BOHR International Journal of Advances in Management Research
2023, Vol. 2, No. 1, pp. 61–67
DOI: 10.54646/bijamr.2023.18
www.bohrpub.com

METHODS

Identification and analysis of low creators in bilibili video based on multivariate statistics

Tingting Liu*, Zhoulei Pan, Jibo Chen and Jiaqi Meng

[1]School of Management, Tianjin University of Technology, Tianjin, China

***Correspondence:**
Tingting Liu,
jnybv39@163.com

Received: 10 June 2023; **Accepted:** 17 June 2023; **Published:** 24 June 2023

Bilibili video website has grown into a giant video platform. With the anime culture that can attract the younger generation, Bilibili has built a large-scale user creation platform. To stimulate users' creative inspiration, Bilibili issued several plans to provide corresponding rewards for the content produced by video creators, attracting more and more people to participate in the creative party. In this context, many excellent works were born, but at the same time, and there are also works with mixed qualities in the video, i.e., "low-innovation" works. "Low-innovation" works hinder personal development and have a bad impact on the production climate of the platform. First, this paper uses the principal component analysis algorithm to preprocess the user data of Bilibili to improve the efficiency of the algorithm. Based on the K-means clustering algorithm, it analyzes and identifies "low-innovation" users. According to the analysis results, it sets different incentive plans for different types of user groups and plays a positive role in the video quality of Bilibili.

Keywords: user identification, Bilibili, clustering, "low innovation"

1. Introduction

The platform industry is a relatively new global industry (1). Bilibili is now the leading youth cultural community in China. The website was launched on June 26, 2009, and was affectionately called "Bilibili" by fans. Since its founding, Bilibili has gradually grown into a "small giant" of the video platform by virtue of its unique bullet screen culture. In 2020, the total revenue of Bilibili, which has been born for nearly 10 years, reached 12 billion yuan. By 2022, its market value has reached $9.358 billion.

One of the core competitive advantages of Bilibili is high-quality content. Based on this, Bilibili tried and successfully built a UGC ecosystem (2). UGC is user-oriented content, which is different from the traditional marketing model. Customer-based marketing network has the characteristics of interactivity, participation, connectivity, openness, and community, and at present, it is recognized by the vast majority of groups. Based on the analysis of the UGC characteristics of the brand and the fierce competition in the domestic UGC video industry, Bilibili is increasingly dependent on the UP master (short video creator) in operation (3). In order to attract more users to create and improve the quality of video content, Bilibili put forward the "Creator Incentive Plan" in 2018 which continues to this day. Its content is upgraded to creators with less than 500,000 fans who can get special promotion bonuses by achieving small goals. Bilibili held such incentive creative activities to stimulate UP owners to produce more high-quality content. The most typical example is the famous anime Spring Festival Gala. More than two-thirds of the broad cast volume of Bilibili comes from user-made and original videos, and there are more than 1 million active creators.

Despite such a large user scale, Bilibili still faces a big problem: many interest drivers only attach importance to exaggerated covers and titles and other means to increase the playback volume while ignoring the quality of video content. Low-innovation video refers to simply splicing and editing network material and adding background music or changing another background music for another person's

CC BY NC ND

secondary creation. Due to its low creation threshold and low production cost, the output of low-innovation video is very large. A user-generated video can help predict viewers‘ online viewing behavior (4). Due to the equal opportunities of video push, the low-innovation video occupies a large amount of space on the platform, resulting in the suppression of high-quality works, and the potential impact on the community environment of the video platform. This is contrary to the original intention of Bilibili to launch the creative incentive plan.

Therefore, only by accurately identifying low-innovation users and strengthening the management system of creators, better video content can be produced and combined with revenue incentives to drive creators to create better works, thus forming a virtuous circle of video websites. Based on the above objectives, this paper triggers from the point of view of extracting the behavioral characteristics of the users in Bilibili, introduces the factor analysis algorithm to sort the characteristics of the low-innovation users, extracts the main characteristics of the low-innovation users, and then carries out user clustering to realize the recognition and division of the user group in Bilibili. While solving the problem of how the video platform can identify low-innovation users, different incentive plans are proposed to accurately guide the rise of creators with different characteristics and reasonably allocate channel resources for users to rise.

