left(\frac{d_{\min}(C_i,C_j)}{max_{1\leqslant l\leqslant k}\ diam(C_l)}\right)\right\} \tag{7}$$

其中，$d_{\min}(C_i,C_j)$ 为簇 C_i 和簇 C_j 最近样本间的距离，$diam(C_l)$ 为簇 C_l 内样本间的最远距离。DI 系数越大代表类间距越大，类内间距越小，从而证明聚类效果越好。

(3)轮廓系数(Silhouette Coefficient，SC)

SC 系数用于衡量每个样本点到其簇内样本的距离与其最近簇结构之间距离的比值。计算公式如下：

$$S=\frac{b-a}{\max(a,b)} \tag{8}$$

$$SC=\frac{1}{n}\sum_{i=1}^{n}s_i \tag{9}$$

其中，S 为单独样本的轮廓系数，a 表示某个样本与其所在簇内其他样本点的平均距离，b 表示某个样本与其他簇样本的平均距离。对所有样本轮廓系数求取平均值作为这个聚类结果的 SC 指标，因此 SC 指标的取值位于[−1，1]，越靠近 1 聚类效果越好。

4.3 实验流程与结果

本文在获取到的有效短视频文本数据的基础上，对文本数据进行清洗分词处理后进行实验。实验内容包括基础实验和对比实验。本文提出了一种基于改进主题下的用户画像构建方法，并将其命名为 LB-GSDMM，因此在基础实验中需要设定参数并将用户兴趣向量可视化展示，以便非专业人员及时把握用户动向。对比实验则是进行主题好坏对比和主题聚类效果对比，LB-GSDMM 将短视频文本进行主题分类和主题词提取，将重要程度较高的主题词用于表征用户兴趣，因此需要在主题提取的性能上和主题分类的优劣上进行对比。

4.3.1 基础实验

(1)过滤器实验

LDA 模型和 GSDMM 模型均适用于文本主题提取，但是 LDA 模型主题提取速度较快、提取效率低，GSDMM 模型更适用于短文本主题词提取，速度较慢。因此采用 LDA 模型对原始预料预处理，粗略感知初始数据特征，并将非主题预料进行剔除，缩小数据规模并避免非主要文本影响后续用户兴趣特征的生成。如图 6、图 7 所示，根据主题困惑度和主题一致性指标，确定当主题数为 28 时 LDA 过滤器生成结果最好。将主题数确定为 28 时，LDA 过滤器具体结果如图 8 所示。

