Research on User Profile of Knowledge Video - Taking Bilibili as an Example

Donghong Yang, Luyao Guo*

College of Economics and Management, Northeast Petroleum University, Daqing, China

*E-mail:1169675129@qq.com

Abstract. Constructing user profiles from relevant information and textual features in the comment sections of the B station platform enables the platform to understand user needs and preferences, thereby improving content, services, and enabling more precise marketing. From each subcategory under the knowledge-based video category, 8 sample videos were selected, yielding 152,823 comments from 104,811 users. These comments were statistically analyzed across various dimensions, including emotional characteristics derived through sentiment analysis. User groups were then clustered based on the similarity values and textual themes of the comments, using text similarity analysis. Based on the clustering results and additional features, user groups were categorized into three distinct segments, whose characteristics were visualized using word cloud diagrams.

Keywords: B station platform; knowledge-based video content; user profile construction; textual analysis.

1 Introduction

In recent years, with the extensive use of online video websites, various types of video content such as information, education, music, knowledge, food, etc. began to emerge, among which knowledge-based videos emerged and gradually flourished [1]. However, with the surge in the number of knowledge-based videos, it also faces some problems: the quality of the content varies, the production threshold is low, and the phenomenon of excessive entertainment is gradually emerging. The content pushed on the home page after users open the software is often difficult to arouse interest, resulting in lower user viscosity. Therefore, there is an urgent need to solve the problem of how to collect and organise user data and personalised recommendation. In this context, user profiling, as an increasingly mature data analysis tool, has become a key step in solving the problem, and can provide brand new ideas for the study of personalised recommendation and precision marketing in online video communities [2].

In this paper, based on the research of previous scholars, we construct a model for the portrait of users in the comment area of knowledge-based videos on B station, analyze their basic attributes and behavioral attributes, and carve an all-around user

Z. Chen et al. (eds.), *Proceedings of the 2024 International Conference on Humanities, Arts, and Cultural Industry Development (HACID 2024)*, Advances in Social Science, Education and Humanities Research 861,
https://doi.org/10.2991/978-2-38476-281-1_33

portrait, so that B station platform managers can intuitively understand the whole picture of users in the online video community, and the results of the research can promote platform managers to improve personalized services, optimize precision marketing strategies, provide targeted services for the users and optimise the video media content ecosystem.

2 Data Processing and Modelling System Construction

2.1 Data Collection

This study utilizes the third-party IFANS platform (B station version) and integrates data from the popular videos on the B station platform to select samples with high engagement and numerous comments in the knowledge category. A total of 8 video samples were selected and numerically labeled as shown in Table 1. Python was employed to develop a crawler for extracting user data from the comment sections, including username, comment time, gender, number of likes on the comment, comment content, RID (unique comment identifier), number of replies, user ID, level, and membership status.

Table 1. Video Selection Details

Video Category	Video ID	Video Title
Science Popularization	S1	Secrets of Myopia Surgery Revealed by High-Speed Camera: What Really Happens to the Eye?
	S2	A Diamond Ring Bought for 18,000 Ten Years Ago Now Sells for Only 180, Depreciating 99%
Business and Finance	F1	Pinduoduo's "Refund Only" Policy: How Many Have Been Angered?[Bad Review]
	F2	[In-Depth]Understanding National Debt: Can We Keep Borrowing Like This?
Campus Learning	C1	Full Series of "Probability Theory and Mathematical Statistics" (Song Hao)
	C2	Data Analysis "Line Test" System Course (Nationwide) — Liu Wen Chao
Professional Workplace	W1	Expert Dialogues: Mere Seconds to React, Zhang Zhongping's Resource Integration Makes All Parties Dignified, Flawless Speech #SocialSkills
	W2	[Half-Buddha]Job-Hunting Guide for Maximizing Benefits for New Graduates

2.2 Data Processing

The initial step involves cleansing the acquired data to eliminate any invalid entries. The second step entails preprocessing the text content—specifically, the comment data—through data cleaning, Chinese word segmentation, de-duplication, and other necessary processing. The final step implements varying processing techniques tailored to the specific methods required by the study. For sentiment analysis, the Python library SnowNLP is utilized to perform iterative sentiment assessments[3]. For user clustering, deep learning techniques are applied, specifically BERT (Bidirectional Encoder Representations from Transformers) for sentence embedding[4]. This method transforms text sentences into vector representations, uses cosine similarity to calculate distances, derives similarity values, and integrates with LDA topic analysis to categorize and match the comment text topics expressed by different users.

