计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(1):69−74 [doi: 10.15888/j.cnki.csa.006728]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于聚类的网络直播群体行为建模分析①

兰荣亨[1], 朱　格[2], 杨　文[2], 田　野[2], 朱　明[1]

[1](中国科学技术大学 信息科学技术学院, 合肥 230027)
[2](中国科学技术大学 计算机科学与技术学院, 合肥 230027)
通讯作者: 田　野, E-mail: yetian@ustc.edu.cn

摘　要: 近年来, 随着互联网技术的不断发展, 以及手机、平板电脑等移动终端的普及, 网络直播逐渐兴起并壮大. 国内众多直播平台基本都有送礼机制, 允许观众购买平台提供的虚拟礼物来打赏主播. 观众的打赏对于主播和平台来说都是主要的收入来源之一, 所以理解观众的行为以挖掘观众的用户价值, 提升用户的变现能力就显得尤为重要. 本文以斗鱼直播平台为例, 聚焦于直播平台上的高消费群体, 通过构建观众特征, 采用聚类方法分析高消费群体的行为. 实验结果表明, 高消费观众可被分为特征有明显差异的三类群体. 对这三类观众的特征, 本文进一步进行详细分析, 为直播平台面向用户的差异化产品服务提供依据.

关键词: 网络直播; 用户行为; 特征挖掘; 聚类分析; 数据挖掘

引用格式: 兰荣亨,朱格,杨文,田野,朱明.基于聚类的网络直播群体行为建模分析.计算机系统应用,2019,28(1):69–74. http://www.c-s-a.org.cn/1003-3254/6728.html

Modeling and Analysis of Community Behavior on Live Streaming Platform Using Clustering Approach

LAN Rong-Heng[1], ZHU Ge[2], YANG Wen[2], TIAN Ye[2], ZHU Ming[1]

[1](School of Information Science and Technology, University of Science and Technology of China, Hefei 230027, China)
[2](School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China)

Abstract: With the continuous development of Internet technology, and the popularization of mobile phones, computer tablets, and other mobile terminals, live video streaming has flourished and expanded over the past few years. Almost every live video streaming platform in China has virtual-gifts donating mechanism, which allows viewers to buy virtual-gifts provided by the platform for rewarding the broadcasters. Viewers' virtual-gifts donation is one of the most important sources of revenue for both the broadcaster and the platform, which makes it important to understand the viewer's behavior, so that it can be used to explore user's value and enhance user's liquidity. In this study, we take Douyu live video streaming platform as a case study, mainly focusing on the high consumption community on the platform. We specifically construct viewer features to analyze their behavior through clustering approach. The experiment result shows that the high consumption community can be divided into three clusters with significant difference in their behavior. We also conduct detailed analysis regarding viewer characteristics for all these three clusters, and offer suggestions for the platform to provide diversified user-oriented services.

Key words: live video broadcasting; user behaviors; feature mining; clustering analysis; data mining

① 基金项目: 国家自然科学基金 (61672486); 国家科技重大专项 (2017ZX03001019-004); 安徽省自然科学基金 (1608085MF126)
Foundation item: National Natural Science Foundation of China (61672486); National Science and Technology Major Program of China (2017ZX03001019-004); Natural Science Foundation of Anhui Province (1608085MF126)
收稿时间: 2018-07-03; 修改时间: 2018-08-13; 采用时间: 2018-08-16; csa 在线出版时间: 2018-12-26

在过去的 20 年间, 视频多媒体应用占据了英特网上大多数网络流量[1], 视频应用逐渐融入人们的日常生活. 随着宽带网络的普及, 上网费用的降低, 终端用户已经不再仅仅是内容消费者, 同时也成为了内容生产者[2]. 网络视频直播逐渐兴起并壮大, 越来越多的人参与其中, 直播并分享自己的生活. 不同于传统的文字、图片、视频等传播形式, 直播紧密的将用户与直播内容交互在一起, 用户本身也成为内容生产的一份子, 所以网络直播得到越来越多用户的推崇.

针对直播系统, 已经有大量文献进行了相关研究. Qiu 等人[3]研究发现, 频道的流行度分布是偏态的并且可以用 Zipf 分布来刻画, 作者也指出流行度的动态变化可以用 Ornstein-Uhlenbeck 过程来建模; Li[4]通过研究网络直播系统的访问日志, 提出了一系列用于直播持续时间, 用户活动, 用户的到来与离开时间建模的模型. 此外还有一系列研究直播系统生态、架构设计、用户行为等的工作[5–10].

