基于弹幕文本挖掘的视频 UP 主研究

朱文静

（湖南大学工商管理学院，湖南、长沙，410000）

摘要：近年来随着视频门户网站的发展，视频自媒体也得到了较大发展。视频因为具有更高的传播示能和内容密度，能够更大程度调动消费者的积极性，为品牌营销带来更多机会。本文以哔哩哔哩弹幕视频网为研究平台，使用python语言对美妆类up主的视频弹幕进行爬取，使用LDA方法对弹幕文本进行主题分析，最后进行词云可视化，从而分析不同美妆up主的特色，为视频创作者提供思路，也为品牌方选择与品牌风格相同的up主提供参考。

关键词：UP 主；视频弹幕；爬虫；LDA 主题分析；词云；

中图分类号：870.3040　　**文献标识码：**B

0 引言

互联网的发展，把我们的三维世界拉成一个平面，当世界是扁平的，这平面上的每一个点上突出来便是“顶峰”，而让我们成为这个“顶峰”最有效地途径就是自媒体。自媒体一词起源谢因·波曼与克里斯·威理斯于2003年7月联合提出的研究报告[0]，之后我国学者张斌对“自媒体”的概念进行了细化，他认为新的媒介工具的产生对自媒体的产生最为关键[0]。近几年随着直播、短视频的出现，视频自媒体成为创业新热点，美妆类自媒体对消费者购买力的影响越来越大，不仅更多的人想要进入美妆自媒体行业，还吸引了越来越多的品牌方和美妆自媒体合作。以优酷、爱奇艺、哔哩哔哩弹幕网[1]为代表的视频门户网站开始向移动端迈进，为了增加视频的互动性，增加了“弹幕[2]”功能，以便用户随时随地发表感受，而“弹幕”作为一种新兴的视频评论形式，因其趣味性和社交性，逐渐被学界和业界关注。目前国内外对于弹幕的研究仍处在起步阶段，关于弹幕的研究主要集中在弹幕的亚文化领域[0]、传播领域[0]，近几年学术界开始将弹幕与文本分析相结合[0]。

本文以哔哩哔哩弹幕网（下文简称B站）为研究平台，对美妆领域的视频自媒体进行研究，通过爬取UP主[3]们视频的弹幕数据，对弹幕文本进行聚类和主题分析，从而分析美妆博主视频流行的原因及视频特色，以期为其他博主的内容创作提供思路，以及为品牌商投放广告提供参考。

1 研究理论

1.1 DBSCAN 聚类算法

DBSCAN(Density-Based Spatial Clustering of Applications with Noise)是基于一组邻域来描述样本集的紧密程度的，参数(ϵϵ, MinPts)用来描述邻域的样本分布紧密程度。

[1]哔哩哔哩弹幕网于2009年6月26日创建，现为国内领先的年轻人文化社区，被粉丝们亲切的称为“B站”。目前拥有动画、番剧、国创、音乐、舞蹈、游戏、科技、生活、娱乐、鬼畜、时尚等分区，并开设直播、游戏中心、周边等业务板块。

[2]弹幕指在视频播放过程中，经用户发送并实时显示在视频播放画面的评论性文本，源自日本弹幕视频分享网站Niconico动画，国内首先引进为AcFun以及后来的bilibili。

[3] UP主（uploader）指在视频网站、论坛、ftp站点上传视频音频文件的人。起先由日本传入，之后被国内ACGN视频网所使用。

其中，$\epsilon\epsilon$描述了某一样本的邻域距离阈值，MinPts描述了某一样本的距离为$\epsilon\epsilon$的邻域中样本个数的阈值。DBSCAN算法原理如下：

（1）DBSCAN通过检查数据集中每点的Eps邻域来搜索簇，如果点p的Eps邻域包含的点多于MinPts个，则创建一个以p为核心对象的簇；

（2）DBSCAN迭代地聚集从这些核心对象直接密度可达的对象，这个过程可能涉及一些密度可达簇的合并；

（3）当没有新的点添加到任何簇时，该过程结束。

DBSCAN算法的特别之处在于不需要输入类别数k，其优势是可以发现任意形状的聚类簇，同时它在聚类的同时还可以找出异常点，聚类结果没有偏倚。一般来说，如果数据集是稠密的，并且数据集不是凸的，用DBSCAN算法的效果较好。