2.3 Establishment of a Multi-dimensional Labelling System

Drawing on the interactive context characteristics of B station community users, and adhering to the label construction principle of "on-demand design without infinite refinement"[5], this study establishes a labeling framework for user profile construction. As shown in Fig. 1. It subdivides indices into three dimensions: basic user attributes, behavioral user attributes.

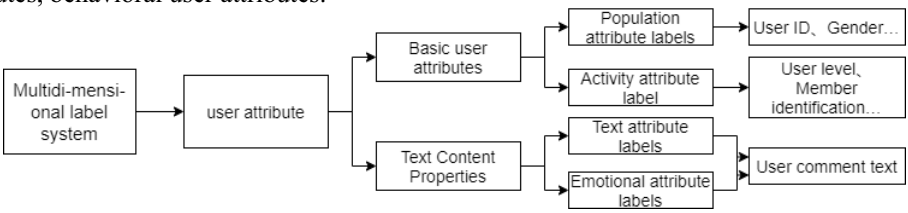

Fig. 1. Multi-dimensional labelling system diagram

3 B Station Knowledge Video User Profile Analysis

3.1 User Dimension

3.1.1 Grade Distribution

The user grade level reflects both the duration of users' memberships at B station and their activity levels. Analysis of user samples from the comment section revealed that 3,128 users (3%) were grade 2, 9,709 users (9.3%) were grade 3, and 91,513 users (87.7%) were grade 4 and above. The majority, 45,264 users, were grade 5, indicating that most users are either long-term or highly active recent registrants familiar with B station's community atmosphere. Level 6 users comprised 27.4% of the total, suggesting that a significant portion of the knowledge category audience consists of B station "natives."

3.1.2 Gender Distribution

Descriptive statistics of gender reveal three categories: male, female, and undisclosed. Specifically, 29,527 users are male (28.3%), 13,843 are female (13.3%), and 60,980 have undisclosed gender (58.4%). This distribution indicates that a majority of users either have not specified their gender or prefer to keep it confidential, accounting for over half of the total user base.

3.1.3 Member Logo Distribution

This study analyzes user data in the knowledge-based video category, noting that '0' represents ordinary users and '1' indicates member users. The analysis revealed that 67% of users are ordinary, not subscribing to paid services, hence lacking access to premium privileges. Conversely, 33% have subscribed to B station's premium services, enabling them to view additional videos and enjoy enhanced privileges.

3.2 The Behavioural Dimension

3.2.1 Commentary Sentiment Analysis

Sentiment analysis of the 133,515 comments below the collected sample videos was performed in Python using the third-party library SnowNLP natural language processing model to enrich the emotional attributes of the users.

1)Sentiment analysis of the comment text in the two videos under the secondary partition of science science, these two results are more obviously different. The video numbered S1 has 51.8% of positive comments, 31.8% of negative comments, and 16.3% of neutral comments, which clearly shows that most of the users commenting on this video still have positive sentiment attributes, while the video numbered S2 is just the opposite, with 34.6% of positive comments, 53.0% of negative comments, and 12.4% of neutral comments. The video number S2 is just the opposite, with 34.6% of positive comments, 53.0% of negative comments and 12.4% of neutral comments.

2)In the finance and business sub-section, two videos, F1 and F2, were analyzed for their sentiment distribution in user comments. Video F1 displayed a predominantly negative sentiment, with 74% of the comments being negative, followed by 17.5% positive and 8.6% neutral. This overwhelming negativity suggests that users, whether as consumers or merchants, resonate emotionally with the video, generally viewing the Pinduoduo platform unfavorably and sharing their individual purchasing experiences. In contrast, the sentiment for video F2 is more balanced, with a majority of 55.2% positive comments, 30.6% negative comments, and 14.2% neutral comments, indicating a different emotional response from viewers.

3)In the videos under the campus learning sub-district, both of which are selected from about teaching courses, it is found that the videos with video number C1 and video number C2 have a larger percentage of positive comments, specifically analysing the sentiment value derived from each comment, there are more comments in this category of videos about appreciation and gratitude to the teacher as well as the return of wishes after learning, and the comments of both of these two kinds of comments

are positive comments, so they account for a larger percentage of the comments, of which in C1 the percentage of positive comments accounted for 59%, negative comments accounted for 22.8%, and neutral comments accounted for 18.2%; in C2 the positive comments accounted for more, with a percentage of 68.5%, negative comments accounted for 16.6%, and neutral comments accounted for 15%.