国内的直播平台在近几年大量涌现, 在 2016 年甚至出现千播大战的局面[11]. 与国外最大的直播平台, Twitch.tv[12]的订阅收费机制不同, 国内如斗鱼、虎牙、熊猫等平台, 基本都引入了送礼机制, 即观众可以通过购买平台提供的虚拟礼物来打赏自己喜欢的主播, 而平台则以抽成的方式分享观众打赏的礼物. 观众的打赏, 成为主播和平台的主要收入来源之一. 所以, 在海量带宽、技术开发、运维等各种成本的巨大压力下, 网络直播平台理解观众的行为, 尤其是在平台上一掷千金的高消费观众, 以采取相应策略来提高用户的变现能力, 就显得尤为重要.

本文以斗鱼直播平台为例, 聚焦于在平台送出高价值礼物的观众, 通过聚类分析研究直播平台上高消费群体的行为, 为挖掘用户潜在价值提供合理依据.

1 平台简介与数据采集

1.1 平台简介

斗鱼直播平台 (Douyu.com)[13]是国内主流直播平台之一, 从最初的游戏直播平台, 至今已发展成为集游戏、娱乐、户外、体育、影视等众多热点的综合性直播平台. 平台上每个主播都有自己独立的直播间, 并可以自主选择直播内容类别. 和国内的众多视频网站一样, 斗鱼允许观众在观看直播的同时, 在直播间内发送弹幕与主播进行互动, 极大增强了观众在直播内容产生中的参与感, 这种参与感是在传统的点播和电视直播中是无法得到的. 除了发送弹幕, 斗鱼也提供了从 0.1 元到上千元价值不等的各种虚拟礼物, 供观众购买用于打赏主播.

在本研究中, 价值大于等于 100 元的礼物被称为高价值礼物. 与普通礼物不同, 当观众送出高价值礼物时, 斗鱼平台会将该事件通过弹幕的形式, 在平台所有的直播间内广播, 其他直播间的观众看到广播消息后可以通过点击广播进入该直播间. 此外, 当观众在某个直播间送出高价值礼物后, 斗鱼平台还会以该观众的名义, 在该直播间派送一些同样可用于赠送给主播的免费虚拟礼物, 所以观众在看到高价值礼物的广播消息后, 往往会点击广播进入直播间, 以领取免费的虚拟礼物.

1.2 数据采集

用户行为分析需要大量的数据支持, 如观众发送的弹幕消息, 观众产生的送礼消息等. 本研究通过维护一个每隔 5 分钟更新一次的开播直播间列表, 结合斗鱼直播平台开放的相关 API[14], 对开播房间列表中的每一个房间实时抓取弹幕消息和礼物消息并存入数据库, 完成数据的采集.

本研究采集了 2016/11/22 至 2016/12/19 连续四周的数据. 数据包含近 750 万位观众发送的 2.5 亿条弹幕数据和送出的 689 万个礼物数据, 以及 24 万个主播产生的近 179 万条开播记录. 经过简单统计, 在这四周内斗鱼观众总共送出了价值近 4700 万元的礼物. 表 1 展示了所采集数据的统计概览.

表 1 数据集概览

数据集	数据值
持续时间	28 天
主播数量	242 697 位
开播数量	1789 027 次
观众数量	7482 937 位
总弹幕消息	250 291 347 条
总礼物数量	6894 747 个
总礼物价值	RMB 46 971 922 元

2 样本选取与特征构建

2.1 样本选取

本研究旨在分析直播平台中高消费群体的行为, 所以首先需要获取高消费群体研究样本.

先给出本研究中高消费群体的定义: 在 2016/11/22

至 2016/12/19 四周内, 送出过高价值礼物, 且送出的总礼物价值超过 500 元, 则称之为高消费观众, 所有高消费观众构成高消费群体.

在不失一般性的前提下, 为了兼顾效能, 本研究根据高消费群体的定义, 从四周数据集中过滤出所有高消费观众, 并从中随机挑选了 324 名 (约占高消费观众的 10%) 观众作为研究样本.

2.2 特征构建

对于研究样本中的每一个观众, 构建如下 8 个特征:

TotalGiftValue: 观众在四周内送出的总礼物价值, 单位为元.

TotalDanmuNum: 观众在四周内发送的总弹幕数量.

BroadcasterNum: 如果某观众在某个直播间发过弹幕或者送过礼物, 称该观众与该主播产生交互. 此特征指观众与之产生交互的所有主播数量.