Section 2 of this paper introduces the existing research results at home and abroad from relevant fields. Section 3 introduces the use of factor analysis to reduce the dimension of user data. Section 4 uses the clustering algorithm to identify low-innovation users. Section 5 summarizes the analysis results and puts forward suggestions for commercial application.

2. Literature review

There have been many studies on the user behavior of Bilibili. Rong et al. (5) and others revealed that appropriability resources are necessary for the user engagement of short video platforms. Liu et al. (6) analyzed the variables including the content and function of the barrage video websites that significantly impact the user experience. Zhang and Wu(7) investigated the game video creators‘ playbour on Bilibili and argues that video creators can barely escape from the unequal labor relationship, constructed by the platform. To sum up, the theoretical research on the behavior characteristics of Bilibili is relatively rich, while the empirical research is still deficient. This paper crawls the real-time data of Bilibili to identify the users of Bilibili, and the results are more authentic, which can better provide professional suggestions for the platform managers.

In this paper, factor and cluster analyzes are used to realize the recognition of user features. Some scholars have studied the research based on these two methods in the field of user feature recognition. Meng and Sun (8) through social network analysis combined with cluster analysis to explore the heterogeneity of user interaction behavior and contribution behavior and finally establish five different user roles. Al-Durgham and Barghash (9) used factor and cluster analyzes as a tool for patient segmentation applied to hospital marketing in Jordan. Liu et al. (10) proposed a new algorithm for identifying users of wild calling based on cluster analysis. From the above, the research system of user identification using clustering and factor analysis methods is mature, and scholars have not applied these two methods to the user data of Bilibili. As the mainstream video platform in China, the economic benefits of Bilibili cannot be underestimated. Therefore, the economic significance of the analysis of Bilibili data in this paper is more profound, which can make up for the gap in the current research system.

3. Feature extraction of low invasive users based on factor analysis

3.1. Data description

The data used in this paper are crawled from the website where the user information is stored in Bilibili. The sample contains 8,579 user information, which has been counted since the second half of 2022. The specific variable description is shown in **Table 1**. In addition, this paper also establishes derivative variables to describe the quality level of creators' works: a_fans = fans/archives, a_like = like_num/archive, and a_view = view/vcount. This paper selects 11 evaluation indicators from three aspects of user's personal information, user's popularity, and user's creative level to analyze the creative characteristics of Bilibili users. The specific variable description is shown in **Table 1**.

3.2. Factor analysis results

Factor analysis uses dimensionality reduction technology to reduce complex original variables to a few common factors, so as to draw the correlation between the original variables. Unlike principal components, the common factors constructed by factor analysis are more clear and easy to explain.

3.2.1. Applicability test

Standardize the four variables that describe the popularity of users and the level of users' creation and the two quantitative variables that concern the number of contributions, and use the KMO test and Bartlett spherical test to test the applicability of user data in Bilibili. The results show that the KMO value = 0.604, which is greater than the

critical value of 0.6. If Bartlett's spherical test approximation χ^2value = 36,298.633, degree of freedom is 28, and significance P = 0.000, then reject the original assumption that the covariance matrix is the unit matrix, indicating that there is a strong correlation between the variables, indicating that the data containing each variable is suitable for factor analysis.

TABLE 1 | Variable description.

Variable type	Variable name	Value range	Remarks
User profile	sex	0: male, 1: female, and 2: confidential	Gender and qualitative variable
	level	0, 1, 2, 3, 4, 5, and 6	Grade and qualitative variable
	vtype	1: none, 2: monthly general member, and 3: Annual and above major members	Member type, qualitative variable (1 represents never being a member, and excluding expired)
	vcount	0∼20,000	Number of contributions and quantitative variable (unit: piece)
	following	0∼2,000	Number of concerns and quantitative variable (unit: piece)
User popularity	fans	0∼4E + 06	Number of fans and quantitative variable (unit: piece)
	likes	0∼3E + 07	Number of likes and quantitative variable (unit: piece)
	view	0∼4E + 08	Playback volume and quantitative variable (unit: piece)
User authoring level	a_ fans	0∼9E + 04	Fans mean and quantitative variable
	a_ like	0∼2E + 06	Liked mean and quantitative variable
	a_view	0∼6E + 06	View mean and quantitative variable

TABLE 2 | Correlation coefficient matrix.