1.2 LDA 主题分析

LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，包含词、主题和文档三层结构。LDA采用词袋的方法，将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。LDA以文档集合D和主题数k作为输入，D中每篇文档d看作个单词序列:

$$\langle w_1, w_2, \cdots, w_i, \cdots, w_n \rangle，\ w_i \text{表示第 i 个单词}$$

D 中涉及的所有不同单词组成一个词汇表大集合 V，假设 V 中共有 m 个词，LDA 以文档集合 D 作为输入，希望训练出的两个结果向量 ：

（1）对每个 D 中的文档 d，对应到不同主题的概率 θ_d:

$$\langle pt_1, \cdots, pt_i, \cdots, pt_k \rangle$$

其中 pt_i 表示 d 对应 k 个主题中第 i 个主题的概率：

pt_i=d 中有多少个词是第 i 个主题也有的/d 中所有词的总数

（2）对每个 T 中的主题 t，生成不同单词的概率向量 φ_t：

$$\langle pw_1, \cdots, pw_i, \cdots, pw_m \rangle$$

其中 pw_i 表示主题 t 生成 V 中第 i 个单词的概率：

pw_i=主题 t 对应到 V 中第 i 个单词出现的次数/主题 t 下的所有单词总数

LDA 的核心公式：

$$P(w|d)=P(w|t)* P(t|d)$$

直观的看这个公式，就是以主题作为中间层，可以通过当前的 θ_d和 Φ_t给出了文档d中出现单词w的概率。其中p(t|d)利用 θ_d计算得到，p(w|t)利用 Φ_t计算得到。算法开始时先随机地给 θ_d和 Φt赋值（对所有的d和t）。然后上述过程不断重复，最终收敛得到LDA的结果。由于LDA属于无监督算法，每个主题并不会要求指定条件，但聚类后通过统计出各个主题上词的概率分布，那些在该主题上概率高的词，能非常好的描述该主题的意义。

2 研究方法

2.1 方法概述

2.1.1 数据爬取与清洗

本文使用了python语言爬取视频弹幕，开发平台是Spyder。编写爬虫代码前首先进行网页结构的分析，确定要抓取的数据属于网页结构的哪一部分。首先解析视频播放页面链接，解析后发现b站的视频链接地址是https://www.bilibili.com/video/+视频av号。然后查找弹幕资源所在地，发现弹幕内容包括在一个xml文件里面。最后通过视频av号找到弹幕对应的xml文件号。爬取B站弹幕之前需要登录B站账号,否则不能查看历史弹幕记录，登陆成功后

将cookie中的信息复制爬虫程序中的cookie文件中，开始弹幕数据的爬取。之后对爬取的弹幕数据进行清洗、去重，得到初步的实验数据。具体流程如下:

（1）登录B站账号，得到cookie信息；

（2）解析B站视频连接地址；

（3）通过视频av号找到弹幕对应的xml文件号；

（4）数据爬取、去重、初步清洗。

2.1.2 分词和去停用词

数据清洗完后，对数据进行分词处理。本文使用的是jieba分词对数据进行分词处理，jieba分词允许开发者使用自定义词典，以便包含jieba词库里没有的词，提高分词准确率。根据弹幕语言特点及美妆领域有其专有的名词，本文对jieba词库进行了扩充，具体操作是从网上开放的词库选择和弹幕和美妆领域相关的词语，之后对这些词语进行人工筛选，最后汇总成实验所需要的词典，表1展示了自定义词典的部分词。分词处理后对实验数据进行去停用词处理，本文整理了“中文停用词库”、“哈工大停用词表”、“四川大学机器智能实验室停用词库”和“百度停用词表”，形成本文的停用词表。

表1 自定义词典添加词（部分）

Tab. 1 Custom dictionary add words

真实、前方高能、颜表立、弹幕护体、合影、开口脆、真香、战歌起、空降成功、多谢款待、鬼畜、666、233、BGM、+1、金坷垃、火钳刘明、awsl、xswl、口红、遮瑕、粉底、睫毛膏、肤质、眉笔、散粉、定妆、精华、洗面奶、痘痘、过敏、卸妆、面霜、喷雾、眼影、腮红、面膜、洗发露、发膜