HGBroadcasterNum: 观众通过高价值礼物交互过的主播数量.

IfTheSame: 观众在某个直播间与主播互动有两种方式, 发弹幕和送礼物. 此特征标记观众发送弹幕最多的直播间与打赏礼物最多的直播间是否相同, 若相同则取值 1, 否则取值 0.

剩下三个特征的构建, 引入了信息论中熵 (Entropy). 在信息论中, 熵用于度量信息的不确定性, 熵越大, 则信息的不确定性越大, 即信息越发散. 记离散事件 X 的概率分布为 $p_1, p_2, \cdots, p_n$, 则该事件 X 的熵 $H(X)$ 定义如下:

$$H(X) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{1}$$

本研究采用熵的标准化形式:

$$NH(X) = \frac{H(X)}{\log(n)} = \frac{-\sum_{i=1}^{n} p_i \log(p_i)}{\log(n)} \tag{2}$$

当 n=1 时, 令 $NH(X)$=0, 则易得 0<=$NH(X)$<=1, 当 p_i=1/n, i=1, 2,⋯, n, 时, $NH(X)$ 取最大值 1.

GiftEntropy: 观众的礼物熵. 若某观众 x 在 n 个直播间送过礼物, 送出的礼物价值分别为 $g_1, g_2, \cdots, g_n$, 则可计算观众 x 的标准化礼物熵:

$$GiftEntropy(x) = \frac{-\sum_{i=1}^{n} \frac{g_i}{g} \log \frac{g_i}{g}}{\log(n)} \tag{3}$$

$$g = \sum_{i=1}^{n} g_i \tag{4}$$

观众的礼物熵越大, 意味着该观众对特定主播的送礼偏好程度越低, 换言之, 该观众越倾向于将礼物平均的送给若干主播.

DanmuEntropy: 观众的弹幕熵. 计算方式与 EntropyOfGift 类似, 用来反映观众在不同直播间的发弹幕行为偏好程度.

CategoryEntropy: 此特征反映观众对某一类直播间的偏好程度. 斗鱼直播平台中的直播内容分为热门游戏、手机游戏、娱乐天地等若干大类, 每一大类下又分为若干小类. 每一个主播都可以自主选择自己的直播间类别, 如直播游戏的英雄联盟类, 直播唱歌的音乐类等. 若观众 x 与若干类别直播间的主播产生过交互, 每一类分别有 $c_1, c_2, \cdots, c_n$ 个主播, 则类似可得该观众的交互主播类别熵:

$$CategoryEntropy(x) = \frac{-\sum_{i=1}^{n} \frac{c_i}{c} \log \frac{c_i}{c}}{\log(n)} \tag{5}$$

$$c = \sum_{i=1}^{n} c_i \tag{6}$$

表 2 简单总结了各个特征的含义. 特征构建完毕后, 研究样本中的每个观众被映射为一个 8 维的特征向量, 最终得到 324×8 维的高消费群体特征数据. 下一章节将对特征数据做聚类分析.

表 2 特征含义

特征名称	特征含义
TotalGiftValue	四周送出的总礼物价值
TotalDanmuNum	四周发出的总弹幕数量
BroadcasterNum	产生交互的主播数量
HGBroadcasterNum	通过高价值礼物交互过的主播数量
IfTheSame	送礼最多主播与发弹幕最多主播是否相同
GiftEntropy	礼物熵
DanmuEntropy	弹幕熵
CategoryEntropy	交互主播类别熵

3 模型设计与结果分析

3.1 模型设计

本研究采用无监督学习方法—聚类, 来分析高消费群体的行为. 聚类的本质是识别并区分数据中的一些代表性群体, 这些由相似个体构成的群体被称为簇

(cluster).

聚类涉及相似性度量, 由于观众的特征数据包含属性变量 (*IfTheSame*), 所以传统的欧式距离无法直接用来计算相似性, 而需要寻求其它适用混合数据的相似性度量方法. Gower[15]通过对数值特征和属性特征分别采用不同的距离度量, 最后对所有特征的距离求加权和的方式, 提供了一种用于混合数据的相似性度量方法. 记 X_i, X_j 为两个具有 N 维特征的变量, 则 Gower 距离可形式化定义如下:

$$S_{ij}=\frac{\sum_{k=1}^{N} w_{ijk}S_{ijk}}{\sum_{k=1}^{N} w_{ijk}} \tag{7}$$

其中, 若特征 k 为属性特征:

$$S_{ijk}=\begin{cases}0, & if\ X_{ik}=X_{jk}\\ 1, & if\ X_{ik}\neq X_{jk}\end{cases} \tag{8}$$

若特征 k 为数值特征:

$$S_{ijk}=\frac{|x_{ik}-x_{jk}|}{r_k} \tag{9}$$

$$r_k=\max(x_{\cdot k})-\min(x_{\cdot k}) \tag{10}$$

w_{ijk} 为赋给各个特征的权值.

