A Validation of the Algorithmic Formula for the Relevance of Douyin Users' Favorite tags

Weijie Cao[1, †], Jin Li[2, †], Xinran Meng[3, *, †]

[1]School of Finance and information, Ningbo University of Finance and Economics, Ningbo, China

[2]Business School, University of Shanghai for Science and Technology, Shanghai, China

[3]Faculty of Business, University of New Brunswick, Saint John, Canada

*Corresponding author: xmeng1@unb.ca

†These authors contributed equally.

Abstract. By June 2022, the number of monthly active users of Douyin has exceeded 800million. Since its launch, it has been favored by many users because of its tag push mechanism. The tag push mechanism has become one of the core characteristics. This paper investigates the accuracy and feasibility of a recommendation formula for the Douyin tags' relevance algorithm. A questionnaire survey is carried out and then empirical analysis in terms of the collected data is carried out to find out the relevant coefficients between different tags and compare it with the actual data collected by Douyin. Based on the results of the questionnaire, the food and life tags had the highest relevant coefficients. Subsequently, the food tag is selected as the main object of study. After further processing and analysis, the relevant coefficients between food and life tags was even higher, confirming the feasibility of the tag relevance recommendation algorithm. These results shed light on guiding further exploration of Douyin algorithm.

Keywords: Neural Network, Prediction Model, Big Data.

1. Introduction

Contemporarily, human beings are in an era of information explosion, where short video applications are rapidly popular, occupying a large number of markets. According to the data of China Internet Information Center, China's short video users have reached 934 million, and the utilization rate is more than 90.2 % in December 2021 [1]. In order to provide users with more accurate product services and improve the real flow of products, short video enterprises use big data technology in the recommendation mechanism of the platform to provide personalized recommendation for users. The "Tik Tok" APP is a product under the headline series today. Today's headlines use a powerful algorithm technology to label user-published content as a tag. After the user opens the vibrating APP, the home page will pop up the video by sliding, which is easy to operate. The vibrating APP can accurately analyze the user's preference according to the user's preference video type and the dwell time of a certain video, and then give the user a short video content of the tag type that the user likes, thereby increasing the user frequency [2]. Recommendation algorithm is a technology to accurately match information content with users by means of big data analysis and information filtering mechanism. It has been widely used in information, e-commerce, search engine, short video and other Internet platforms. The content recommendation of algorithm accounts for 70 % of the whole Internet content distribution [3]. In order to improve the single stay time and stickiness of user, the recommendation mechanism analyzes the user's historical behavior data, labels each user, depicts the user portrait, and combines the brand product category for personalized recommendation. Personalized recommendation is to reclassify the fragmented information precipitated by the platform, filter and reorganize the directional output. When the algorithm enters the field of information dissemination, the decision-making power of information delivery gradually shifts from people to machines. The recommendation mechanism provides personalized recommendation for customers according to the logic of "who are you" deciding "what to give you" [4]. In short, Douyin controls your menu of entertainment by observing your reactions to each past video. Therefore, Douyin users don't need to think and search for the videos but are fed personal preference-based videos, which is

a crucial part of Douyin [5]. The current mainstream recommendation algorithms are roughly the following categories: content-based recommendation, collaborative filtering algorithm and hybrid recommendation [6]. Douyin currently uses a hybrid recommendation algorithm based on content and collaborative filtering recommendation, and develops a special patent algorithm to calculate the relevance of user favorite tags and recommend content. The goal of this paper is to explore the relevant coefficients between user tags of Douyin based on the tag relevant coefficients formulae loved by users of Douyin platform. The rationality of the formula is verified by collecting the data of users. The main theoretical basis of this study is the current algorithm formula of Douyin's relevance to users' labels. By using the questionnaire and analyzing the data, the relevance between labels is obtained, and the rationality of the results is studied by collecting user data. Douyin invented a patent algorithm to analyze and predict user behavior. The simplified formula is:

$$Rel(t1, t2) = \frac{2N(t1, t2)}{N(t1) + N(t2)} \tag{1}$$

where the relevant coefficients coefficient of t1 and t2 is equal to twice the number of t1 and t2 tags appearing at the same time, except the sum of the number of t1 tags appearing alone and the number of t2 tags appearing alone. Accordingly, if the greater the relevant coefficients coefficient, the greater the relevance before the two labels, system will recommend more t2 label content to users who like t1 labels; if the relevant coefficients coefficient is smaller, the number of users who like two tag videos will be less, and the system will reduce the corresponding recommendation. This is also an important theoretical basis for this study.