2.1.3 DBSCAN 与 LDA 相结合的主题分析

本文的主要研究目的是对不同UP主的弹幕数据进行主题分类，从而得到不同UP主的弹幕主题。传统的LDA主题分析需要人工确定主题数，这就意味着输入不同的主题数会得到不同的实验结果。为了避免人工确定主题数，本文在进行LDA主题分析之前，先用DBSCAN算法得出聚类的数目，作为LDA分类数的输入。本文使用sklearn来实现DBSCAN算法，用tf-idf方法进行特征表示，再使用TruncatedSVD进行降维，得到DBSCAN聚类算法的输入数据。为检验tf-idf特征表示后的聚类效果，本文选择了部分实验数据进行测试，调用Calinski-Harabaz Index[4]来检验聚类效果，实验中测得Calinski_harabaz_score为773720467.3643986，聚类效果良好。在此基础上又用不同维数进行实验，维数参数无论是5000还是15000，都得到了相同的聚类数目，进一步证明了算法的可靠性。DBSCAN得到的聚类数目作为LDA主题分析的num_topics参数，进行最后的主题分析。

2.1.4 词云可视化

词云就是对文本中出现频率较高的关键词予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而使读者快速掌握信息点。本文使用python的第三方库wordcloud进行词云可视化，调用generate_from_frequencies方法，为每个词赋予权重。为了使得词云效果更加清晰，对每个主题中的关键词指定了颜色，最后生成每个主题的词云展示图。

[4] Calinski-Harabaz Index通过计算类中各点与类中心的距离平方和来度量类内的紧密度，通过计算各类中心点与数据集中心点距离平方和来度量数据集的分离度，最后由分离度与紧密度的比值得到。CH越大聚类效果更优，实验中测得Calinski_harabaz_score为773720467.3643986。

3 研究过程

3.1 实验对象

本文选择B站为研究平台。作为目前国内最大的二次元综合娱乐社区平台，B站拥有着过亿用户群体，这些用户群体带来的巨大的流量，吸引了众多内容创作者进驻B站。不仅如此，B站为了吸引更多创作者，还制定了“bilibili创作激励政策[5]”，这进一步刺激了更多的人进入B站进行创作。本文选取研究对象的依据是“看哔哩哔哩”网站（kanbilibili.com），网站会对B站的视频播放情况每日更新“bilibili总榜”和“UP榜”，帮助用户了解网站排名。

本文主要研究美妆类UP主，所以选择“UP榜”的“时尚”版块为依据，根据排名UP主排名高低来选择研究对象。综合“粉丝数”和“播放数”排名并排除非美妆类UP主[6]后，分别选取了“机智的党妹”、“宝剑嫂”、“千户长生”以及“Vivekatt”四位UP主作为研究对象。

表2 美妆类UP主信息

Tab. 2 Beauty Uploaders information

综合排名	UP主	视频数	粉丝数	播放数
1	机智的党妹	138	361.01万	14366.87万
2	宝剑嫂	153	287.24万	12967.99万
3	千户长生	219	160.49万	10364.75万
4	Vivekatt	145	135.10万	4938.72万

数据更新时间截止到2019年9月15日

3.2 弹幕主题分类及词云可视化

经过DBSCAN和LDA相结合的主题分析，初步的主题分类结果中仍存在一些意义不明、重复或是对实验结果没有意义的数据，所以需要进一步筛选、归纳，形成最终的主题分析表。为了使实验结果更加直观，对结果进行了词云可视化。

3.2.1 机智的党妹

实验数据显示“机智的党妹”的弹幕被分成了三个主题：

表3 “机智的党妹”弹幕主题

Tab. 3 Barrage theme of “Clever Dangmei"