可以看到, 当特征为数值型时, S_{ijk} 为曼哈顿距离 (Manhattan Distance), 且通过除以尺度因子 r_k 使其规约化到了 0~1 之间.

本研究采用 Gower 距离作为个体间的相似性度量, 采用的 PAM (Partition Around Medoids)[16]作为聚类方法.

除了相似性度量与聚类方法, 聚类的另一核心问题为聚类个数的选取. 本研究通过自定义目标函数 *Object*, 结合手肘法 (Elbow method)[17]来确定最佳聚类个数. 记 m_l 为聚类 C_l 的中心个体, 则目标函数 $Object(k)$ 定义如下:

$$Object(k)=\sum_{l=1}^{k}\sum_{j\in C_l} S_{jm_l} \tag{11}$$

算法 1 描述了整个建模过程.

算法 1. 直播平台中高消费群体行为分析算法

1) 从四周数据中过滤出所有高消费群体, 并随机选择 324 名 (10%) 观众作为研究样本 U;

2) 对于 U 中每个研究样本, 构建特征, 得到 324×8维的高消费群体特征数据 F;

3) 根据特征数据 F 计算 Gower 距离, 得到个体间相似性度量矩阵 S.

4) 根据相似性度量矩阵 S, 使用 PAM 方法对高消费群体聚类, 采用手肘法确定最佳聚类个数 k.

3.2 结果分析

根据目标函数 $Object(k)$, 作出 $Object$~k 曲线, 如图 1 所示, 可见曲线在 k=3 时出现明显拐点, 结合手肘法, 可以得到最佳聚类个数为 3.

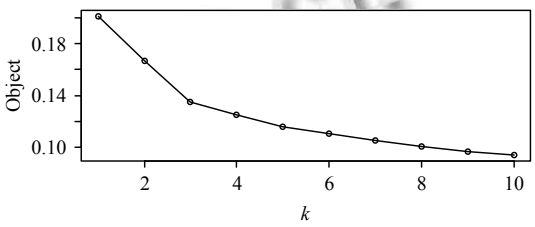

图 1 目标函数曲线

为了进一步检验聚类效果, 本研究使用 Maaten[18]等人提出的 t-SNE 高维数据可视化算法, 来直观的展示聚类结果. 如图 2 所示, 可见高消费观众可明显被聚成 3 类, 且算法也成功的区分出了这 3 类观众.

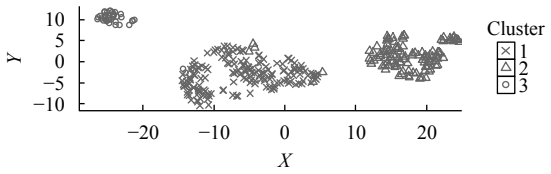

图 2 聚类结果可视化

根据聚类结果, Cluster1、Cluster2、Cluster3 分别包含 170 名、114 名、40 名观众. 计算每一类观众各个特征的统计描述, 如均值 (Mean), 分位数 (Quantile) 等, 结果如表 3 所示. 下面分析这三类观众的行为.

从四周内送出的总礼物价值来看, Cluster1 和 Cluster2 的消费能力最高, 送出总礼物价值的均值都超过了 2 万元, Cluster2 中有观众甚至在四周内送出了超过 100 万元的礼物 (1192 234 元). 相比之下, Cluster3 的消费能力则低一些, 送出总礼物价值的均值不到 1 万元. 四周发送的总弹幕量特征与送出的总礼物价值特征类似, Cluster1 和 Cluster2 发送的弹幕量远多于 Cluster3, 表明 Cluster1 和 Cluser2 中的观众总体上要比 Cluster3 活跃. 另一方面, 从表中可以看到, Cluster3 交互主播数远小于 Cluster2 和 Cluster1, 其均

值为 2, 最大交互主播数也仅为 7 个. 而 Cluser1 和 Cluster2 的平均交互主播数都是数十倍于 Cluster3. 其中 Cluster2 的交互主播数最多, 除了最大最小值, 各项指标都在 Cluster1 的两倍之上. 所以如果从平均意义上看, Cluster3 给每个房间送的礼物价值和发送的平均弹幕量都远高于 Cluster1 和 Cluster2. 这说明 Cluster3 具有很强的主播偏好性, 他们通常只在 1~2 个房间内送出大量礼物, 以及发送大量弹幕. 如果只关注观众送过高价值礼物的房间, Cluster3 的主播偏好性表现得更加明显, 他们几乎都只给一个主播送高价值礼物. 还可以看到, 虽然 Cluser1 和 Cluser2 的交互主播数量较大, 但送过高价值礼物的主播数量并不多, 他们只对约 20% 的交互主播送高价值礼物.