	Vcount	Fans	Likes	Following	View	A_view
Vcount	1.000					
Fans	0.156	1.000				
Likes	0.174	0.845	1.000			
Following	0.020	0.022	0.038	1.000		
View	0.583	0.532	0.471	0.008	1.000	
A_view	0.009	0.334	0.185	−0.008	0.308	1.000

3.2.2. Extract common factors

Calculate the simple correlation coefficient of the variables used above, and obtain the correlation coefficient matrix to describe the correlation degree between variables. The correlation coefficient matrix is shown in **Table 2**. The matrix results show that the total playback volume has a strong

TABLE 3 | Characteristic root and variance contribution of each factor before rotation.

Principal component	Characteristic root	Variance contribution rate (%)	Cumulative contribution rate (%)
F_1	3.389	0.424	42.36
F_2	1.511	0.189	61.24
F_3	1.006	0.127	73.81
F_4	0.903	0.113	85.10

TABLE 4 | Characteristic root and variance contribution of each factor after rotation.

Principal component	Characteristic root	Variance contribution rate (%)	Cumulative contribution rate (%)
F_1	2.234	27.92	27.92
F_2	1.001	12.52	40.44
F_3	1.955	24.44	64.88
F_4	1.618	20.22	85.10

TABLE 5 | Factor load matrix after rotation.

	Factor 1	Factor 2	Factor 3	Factor 4
Vcount'	0.0317	0.0246	−0.0487	0.8829
View	0.3425	−0.0160	0.1699	0.8318
Likes	0.9197	0.0178	0.0727	0.2046
Following	0.0153	0.9991	0.0004	0.0122
Fans	0.8725	−0.0004	0.3059	0.2153
A_vie	−0.0357	−0.0228	0.8631	0.2010
A_like	0.5546	0.0328	0.6816	−0.1104
A_fans	0.4453	0.0136	0.7845	−0.0733

TABLE 6 | Factor score.

	User feedback factor	Hobbies multiple factor	Video quality factor	User impact factor
Sample 1	1.697	0.004	2.152	−1.108
Sample 2	0.301	4.527	−0.620	4.970
Sample 3	4.897	−0.357	2.034	1.762
Sample 4	−8.395	3.458	−0.109	−0.085

correlation with the number of fans and likes, indicating that users tend to pay more attention to and like high-quality works.

There is a high correlation between the number of fans and the number of likes, which indicates that high-quality creative works often make users prefer to make attention and likes. The correlation coefficient between the total playback volume and the number of contributions is high, and the correlation between the total playback volume and the average number of fans, the average number of playback, and the average number of likes is low, which indicates that the overall data of users with a large number of contributions is considerable, but it is not equal to the average creation level of a single video. There may be some users who mainly focus on quantity and create a large number of inferior videos through a few gags on the cover of works or by moving other website resources. Due to the lack of good content and meaning of videos, the average number of fans and the average number of liked videos are not high, which is one of the representative characteristics of low-innovation users.

According to **Table 3**, the principal component analysis method is used to calculate the component matrix and characteristic root and select the components with characteristic root greater than 1, namely, the first three common factors, with a cumulative contribution rate of 85.10%. It shows that the first three common factors can explain 85.10% of the total variance. According to the principle that the cumulative variance contribution rate is greater than 80% (11), the information contained in the sample can be basically extracted. On this basis, in order to better explain the meaning of the factor, the factor rotation is carried out, and the variance maximization orthogonal rotation is adopted. The results are shown in **Table 4**. The number of common factors is determined to be 4, which are, respectively, recorded as F_1, F_2, F_3, and F_4. The factor load matrix after rotation is shown in **Table 5**.