topic1	0.275*"哈哈哈哈哈哈"+0.045*"可爱"+0.028*"笑"+0.026*"喜欢"+0.010*"表白"+0.008*"666666"+0.007*"猫"+0.005*"螺蛳粉"+0.003*"恍恍惚惚"+0.002*"红红火火"+0.002*"傻子"+0.001*"睡裤"+0.001*"表情包"+0.001*"日常"+0.001*"看猫"+0.001*"布偶"+0.001*"撸猫"
topic2	0.074*"党妹"+0.041*"好看"+0.009*"党哥"+0.005*"哥哥"+0.005*"直男"+0.003*"男朋友"+0.003*"好美"+0.003*"漂亮"+0.003*"男生"+0.003*"帅"+0.002*"男人"+0.002*"颜值"+0.002*"帅哥"+0.002*"少女"+0.001*"太帅"+0.001*"女人"+0.001*"女朋友"+0.001*"女装大佬"+0.001*"大哥"+0.001*"姑娘"+0.001*"温柔"+0.001*"好听"+0.001*"撩"
topic3	0.044*"啊啊啊"+0.042*"!"+0.008*"卧槽"+0.004*"开心"+0.004*"心动"+0.002*"仙女"+0.002*"种草"+0.002*"弯了"+0.002*"截图"+0.002*"打call"+0.002*"疯狂"+0.001*"舔"+0.001*"加一"+0.001*"恋爱了"+0.001*"激动"+0.001*"忍不住"+0.001*"既视感"+0.001*"技术"+0.001*"兴奋"+0.001*"前方高能"+0.001*"牛逼""

topic1的“哈哈哈”、“笑死”等词说明视频中有引发观众笑的内容，“猫”、“螺蛳粉”、“睡裤”等词显示视频中出现较多日常生活场景，接近观众生活，而这些可能是引发观众发笑的原因；topic2交替出现“党妹”和“党哥”，结合UP主的视频封面分析，UP主经常模仿男性妆容，风格在少女和少年之间变化；topic3的“啊啊啊”、“卧槽”等词说明UP

[5] “bilibili创作激励计划”是指B站推出的针对up主创作的自制稿件进行综合评估并提供相应收益的系列计划。up主们只要拥有 1000 粉丝或 10 万累计播放量(以创作中心数据为准)就可以加入激励计划。

[6]本文根据“时尚区”排名选择了排名前五的UP主，因为“周六野Zoey”是健身类UP主，所以排除。

主视频中有让观众意想不到的画面，结合topic2分析，可能是UP妆前妆后形象巨大反差极大调动了粉丝情绪。

对上述主题分析结果进行词云可视化：

图 1 “机智的党妹”词云

Fig. 1 Word cloud of “clever Dangmei”

3.2.2 宝剑嫂

实验数据显示“宝剑嫂”的弹幕被分成了三个主题：

表4 “宝剑嫂”弹幕主题

Tab. 4 Barrage theme of “Baojiansao"

topic1	0.153*"哈哈哈哈哈哈"+0.079*"互动"+0.046*"好看"+0.045*"喜欢"+0.043*"激情"+0.033*"爱"+0.030*"嫂子"+0.005*"棒"+0.004*"心动"+0.004*"谢谢"+0.004*"口音"+0.004*"三连"+0.004*"啦啦啦"+0.002*"普通话"+0.002*"红红火火"+0.002*"发际线"+0.002*"恍恍惚惚"+0.001*"巨蟹座"+0.001*"鹿晗"+0.001*"女装大佬"+0.001*"好玩"+0.001*"甜"+0.001*"激动"+0.001*" 电影"
topic2	0.049*"可爱"+0.048*"想要"+0.044*"!"+0.030*"啊啊啊"+0.025*"我我我"+0.015*"拉低"+0.015*"中奖率"+0.007*"哇哇"+0.006*"视频"+0.006*"中奖"+0.005*"我来了"+0.004*"开心"+0.002*"太棒了"+0.002*"优秀"+0.002*"承包"+0.002*"超棒"+0.002*"努力"+0.001*"投币"+0.001*"幸运"+0.001*"抽奖"+0.001*"抽到"+0.001*"抽中"+0.001*"机会"+0.001*"支持"+0.001*"大方"+0.001*"好评"
topic3	0.013*"英语"+0.007*"朋友"+0.005*"大学"+0.005*"宿舍"+0.004*"社恐"+0.004*"毕业"+0.004*"青协"+0.003*"舍友"+0.003*"女孩"+0.002*"闺蜜"+0.002*"数学"+0.002*"社团"+0.002*"三观"+0.002*"同学"+0.002*"基金"+0.002*"初中"+0.002*"独处"+0.002*"习惯"+0.002*"学习"+0.002*"专业"+0.002*"学校"+0.002*"姐妹"+0.002*"讨好"+0.002*"大二"+0.001*"存钱"+0.001*"迪士尼"+0.001*"大三"+0.001*"四级"+0.001*"考生"

topic1的“哈哈哈”、“好看”等词说明视频内容有趣，“互动”、“口音”、“普通话”、“女装大佬”等词可能是引发观众笑的原因；topic2的“中奖率”、“抽中”等词说明视频中可能有抽奖活动，这触发了观众发弹幕的行为；topic3的“大学”、“宿舍”、“毕业”等词都和学生群体有关，说明宝剑嫂的视频中会经常出现与学生群体相关的话题，而其粉丝群体可能以学生为主，从而引发了观众的共鸣。