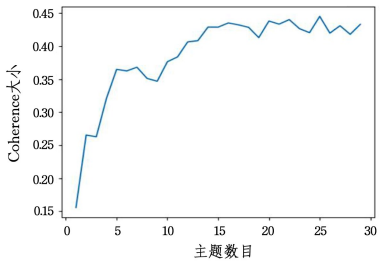

图 6 主题-主题一致性指标变化情况

Fig. 6 Changes of topic-topic consistency index

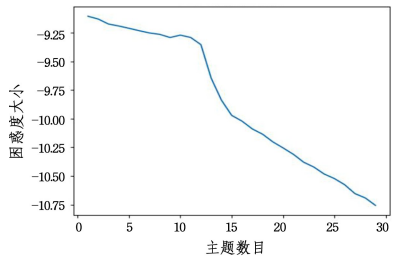

图 7 主题-主题困惑度指标变化情况

Fig. 7 Changes of theme-theme confusion index

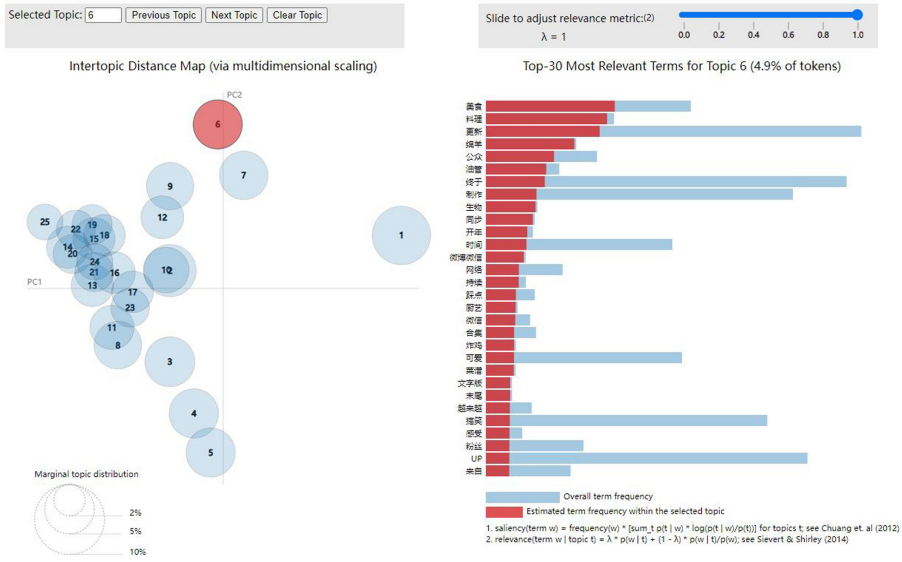

图 8 LDA 过滤器结果

Fig. 8 LDA filter results

根据 LDA 模型生成的结果，将不包含主题词的文本剔除，形成预备生成文档库。

(2)词权重提取实验

词权重主要根据由 TF-IDF 算法形成的语义权重和由神经网络 Bi-GRU 形成的语境权重形成。TF-IDF 算法主要根据词本身在文本中出现的频次实现，而 Bi-GRU 则需要将预备文档库中的词转化为词向量，本文采用 Word2Vec 算法，将每个词转化为含有 100 维的词向量。将每个词向量输入到 Bi-GRU 网络中进行训练学习，获得包含上下文特征的多维矩阵，Bi-GRU 实验相关参数设定如表 2 所列。

表 2 Bi-GRU 的主要参数

Table 2 Main parameters of Bi-GRU

算法	参数	解释	取值
Word2Vec 算法	vector_size	特征向量的维度	100
	window	在每个单词周围的最大距离	5
	min_count	单词的最小出现次数	1
	workers	并行训练的线程数	4
	sg	使用的算法类型(sg=1 表示使用 Skip-gram 算法进行训练，而 sg=0 表示使用 CBOW 算法进行训练)	1
Bi-GRU 神经网络	input_dim	输入数据的特征数	1
	hidden_dim	GRU 层的隐藏层单元数	100
	output_dim	输出数据的特征数	100
	num_layers	GRU 层的堆叠层数	1

针对神经网络学习获得的高纬矩阵，在网络外部嵌入仿池化步骤，利用 PCA 算法将每个词向量学习的矩阵降低成一维列向量。文献[15]提出的语境权重计算方法指出，某词在该语境中与其他词的粘连度决定了该词的语境权重，而粘连程度则是通过向量间的余弦相似度获取。综上所述，某词的语境权重计算公式如下：

$$Word_{Context(w_i)}=\frac{\sum_{j=0}^{k} Similarity(PCA_i, PCA_j)}{k-1}, i \neq j \quad (10)$$

其中，k 为某词所在的短文本中分词的数量，PCA_i 为第 i 个词经过 PCA 算法降维后的结果，$Similarity(PCA_i, PCA_j)$ 代表第 i 个词和第 j 个词的余弦相似度，词 i 的语境权重由该词与其他词的余弦相似度加权平均获取。

根据词权重计算公式，取 α 为 0.5，表示语境权重和语义权重的重要程度各占据词权重的一半。

(3)用户画像可视化实验

根据 LB-GSDMM 模型生成的用户特征向量结果形成用户画像的可视化展示，根据主题概率和词权重形成画像词云，取重要程度排名前 20 的主题词汇作为用户的兴趣标签，具体结果如图 9 所示。

图 9 用户画像可视化结果

Fig. 9 Visualization results of user portraits

4.3.2 对比实验

(1)主题对比实验

实验采用控制变量的方式，将处理完备的词预料注入到不同的主题模型进行关键词提取，对比结果如表 3 所列，从实验结果来看，LB-GSDMM 模型的主题提取性能高于其他模型。

表 3 主题对比结果

Table 3 Theme comparison results

模型	主题连贯性得分
LB-GSDMM	0.62
LB-LDA	0.44
LB-TextRank	0.33
LB-TF-IDF	0.40

(2)聚类对比实验

为了证明本文提出的 LB-GSDMM 进行主题分类时能够更好地解决短视频下用户画像构建问题，本文引入多个聚类算法进行对比，将经过前 2 个步骤处理的预料结果输入 *K*-means，DBSCAN，BIRCH 等聚类算法中进行比较，结果如表 4 所列。综合考量 DBI，DI，SC 等聚类评价指标，可以得出本文使用的 LB-GSDMM 模型对于短视频文本特征提取的性能表现优异，能够较为完备地提取用户兴趣标签。