Based on the elaboration and analysis of cases, this paper finds out the problems that need to be solved, collects the original data by questionnaire survey, and uses database management system to find out the data that is essential for the research. Then, using the hypothesis argument method, assuming that the algorithm formula is established, and put the existing data results in. Finally, the actual users' favorite video labels are collected and verified the rationality of the algorithm formula by frequency statistics. The rest part of the paper is organized as follows. The Sec. 2 will summarize the previous literature. Afterwards, the Sec. 3 will discuss the methodology for this paper. Subsequently, the empirical analysis results will be demonstrated in Sec. 4. Eventually, a brief summary is given in Sec. 5.

2. Literature Review

User preference is a subjective biased choice when users choose goods or services. Preference is a basic concept of modern consumer behavior theory. With the development of big data and machine learning technology, scholars have begun to cross-study user preferences with big data and machine learning technology. The recommendation algorithm is an algorithm in computer science. Through some mathematical algorithms, one can infer the preference of users. The better application of the recommendation algorithm is mainly the network. The so-called recommendation algorithm speculates the users preference based on some of the user's behavior via mathematical algorithms.

Ratsch et al. proposed a personalized clothing recommendation system based on user sentiment analysis by using emotion calculation and intelligent human-machine interface technology. Through the hybrid recursive convolution neural network, the accuracy of personalized clothing recommendation was greatly improved [7]. Ge believes that the recommendation algorithm of Douyin is based on the user ' s browsing history and basic personal information to calibrate the user, and then push the content to the user through a filtering algorithm. It has resulted in the homogenization of audio and video jitters, as creators tend to choose widely acclaimed short video categories when making videos, while more long-tailed high-quality videos are ignored [8]. Zhong believes that the most distinctive feature of Douyin's recommendation algorithm is to seek differences and differences. He believes that the algorithm is a set of machine plus manual dual audit algorithm mode. In addition

to manual audit to pass to the user platform recognized value. AI analyzes the user ' s preferences through the process of input data-machine learning-prediction, finds the video content pool that belongs to the user, and achieves thousands of recommendations [9]. In addition to using a hybrid recommendation algorithm based on content and collaborative filtering, Liu argues that Douyin also uses a flow pool recommendation method, which is similar to A/B Test. When a work is published, the algorithm will distribute the work to some users, and obtain the evaluation of these users for the work, such as point praise, comment number and forwarding number. If the work performs well, the algorithm will expand the scope of recommendation for the work, and then label the work and push it to users with the same label [10]. Huang argues that the recommendation algorithm of Douyin is a recommendation algorithm based on user preferences and comments. Douyin will make relevant coefficients recommendations based on videos that users have browsed, searched or interested before, and label customers in the form of "definition tree" [11]. Zhao stated that the main reason for the success of Douyin in the market is its unique patent algorithm. Buffet will label each user to facilitate its recommendation algorithm. First, the content of user browsing will be divided into videos in different fields. Based on the amount of customer browsing and feedback (including praise, collection, comment, and forwarding), the technology of big data is used to copy and add thousands of data [12]. Therefore, a complete database is composed of hundreds of millions of video labels that customers like.

Overall, Although the recommendation algorithm of Douyin is very complex, the basic logic of the algorithm is still content-based algorithm recommendation. It will learn the user' s viewing preference according to the browsing history of the user watching the video, and use the algorithm to label the user and classify the users. On the basis of this label, Douyin uses collaborative filtering algorithms to divide users and push similar content for users with similar interests. At the same time, Buffet will use the method similar to A/B Test to test the content of the push, and select a better video for large-scale push. However, the recommendation mechanism of Douyin may also lead to the homogenization of platform content, resulting in more high-quality long tail content being buried.