First factor F_1—the load on the number of fans and the number of likes is high. These two variables reflect the comprehensive evaluation of creators and are related to the overall level of creators. Therefore, this factor F_1 is defined as a user feedback factor. Second factor F_2 in the field of high attention load, which mainly reflects the hobbies of users in Bilibili. F_2 is defined as a hobby multiple factor. Third factor F_3—the average load of fans, likes, and plays is high, which mainly reflects the user's creative level. F_3 is defined as the video quality factor. Fourth factor F_4—the load of the total number of submissions and the total broadcast volume is high, which mainly reflects the production scale of users. F_4 is defined as a user influence factor. According to the weighted average of the four common factor scores and the contribution rate, the result is the comprehensive score calculation formula:

$$F=\left(F_1 \times 0.2792+F_2 \times 0.1252+F_3 \times 0.2444+F_4 \times 0.2022\right)/0.8510$$

The maximum-likelihood method factor analysis of regression estimation method is used to solve the factor scores of all samples and extract representative samples with obvious characteristics. The results are shown in **Table 6**.

It can be seen that the score of the user impact factor in Sample 1 is low, but the score of the user feedback factor and video quality factor is high, which indicates that the works created by such users are of high quality and can attract more fans with less contributions. They should be users who can promote the upgrading of video quality and content in Bilibili and belong to high-quality creators. Therefore, such users are called "high-tech" users. The user impact factor score of Sample 2 is high, but the user feedback factor and video quality factor score are low, showing the characteristics of high comprehensive evaluation, low fan retention rate, and low praise rate of a single video. Such users can be defined as "low-innovation users." In Sample 3, the scores of three factors are high, except for the love factor, which indicates that such users are highly popular and professional. This type of user is often able to obtain a lot of economic benefits from Bilibili and also bring traffic to Bilibili. This article defines this type of user as "UP master" user. In Sample 4, only the multiple factor of hobbies has a high score. Such users often only browse the videos posted on Bilibili without producing original works. Such users are defined as "transparent" users. The rest of the samples have no obvious features or the meaning of their index features is not clear, and the composition is complex. In fact, they occupy a large part of the user composition. This category of users is defined as "ordinary" users.

4. User classification and recognition based on cluster analysis

4.1. KMeans clustering algorithm introduction

After factor analysis, cluster analysis is carried out on the samples, and the clustering diagram can be drawn to make the analysis results more intuitive.

Clustering technology is to divide the classified objects into several categories based on certain rules. It is the most commonly used method for user identification analysis. KMeans clustering divides the original data set into K groups according to the similarity of service characteristics. The input of KMeans clustering includes the number of categories K and training sample data, and the output is the group set of K categories (12). KMeans is a classic algorithm to solve the clustering analysis problem (13). When the sample size is large, the results obtained by this method are fast and easy to understand, so it is widely used.