对上述主题分析结果进行词云展示：

图 2 “宝剑嫂”词云

Fig. 2　Word cloud of “Baojiansao”

3.2.3 千户长生

实验数据显示“千户长生”的弹幕被分成了三个主题：

表5　“千户长生”弹幕主题

Tab. 5 Barrage theme of “Qianhuchangsheng"

topic1	0.388*"哈哈哈哈哈哈"+0.056*"可爱"+0.022*"喜欢"+0.017*"啊啊啊"+0.007*"恍恍惚惚"+0.006*"红红火火"+0.005*"好帅"+0.004*"温柔"+0.004*"爱你"+0.004*"姐妹"+0.004*"厉害了"+0.002*"什么鬼"+0.002*"好美"+0.002*"吐槽"+0.002*"贵妇"+0.001*"做作"+0.001*"骂人"+0.001*"蜡像"+0.002*"猪叫"+0.001*"非洲"
topic2	0.095*"本尼"+0.029*"霹雳"+0.020*"!"+0.011*"骚鸡"+0.007*"人老珠黄"+0.007*"清纯"+0.004*"魔鬼"+0.004*"美猴王"+0.002*"芭比"+0.002*"戏精"+0.002*"吓死"+0.002*"妖艳"+0.001*"gay"+0.001*"猴哥"+0.001*"女人"+0.001*"妖艳贱货"+0.001*"牛逼"+0.001*"面具"+0.001*"男人"
topic3	0.171*"好看"+0.010*"好像"+0.006*"面膜"+0.004*"眼妆"+0.004*"喷雾"+0.003*"定妆"+0.003*"头发"+0.003*"好用"+0.003*"素颜"+0.002*"衣服"+0.002*"嫩"+0.002*"粉底"+0.002*"防晒"+0.002*"口红"+0.002*"内双"+0.002*"高级"+0.002*"遮瑕"+0.001*"卸妆"+0.001*"味道"+0.001*"牌子"+0.001*"眼线"+0.001*"散粉"+0.001*"补水"+0.001*"贵"+0.001*"单眼皮"+0.001*"亮片"+0.001*" 油皮”

topic1的“哈哈哈哈哈哈”、“喜欢”等词说明视频内容有趣，“吐槽”、“贵妇”、“做作”等词可能是引发观众笑的原因；topic2的“本尼”、“benny”等词是粉丝对千户长生的昵称，“戏精”、“妖艳”等词说明UP主化妆风格可能比较浮夸艳丽，而“gay”一词进一步说明了UP主的身份特征；topic3的“面膜”、“眼妆”、“喷雾”等词说明UP主使用的产品被粉丝密切关注。

对上述主题分析结果进行词云展示：

图 3　“千户长生”词云

Fig. 3　Word cloud of “Qianhuchangsheng”

3.2.4 Vivekatt

实验数据显示“Vivekatt”的弹幕被分成了三个主题：

表6　“Vivekatt”弹幕主题

Tab. 6 Barrage theme of “Vivekatt"

topic1	0.107*"哈哈哈哈哈哈"+0.069*"喜欢"+0.032*"!"+0.015*"吃"+0.009*"男朋友"+0.008*"美"+0.008*"开心""+0.006*"笑"+0.005*"小姐姐"+0.005*"啊啊啊啊"+0.004*"温柔"+0.004*"素颜+0.003*"好听"+0.003*"辛苦"++0.003*"自然0.002*"舒服"+0.002*"干净"+0.001*"心动"+0.001*"口音"+0.001*"少女"+0.001*"好甜"
topic2	0.056*"好看"+0.041*" 想看"+0.008*"女神"+0.007*"教程"+0.007*"种草"+0.005*"表白"+0.004*"韩妆"+0.004*"厉害"+0.004*"羡慕"+0.004*"美瞳"+0.004*"眉毛"+0.004*"肤质"+0.003*"眼影"+0.003*"推荐"+0.003*"卷发"+0.003*"遮瑕"+0.003*"面膜"+0.002*"学到"+0.002*"睫毛膏"+0.002*"指甲"+0.002*"好酷"+0.002*"分享"+0.001*"安利"
topic3	0.047*"可爱"+0.037*"好好看"+0.017*"啊啊啊"+0.009*"嘴唇"+0.008*"声音"+0.008*"护肤"+0.007*" 脖子"+0.004*"眼睛"+0.003*"痘"+0.003*"痣"+0.003*"油皮"+0.002*"腿"+0.007*"单眼皮"+0.002*"干皮"+0.002*"暗沉"+0.001*"敏感"+0.001*"混油"+0.001*"闭口"+0.001*"肤色"+0.001*"胸"+0.001*" 发型”