表 3 三类观众各个特征的统计描述

Cluster	*TotalGiftValue*	*TotalDanmuNum*	*BroadcasterNum*	*HGBroadcasterNum*	*GiftEntropy*	*DanmuEntropy*	*CategoryEntropy*	*IfTheSame*
1	Min	500.0	7	2	1	0.00	0.00	0.40	0:0 1:170
	1st Qu	2658.0	158	6	1	0.01	0.29	0.77	
	Median	6328.0	674	12	2	0.17	0.48	0.86	
	Mean	20 518.0	1479	24	5	0.21	0.46	0.83	
	3st Qu	18 366.0	1938	27	6	0.36	0.63	0.92	
	Max	254 929.0	11 768	228	51	0.87	1.00	1.00	
2	Min	500.0	0	2	1	0.00	0.00	0.35	0:109 1:5
	1st Qu	3050.0	166	10	2	0.11	0.58	0.72	
	Median	9073.0	841	30	5	0.30	0.69	0.82	
	Mean	27 851.0	1851	51	9	0.31	0.65	0.79	
	3st Qu	19 070.0	2317	72	12	0.46	0.75	0.90	
	Max	119 2234.0	12 285	282	68	0.82	0.91	1.00	
3	Min	500.0	0	1	1	0.00	0.00	0.00	0:1 1:39
	1st Qu	976.1	1	1	1	0.00	0.00	0.00	
	Median	4284.0	21	1	1	0.00	0.00	0.00	
	Mean	9099.7	115	2	1	0.04	0.13	0.05	
	3st Qu	11 722.9	115	2	1	0.00	0.14	0.00	
	Max	68 300.0	1099	7	4	0.54	0.85	0.81	

进一步观察礼物熵、弹幕熵和类别熵三个特征, 可以得到 Cluster1 的礼物熵和弹幕熵均小于 Cluster2, 这表明与 Cluster2 相比, Cluster1 中观众对某一小部分主播的偏好性会较强一些. 横向对比礼物熵和弹幕熵两个特征, 可以看到, 礼物熵要明显小于弹幕熵, 这说明和发弹幕这种几乎不耗费成本的行为相比, 观众对金钱的分配则更为慎重, 他们会将金钱打赏给那些真正喜欢的主播. 而对于类别熵, Cluster1 与 Cluster2 相近, 并且数值都较大, 可见他们对直播间类别并无明显偏好, 即他们的偏好性是面向主播, 而非类别. 由于 Cluster3 中观众的交互主播几乎都只有 1~2 个, 所以 Cluster3 的礼物熵、弹幕熵及类别熵都很小, 几乎都为 0, 这与前面得出 Cluster3 中观众具有强偏好性的结论是一致的.

最后一个特征一定程度上反映观众送礼行为和发弹幕行为的一致性. 可以看到 Cluster1 和 Cluster3 中观众的这个特征几乎都取值为 1(只有 Cluster3 中的一个观众取值为 0), 即观众送礼最多的主播与发弹幕最多的主播相同. 有趣的是, Cluster2 中观众的这个特征取值基本为 0, 这反映了 Cluster2 中观众两种行为的不一致性. 对这类观众而言, 他们可能在某个直播间很活跃, 发送大量弹幕与主播交互, 但他们不一定想用金钱支持与推广这个主播.

总结分析结果, 得到 Cluster1、Cluster2、Cluster3 的特征如下:

Cluster1: 消费能力较高, 活跃较多房间, 对其中一些主播有一定偏好性, 金钱基本用于打赏这些偏好的主播, 送礼行为与发弹幕行为表现一致性;

Cluster2: 消费能力最高, 活跃在大量房间, 对很多的主播感兴趣, 金钱用于打赏较多主播, 送礼行为与发弹幕行为表现不一致;

Cluster3: 消费能力较低, 只在 1~2 个房间内活跃, 金钱基本只用于打赏 1 个主播, 对某特定主播具有强偏好性, 送礼行为与发弹幕行为表现一致性.