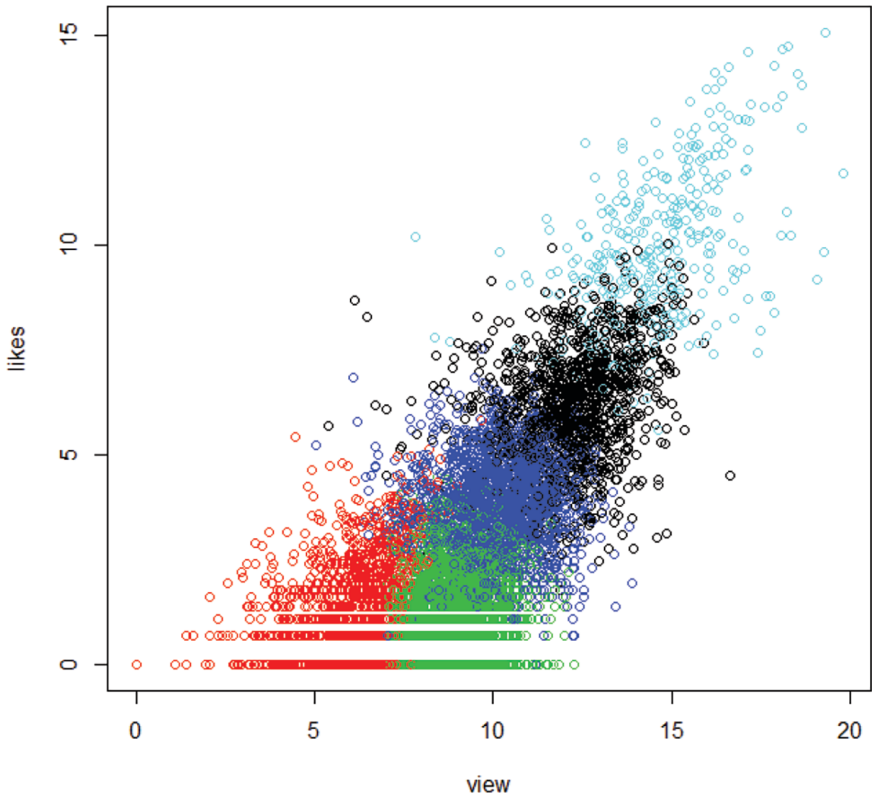

FIGURE 1 | Clustering results of number of plays (logarithm) and number of likes (logarithm).

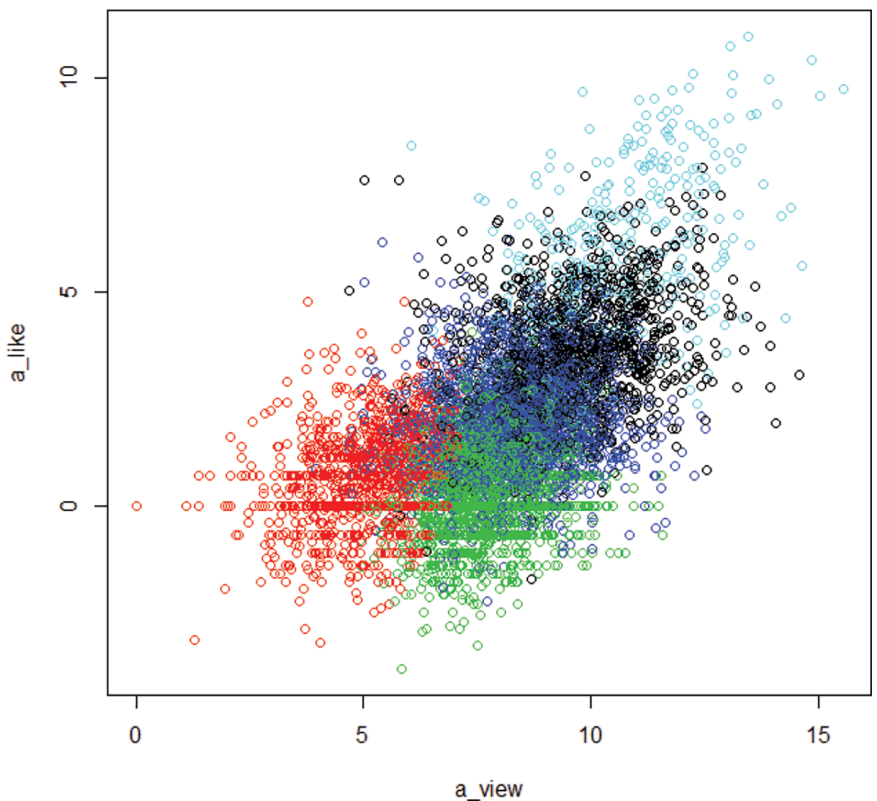

FIGURE 2 | Clustering results of average number of plays (logarithm) and average number of likes (logarithm).

4.2. KMeans clustering algorithm process

Clustering is a kind of unsupervised learning method, which aims to make the similarity within clusters higher, while the similarity between clusters lowers. Considering that the data set in this paper contains both qualitative and quantitative variables, the order of magnitude between the data is large. In the process of clustering, all quantitative variables are logarithmically transformed.

The KMeans algorithm includes the following steps.

Step 1: Divide the initial sample into K classes. According to the results of factor analysis in Section 3, the value of K is set to 5. Namely, five samples are randomly selected as the initial class.

Step 2: Calculate the sum of squares of the distance from all samples to the center of the class through Euclidean distance, divide each sample into the nearest class to the center, and recalculate the center target for the new class.

Step 3: Repeat Step 2 until all samples can no longer be classified.