topic1的“哈哈哈哈哈哈”、“喜欢”等词说明视频内容有趣，“吃”、“男朋友”、“口音”等词可能是引发观众笑的原因；topic2的“想看”、“教程”、“种草”等词说明

UP主的化妆技术被观众所认可，使用的产品也被观众所关注。topic3的“嘴唇”、“脖子”、“眼睛”等词说明UP主的长相特征被观众所关注。

对上述主题分析结果进行词云展示：

图 4 “Vivekatt”词云

Fig. 4 Word cloud of “Vivekatt”

4 结论与展望

经过上述四位UP主的弹幕主题分析，从视频创作者角度来说，想要成为一个受欢迎的up主，有以下三点建议：一是视频内容一定要有趣，这样观众才会有继续看下去的欲望。综合分析四位UP主的主题分类结果，发现topic1都是“哈哈哈”，譬如“机智的党妹”弹幕中的“螺蛳粉”、“睡裤”、“猫”等日常生活中常见的场景；“宝剑嫂”弹幕中的“口音”、“普通话”、“巨蟹座”等；“千户长生”弹幕中的“吐槽”、“贵妇”、“做作”等；以及“Vivakatt”弹幕中出现的“吃”、“男朋友”等。这些词都与日常生活贴近，更能拉近与观众的距离。二是up主本身要有鲜明的个人特征，譬如“机智的党妹”的软妹外表和模仿男性妆容所造成的反差；“宝剑嫂”经常讨论学生群体有关话题，乐于分享个人经验的“知心大姐”形象；“千户长生”的“gay”、“妖艳贱货”标签；以及“Vivekatt”的具有个人特色的长相特征。三是视频内容要能调动观众的情绪，触发观众发弹幕的动机。譬如“机智的党妹”的妆前妆后的强烈对比总是能让粉丝兴奋惊叹；“宝剑嫂”经常选择学生群体相关的话题进行讨论，引发粉丝共鸣，而且会经常进行抽奖活动，鼓励粉丝发弹幕；“千户长生”夸张的化妆风格；“Vivekatt”专业的化妆技巧，引发粉丝发弹幕留言求教程。

从品牌方来角度分析，主题分类能帮助品牌方了解UP主的个人风格和粉丝群，从而更精准的投放广告。譬如“机智的党妹”就是青春洋溢、古灵精怪少女/少年，镜头前不拘小节，嘴贫戏精，需要展示个人技能的时候又认真负责；“宝剑嫂”的形象则是“知心大姐”，乐于和粉丝分享关于学业爱情以及成长路上的心得；“千户长生”的风格较为浮夸，因为“gay”属性，所以展现出“内心强大敢于做自己”的形象；“Vivekatt”的形象是温柔的女神，其典型的单眼皮长相和娴熟的化妆技巧被粉丝认可，所以使用的产品也被大家关注。

关于研究的局限性，因为本文是以B站位研究平台，而B站用户群体以年轻群体为主，所以实验放在不同平台可能存在差异性。关于之后的研究，可以选择不同平台进行研究，还可以增加研究对象的数量，增加研究的可靠性。

参考文献

【1】Bowman Chris S . We Media: How Audience are Shaping the Future of News and Information[J]. 2003.

【2】张彬.对“自媒体”的概念界定及思考[J].今传媒,2008(08):76-77.

【3】陈一, 曹圣琪, 王彤. 透视弹幕网站与弹幕族:一个青年亚文化的视角[J]. 青年探索, 2013(6):19-24.

【4】陈席元. 弹幕话语建构的青年亚文化网络社群研究——以哔哩哔哩网对Keyki事件反应为例[J]. 电脑知识与技术, 2014(20):4667-4669.