4)In the videos under the Career and Workplace sub-division, the two sentiment percentages basically converge. Under video number W1, 54.3 per cent of comments were positive, 31.2 per cent negative and 14.6 per cent neutral, while under video number W2, 57.5 per cent of comments were positive, 28.1 per cent negative and 14.5 per cent neutral.

3.2.2 Comment Text Similarity Analysis

The need for a vectorised representation of the text based on BERT sentence embedding was introduced above, and the similarity analysis between the comment texts was done individually for each of the video captured comment content and saved in the form of a matrix. In the n*n dimensional matrix, the degree of similarity between each comment and other texts posted under that video can be derived, with similarity score values ranging from (0,1). cosine_similarity(embeddings) returns a matrix where each element (i,j) contains the cosine similarity between sentence i and sentence j. The cosine similarity is calculated as a function of the cosine similarity of each comment. If the value is closer to 1, it means that the comments between two and two are more similar, and if it is closer to 0, it means that the comments between two and two are not similar[6]. In the matrix it can be seen that the matrix ai (i=1,2,.... .n) is the similarity between the ith comment and the ith comment, so it is basically around 1, e.g. a11=0.9999, due to the matrix dimension is too big, the middle part are used... shows that of the similarity values that can be seen, basically the values are all above 0.5. Convert the similarity into distance, the specific practice is to use 1 to subtract the resulting cosine value, the resulting value to determine the distance, set the K value = 5, the use of K-Means clustering method will be close to its distance from the comment text gathered together, each video is divided into five different class sign. Open the file after the completion of clustering with the EXCEL form of the screening function, one by one to see the classification of each video, found in the class between the class sign and the class sign of the text has a similar text, but belongs to a different class, for the phenomenon we merge them, in order to be able to have a better extraction of the subject matter after the text of the clustering effect, we use the LDA theme model to extract the subject matter, which will be divided into two classes. This is in order to do the next step of video information data and the text after clustering better similarity analysis, to come up with the users who fit and do not fit the text of the video release comments.

4 Results

In this study, based on the characteristics of the analysed content of the text: text similarity and LDA thematic analysis, and the results obtained by clustering with the help

of K-Means, we classify its users into the following three types. Based on the combination of filtering and sorting functions in EXCEL, we enrich the characteristics of the three types of users more deeply, and obtain the group portrait and content tendency of different types of users: Group 1 is a deep explorer of the content, willing to carry out scientific knowledge and in-depth discussion of the video; As shown in Fig. 2. .Group 2 is an active atmosphere-creation user, good at creating interesting comments that are not too related to the content of the video; As shown in Fig. 3. and an active Group 3 are experience-sharing users, mostly respondents in the comments, who post complicated contents and are generalised users; As shown in Fig. 4. The three groups of users have different characteristics.

Fig. 2. Content Exploration User Profile

Fig. 3. Atmospheric User Profile

Fig. 4. Experience Sharing User Profile

5 Conclusions

Although this paper has certain contributions, there are still some research limitations that need to be further improved: when selecting samples, we only chose two videos under each partition, and we chose videos with more comments, and videos with fewer comments as well as other videos in each partition were not included in the scope of this research, which lacks universality. When obtaining the information of B station users, we did not obtain their data for analysis because part of the data involves the anti-crawler mechanism, and the angle of analysis is limited to the research perspective of this paper. In the future, we need to continue to read other literature to make corrections and construct a portrait with richer and more comprehensive features.

Reference

1. Liao, Mingzhu & Wang, Chaoqun. (2023). The Intrinsic Mechanisms and Future Pathways for Sustainable Development of Knowledge-Based Videos on B Station from a Co-Creation of Value Perspective. Audio-Visual World, (06), 27-31.
2. Yao, Sumei. (2021). Research on User Portrait Models for Online Video Community Groups Based on Bullet Comments (Master's thesis, Yanshan University).

3. Bao, T., & Gao, J. (2019, September). Research on American contemporary art review based on sentiment analysis technology. In Proceedings of the 2019 International Symposium on Signal Processing Systems (pp. 136-141).
4. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
5. Chen, Tianyuan. (2018). Empirical Construction of User Portraits for University Mobile Libraries. Library and Information Service, (07), 38-46.
6. Pradhan, N., Gyanchandani, M., & Wadhvani, R. (2015). A Review on Text Similarity Technique used in IR and its Application. International Journal of Computer Applications, 120(9), 29-34.