After using factor analysis to reduce the data dimension, the four common factor scores of all samples are used as new variables for cluster analysis, and the KMeans clustering method is selected. A total of 8,579 samples are divided into five categories by clustering, and the number of samples is 11, 52, 953, 18, and 7,545, respectively. **Table 5** shows the final average value generated by each column value in each cluster. Based on the conclusion of factor analysis, the first category of users is characterized by a high creation level and large audience size and belongs to high-quality creators, so this category of users is "high-quality" users. The second type of users presents the characteristics of wide hobbies but low creative intention or low creative level and usually only appreciate the videos of Bilibili. Therefore, such users are "transparent" users. The third category of users is characterized by high playback volume and low single-video data. This category of users is "low-innovation" users. The fourth category of users has a large scale of fans, high video quality, and has formed an operational production team. This category of users is often " UP master " users. The fifth category of users has no obvious characteristics, but the number of samples accounts for 87.95% of the total sample, which is a relatively high proportion of ordinary users.

TABLE 7 | Clustering results.

	F_1	F_2	F_3	F_4
1	23.457	0.043	4.346	2.560
2	−0.052	8.522	−0.019	0.086
3	−0.002	1.574	−0.023	0.031
4	−2.980	−0.273	11.767	9.591
5	−0.026	−0.257	−0.031	−0.031

According to the results, Bilibili has the highest proportion of ordinary users, up to 87.95%. The second is low-innovation users, accounting for 11.11%. Although the proportion of low-innovation users is relatively small, the number of them is far more than that of UP master users and high-innovation users, which indicates that the prospect of Bilibili is not optimistic, and the situation of being driven by interests to make up the number is not significant, which will put Bilibili in a disadvantage. Therefore, Bilibili should subdivide the user's creative ability as soon as possible, set incentives differently, ensure the quality of works, and create a good atmosphere for creation.

4.3. Cross-analysis of various users

In order to visually display the characteristics of the five categories of users, **Figures 1**, **2** show the distribution of the four discriminating factors on the five categories of users: liked number, played number, liked average number, and played average number.

From **Table 7**, using the clustering model to subdivide users can better identify "low-innovation" users, "high-innovation" users, "transparent" users, and "UP master" users and can effectively achieve the research goal of this paper: identify and classify the "low-innovation" users of Bilibili.

It can be seen from the figures that the low-innovation users are similar to the "UP master" users in terms of playback volume and average playback volume, but they are similar to transparent users in terms of average liked volume and liked volume. This shows that the reason for low user turnover is probably low video quality, low fan conversion rate, and low retention rate. Ordinary users are similar to high-tech innovation users in terms of average playback and volume, but there are obvious differences in terms of average and volume of likes between ordinary users and high-tech innovation users, indicating that ordinary users cannot improve their average creative level even if they increase their playback volume by increasing the number of contributions. There is no obvious difference between the average broadcast and fans of high-tech users and "UP master." The difference is mainly reflected in the overall number of likes. For high-tech users, if they want to become influential and have a large number of fans, they can expand their popularity and attract more traffic fans by increasing their contributions.

5. Conclusions and suggestions

This article mainly starts from two aspects: (1) monitor and identify "low-innovation" users and (2) set stricter incentive thresholds for "low-innovation" users.

1) Monitoring and identification: identify "low-innovation" users through the clustering model.

For users who have low average fans and low average but still produce a large number of videos, appropriately recommend them to participate in some incentive activities launched by Bilibili, or delete some of their "low-innovation" videos to improve the overall video quality of Bilibili.

2) Incentive measures: propose different incentive measures for different user characteristics.

For "low-innovation" users, you can set the threshold value of the average value of praise and the average value of fans in the incentive plan, so as to encourage these users to improve the quality of their works, rather than just increasing the amount of their works played and the number of works.

For ordinary users, we should focus on their video quality. Because these users have the characteristics of high fans and high likes, they are the most likely to become UP master users. At the same time, it is noted that the average number of fans and the average number of likes of these users are low, which indicates that the video quality of these users is uneven. Therefore, it is necessary to strengthen the video quality monitoring of these users, so as to improve the overall video quality of Bilibili.