聚类结果的一个很重要的用途, 在于辅助用户分析, 挖掘用户的潜在价值, 从而提升平台的用户变现能力. 如对于 Cluster1 和 Cluster3 中观众, 利用他们对某一些甚至一个主播的偏好性, 平台可以和这些观众偏好的主播合作, 让主播为其他产品做营销推广、广告植入, 以实现精准挖掘用户消费能力. Cluser2 中的观众具有最高消费能力且主播偏好性较低, 利用这点, 平台可以通过他们的历史观看信息, 挖掘这些观众的兴趣, 向他们推荐类似的主播, 从而进一步刺激用户消费能力.

4 结论与展望

本文以斗鱼平台为例, 研究了直播平台高消费群体的行为. 利用从斗鱼平台抓取的直播数据, 构建观众特征, 使用 Gower 距离度量混合特征的相似性, 并采用 PAM 聚类方法对高消费群体做聚类分析. 结果表明, 高消费群体可以明显被聚成三类. 对聚类结果中的三类观众做详细分析, 得到了三类观众的特征刻画, 并简单介绍了如何利用聚类结果来提升平台用户的变现能力. 在聚类特征的选择上, 本研究人工选择了文中提到的 8 个特征用于模型训练. 接下来的工作是尝试挖掘更多的用户特征, 并引入机器学习中的特征选择方法自动选择最佳特征组合, 进一步改善聚类效果.

参考文献

1 Li BC, Wang Z, Liu JC, *et al*. Two decades of Internet video streaming: A retrospective view. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2013, 9(S1): 33.

2 He QY, Liu JC, Wang CG, *et al*. Coping with heterogeneous video contributors and viewers in crowdsourced live streaming: A cloud-based approach. IEEE Transactions on Multimedia, 2016, 18(5): 916–928. [doi: 10.1109/TMM.2016.2544698]

3 Qiu TQ, Ge ZH, Lee S, *et al*. Modeling channel popularity dynamics in a large IPTV system. ACM SIGMETRICS Performance Evaluation Review, 2009, 37(1): 275–286.

4 Li ZY, Kaafar MA, Salamatian K, *et al*. Characterizing and modeling user behavior in a large-scale mobile live streaming system. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(12): 2675–2686. [doi: 10.1109/TCSVT.2016.2595325]

5 周传华, 江超, 赵伟. 混合云可扩展视频编码的视频直播机制研究. 计算机系统应用, 2017, 26(7): 258–262. [doi: 10.15888/j.cnki.csa.005832]

6 崔虹燕. P2P 视频直播系统中的分布式负载均衡算法. 计算机系统应用, 2009, 18(12): 95–97. [doi: 10.3969/j.issn.1003-3254.2009.12.023]

7 Zhang C, Liu JC. On crowdsourced interactive live streaming: A Twitch.tv-based measurement study. Proceedings of the 25th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video. Portland, OR, USA. 2015. 55–60.

8 Wang BL, Zhang XY, Wang GG, *et al*. Anatomy of a personalized livestreaming system. Proceedings of 2016 Internet Measurement Conference. Santa Monica, CA, USA. 2016. 485–498.

9 Kaytoue M, Silva A, Cerf L, *et al*. Watch me playing, I am a professional: A first study on video game live streaming. Proceedings of the 21st International Conference on World Wide Web. Lyon, France. 2012. 1181–1188.

10 Elhabian SY, El-Sayed KM, Ahmed SH. Moving object detection in spatial domain using background removal techniques-state-of-art. Recent Patents on Computer Science, 2008, 1(1): 32–54. [doi: 10.2174/2213275910801010032]

11 千播大战过去后，直播还是风口么？ https://36kr.com/p/5061698.html [2018-04-21]

12 Twitch. tv. https://www.twitch.tv/. [2018-04-12]

13 斗鱼. https://www.douyu.com/. [2018-04-12]

14 斗鱼 API 文档. https://coapi.douyucdn.cn/125.html. [2018-04-12]

15 Gower JC. A general coefficient of similarity and some of its properties. Biometrics, 1971, 27(4): 857–871. [doi: 10.2307/2528823]

16 Kaufman L, Rousseeuw PJ. Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, NJ: John Wiley & Sons, 2005.

17 Elbow method. https://en.wikipedia.org/wiki/Elbow_method_(clustering). [2018-04-12]

18 Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9: 2579–2605.