表 4 聚类对比结果

Table 4 Cluster comparison results

模型	DBI	DI	SC
LB-GSDMM	0.55	0.83	0.91
LB-Kmeans	0.66	0.67	0.88
LB-DBSCAN	0.35	3.57×10^{-8}	0.79
LB-BIRCH	1.99	3.46×10^{-8}	0.86

结束语 面对目前热门的短视频领域中对于用户画像的形成缺乏比较完整合理的流程和方法等问题，本文提出了一种基于改进主题模型方法的三级短视频用户画像构建模型，融合两种主题分析方法，并引入了语义和语境权重来丰富词向量内涵。在用户画像构建方面，本文提出了一种短视频平台下的三级标签模型，并融合 LDA 和 GSDMM 模型对视频信息进行精确聚类和主题提取。根据实际的用户注意力权重，引入兴趣权重偏移机制，加大用户喜好占比，使得结果更加精准。此外，使用 TF-IDF 算法计算词频权重并结合由双向循环神经网络训练出的上下文权重，获取高质量的特征替代庞大的初始语料库。实现基本流程的同时进行了两次对比实验对本文提出的模型进行评估，多次实验结果表明本文方法能够很好地刻画短视频平台下的用户画像。

参 考 文 献

[1] ZHAO Y H，LIU F L，LUO L. A Review of User Portrait Research in the Context of Big Data：Knowledge System and Research Prospects [J]. Library Science Research，2019(24)：13-24.

[2] SHAN X H，ZHANG X Y，LIU X Y. Research on User Portraits Based on Online Reviews-A Case Study of Ctrip Hotel [J]. Intelligence Theory and Practice，2018，41(4)：99-104，149.

[3] WANG L X，SHEN Z，LI Y. Social Q & A community user portrait construction [J]. Information theory and practice，2018，41(1)：129-134.

[4] WANG Q F. Research on Bayesian network in user interest model construction [J]. Wireless Internet Technology，2016(12)：101-102.

[5] ZHANG Y. Practical analysis of statistical methods for user portraits in the context of big data [J]. Modern Business，2020(6)：9-10.

[6] WAN J P. Design and implementation of real - time game user portrait system based on big data [D]. Beijing：China University of Geosciences，2021.

[7] ZHANG H X，SHENG F F，XU P Y，et al. Visualization of population characteristics based on mobile terminal log data [J]. Journal of Software，2016，27(5)：1174-1187.

[8] COOPER A. The inmates are running the asylum [M]. Vieweg+Teubner Verlag，1999.

[9] GAO G S. A review of user portrait construction methods [J]. Data Analysis and Knowledge Discovery，2019，3(3)：25-35.

[10] NIELSEN L. Personas-user focused design [M]. London：Springer，2013.

[11] BLYTHE M A，WRIGHT P C. Pastiche scenarios：Fiction as a resource for user centred design [J]. Interacting with Computers，2006，18(5)：1139-1164.

[12] MIDDLETON S E，SHADBOLT N R，DE ROURE D C. Ontological user profiling in recommender systems [J]. ACM Transactions on Information Systems(TOIS)，2004，22(1)：54-88.

[13] LEUNG K W T，LEE D L. Deriving concept-based user profiles from search engine logs [J]. IEEE Transactions on Knowledge and Data Engineering，2010，22(7)：969-982.

[14] FENG Y，ZOU B X，XU H Y. Short video recommendation model based on video content features and barrage text [J]. Journal of Liaoning University(Natural Science Edition)，2021，48(2)：108-115.

[15] HU Q，SHEN J J，JING G H，et al. Service clustering method based on describing context feature words and improved GSDMM model [J]. Communication Journal，2021，42(8)：176-187.

[16] ZU X，XIE F. A keyword extraction algorithm based on global and local feature representation [J]. Journal of Yunnan University(Natural Science Edition)，2023，45(4)：825-836.

[17] CAI M D，SHEN G H，HUANG Z Q. A semi-supervised learning keyword extraction method without manual labeling [J]. Journal of Chinese Computer Systems，2024，45(1)：69-74.

[18] CHEN L Y，WU T. A short text sentiment analysis method combining topic model and self-attention mechanism [J]. Foreign Electronic Measurement Technology，2021，40(11)：18-23.

[19] FAN H，LI P F. Research on short text sentiment analysis based on FastText word vector and bidirectional GRU recurrent neural network-Taking Weibo comment text as an example [J]. Information Science，2021，39(4)：15-22.

HUANG Yumin，born in 1998，postgraduate. His main research interests include data mining and personalized recommendations .

ZHAO Chanchan，born in 1982，Ph.D，associate professor. Her main research interests include computer network and software defined network.