References

1. Xie X-Z, Tsai N-C, Xu S-Q, Zhang B-Y. Does customer co-creation value lead to electronic word-of-mouth? An empirical study on the short-video platform industry. *Soc Sci J.* (2018) 8:10.
2. Bahtar AZ, Muda M. The Impact of User – Generated Content (UGC) on product reviews towards online purchasing – a conceptual framework. *Procedia Econ Finan.* (2016) 37:337–42.
3. Jia AL, Shen S, Li D, Chen S. Predicting the implicit and the explicit video popularity in a User Generated Content site with enhanced social features. *Comput Netw.* (2018) 140:112–25.
4. Xi D, Xu W, Chen R, Zhou Y, Yang Z. Sending or not? A multimodal framework for Danmaku comment prediction. *Inform Proc Manag.* (2021) 58:102687.
5. Liu Z, Zhou C, Chen H, Zhao R. Impact of cost uncertainty on supply chain competition under different confidence levels. *Int Trans Operat Res.* (2021) 28:1465–504.
6. Zhou C, Tang W, Zhao R. Optimal consumption with reference-dependent preferences in on-the-job search and savings. *J Ind Manag Optim.* (2017) 13:503–27.
7. Rong K, Xiao F, Zhang X, Wang J. Platform strategies and user stickiness in the online video industry. *Technol Forecast Soc Chang.* (2019) 1:23.
8. Zhou C, Xu G, Liu Z. Incentive contract design for internet referral services: cost per click vs cost per sale. *Kybernetes.* (2020) 49:601–26.
9. Liu W, He Z, Liu M. *An empirical study of the influencing factors on user experience for barrage video website - a case study of bilibili*. Cham: Springer (2021).
10. Zhang N, Wu Y. Platformed playworkers: game video creators' affective labour and emotional labour on Bilibili. *Glob Media China.* (2022) 7:319–39.
11. Zhou C, Leng M, Liu Z, Xin C, Jing Y. The impact of recommender systems and pricing strategies on brand competition and consumer search. *Electron Commerce Res Appl.* (2022) 53:1–15.
12. Yu J, Zhao J, Zhou C, Ren Y. Strategic business mode choices for e-commerce platforms under brand competition. *J Theor Appl Electron Commerce Res.* (2022) 17:1769–90.
13. Meng Q, Sun N. Research on User Roles Identification of Crowdsourcing Innovation Virtual Community. *J Serv Sci Manag.* (2019) 12:421–38.
14. Yu J, Song Z, Zhou C. Self-supporting or third-party? The optimal delivery strategy selection decision for e-tailers under competition. *Kybernetes.* (2022). doi: 10.1108/K-02-2022-0216 [Epub ahead of print].
15. Al-Durgham LM, Barghash MA. Factor and cluster analysis as a tool for patient segmentation applied to hospital marketing in jordan. *Am J Operat Res.* (2015) 5:293–306.
16. Zhou C, Tang W, Zhao R. Optimal consumer search with prospect utility in hybrid uncertain environment. *J Uncertain Anal Appl.* (2015) 3:1–20.
17. Liu XY, Wang SY, Wang SY, Wang B, Bai ZY. Research on identification model of wild call user based on cluster analysis algorithm. *Telecom Eng Tech Stand.* (2022) 35:18–21.
18. Zhou C, Ma N, Cui X, Liu Z. The impact of online referral on brand market strategies with consumer search and spillover effect. *Soft Comput.* (2020) 24:2551–65.
19. Chu M, Zhou C, Yu J. The impact of online referral services on cooperation modes between brander and platform. *J Ind Manag Optim.* (2022). 19:5306–30. doi: 10.3934/jimo.2022174
20. Liu Z, Gao R, Zhou C, Ma N. Two-period pricing and strategy choice for a supply chain with dual uncertain information under different profit risk levels. *Comput Ind Eng.* (2019) 136:173–86.
21. Poole D. *Mack worth AK. Artifical Intelligence; Foundations of Computational. Agents*. Cambridge, UK: Cambridge University Press (2010).
22. Karcamarek P, Kiersztyn A, Pedrycz W, Al E. K-Means-based Isolation Forest. *Knowl Based Syst.* (2020) 195:105659.