【5】王佳琪. 基于弹幕视频网站的弹幕文化研究[D]. 山东师范大学.

【6】周舟. 传播学视野下的网络青年亚文化——“弹幕文化”解读[D]. 2015.

【7】高雪. 抵抗与收编：弹幕亚文化与主流文化的关系研究[D]. 暨南大学.

【8】Hao X , Xu S , Zhang X . Barrage participation and feedback in travel reality shows: The effects of media on destination image among Generation Y[J]. Journal of Destination Marketing & Management, 2019, 12:27-36.

【9】Chen Y , Gao Q , Rau P L P . Watching a Movie Alone yet Together: Understanding Reasons for Watching Danmaku Videos[J]. International Journal of Human–Computer Interaction, 2017:1-13.
Tang Y , Gong Y , Xu L , et al. Is Danmaku an Effective Way for Promoting Event based Social Network?[C]// Companion of the 2017 ACM Conference. ACM, 2017.

【10】江含雪. 传播学视域中的弹幕视频研究[D]. 华中师范大学, 2014.

【11】谢梅, 何炬, 冯宇乐. 大众传播游戏理论视角下的弹幕视频研究[J]. 新闻界, 2014(2):37-40.

【12】马志浩[1], 葛进平[2]. 日本动画的弹幕评论分析：一种准社会交往的视角[J]. 国际新闻界, 2014(8):116-130.

【13】史蓉蓉, 张宁. “四元律”理论下的弹幕视频分析[J]. 传媒, 2015(7):75-77.

【14】Wu Z , Ito E . [IEEE 2014 IIAI 3rd International Conference on Advanced Applied Informatics (IIAIAAI) - Kokura Kita-ku, Japan (2014.8.31-2014.9.4)] 2014 IIAI 3rd International Conference on Advanced Applied Informatics - Correlation Analysis between User\"s Emotional Comments and Popularity Measures[C]// Iiai International Conference on Advanced Applied Informatics. IEEE, 2014:280-283.

【15】Jianwei Niu, Shijie Li,”Shasha Mo, Sen Yang, Boyu Fan.Affective Content Analysis of Online Video Clips with Live Comments in Chinese”,2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovations, guangzhou,China,2018,pp.849-856.

【16】郑飏飏, 徐健, 肖卓. 情感分析及可视化方法在网络视频弹幕数据分析中的应用[J]. 现代图书情报技术, 2015(11).

【17】邓扬,张晨曦,李江峰.基于弹幕情感分析的视频片段推荐模型[J].计算机应用,2017,37(4):1065-1070.

【18】洪庆,王思尧,赵钦佩,李江峰,饶卫雄.基于弹幕情感分析和聚类算法的视频用户群体分类[J].计算机 工程与科学,2018,40(6):1125-1139.

【19】何明. 面向在线视频弹幕数据的挖掘方法研究[D]. 2018.

【20】庄须强, 刘方爱. 基于AT-LSTM的弹幕评论情感分析[J]. 数字技术与应用, 2018, v.36；No.332(02):220-222.

【21】王晓艳. 基于图像分析的网络视频弹幕的情感分类研究与应用[D]. 2018.

【22】邱宁佳, 丛琳, 周思丞, et al. 结合改进主动学习的SVD-CNN弹幕文本分类算法研究[J]. 计算机应用, 2018.

【23】朱海澎. 弹幕文本挖掘:一种影视内容定量测评方法[J]. 传媒观察, 2019, 422(02):87-93.

Research on video uploaders based on barrage text mining

Zhu Wenjing

（School of Business Administration, Hunan University, Changsha 410000，China）

Abstract： In recent years, with the development of video portals, video self-media has also been greatly developed. Because of its higher communication performance and content density, video can arouse consumers' enthusiasm to a greater extent and bring more opportunities for brand marketing. This article uses the Bilibili barrage video website as a research platform, uses python language to crawl the video barrage of the main makeup uploaders, uses the LDA method to analyze the theme of the barrage text. Finally, the word cloud is visualized to analyze the characteristics of different beauty uploaders, provide ideas for video creators, and provide references for brands to choose uploaders with the same style as the brand.

Keywords: Uploaders;Video Barrage;Web Crawler;LDA Topic Analysis;Word Cloud;