3. Methodology

3.1 Tag Relational Algorithm

According to the Analysis on the "Douyin (Douyin) Mania" Phenomenon Based on Recommendation Algorithms, Douyin tags users based on the type of video they prefer and places A subset of users are grouped under the same tag. For example, if user A likes to watch car-oriented videos, Douyin will categorize all the videos he or she views by different tags to make further algorithmic recommendations. In addition to car-oriented videos, user A also likes sports-oriented videos (NBA, football, skiing, etc.). The tag relevance algorithm of Douyin will record this phenomenon after analyzing thousands of "User A's" and obtaining a few basic relevant coefficients tags, and using the Eq. (1).

To find the relevant coefficients coefficient between the individual tags. In this formula, Rel (t1, t2) represents the relevant coefficients coefficient between tags t1 and t2, N (t1, t2) represents the number of times tags t1 and t2 appear together, N(t1) represents the number of times t1 appears alone (i.e., users are not interested in tag t2), and N(t2) represents the number of times tag t2 appears alone. This formula can derive the relevant coefficients coefficient between users' preferences for different tags, which helps Douyin's algorithm to recommend videos to users more accurately. For example, the relevant coefficients coefficient between cars and sports is high, so when a new user likes cars, the Douyin system will often try to recommend sports videos in order to capture the new user's interest.

3.2 The experimental design

Based on Douyin's tag relevance recommendation algorithm, we decided to verify the accuracy of this algorithm for practical applications. To be specific, a two-part experiment is designed to validate the algorithm. Primarily, a questionnaire was used to collect the types of videos that users

like to watch. The questionnaire provides 14 options for the question "What types of videos do you like to watch", the tags are: Food, Life, Cosmetics, Games, ACGN (Animation, Comic, Game, Novel), Dressing, Funny, Music, Facial Attraction, Film & TV, Emotion, Sports and Pets. Once the data was collected, the 14 tags were correlated with each other and visualized according to the algorithm Eq. (1). Subsequently, in terms of the obtained relevant coefficients, we selected the highest scoring parts for secondary collection of the actual Douyin data. In this paper, Food and Life had the highest relevant coefficients coefficient, so we selected the food tag, and observed and collected the 10 most recent videos liked by the food bloggers among their followers, recorded and counted them. (Note: If the video is about food, it will be postponed by one video.) Finally, we used frequency statistics to find out what types of videos, other than food, were most popular among users who liked food videos, and compared them with the obtained relevant coefficients.

3.3 The Data Collection of questionnaire

According to the research needs, a questionnaire on users' interest preference for portraits is designed. By setting the single and multi-choice questions to understand the user's use of Douyin behavior and Douyin content preferences. The first three questions can understand the basic portraits of Douyin users. The fourth question is the validity test of the questionnaire. The fifth and sixth questions are to understand the use behavior of the users. The seventh and ninth questions are to understand the use and content preferences of the users. The data type recovered in the seventh question is text data, which is used to verify the formula of tag relevance of the users. On account of the epidemic, the questionnaire was distributed through the Internet, a total of 268 copies were recovered, including 268 valid questionnaires.

4. Results

The design of the questionnaire and the collection of data allowed us to calculate the specific values of 91 relevant coefficients for the combination of 14 labels, as shown in the Table. 1. As Table.1 shows, Food has the highest relevant coefficients coefficient with Life tag at 2.62, while ACGN has the lowest relevant coefficients coefficient with Dressing and Cars, both at 0.35. The visualization of above results is shown in Fig.1.

Table.1. relevant coefficients.

Rel	Food	Live	Cosmetics	Game	ACGN	Dressing	Funny	Music	Facial attraction
Food	1.00								
Live	2.62	1.00							
Cosmetics	1.03	0.68	1.00						
Game	1.07	0.99	0.68	1.00					
ACGN	0.36	0.36	0.53	0.48	1.00				
Dressing	1.43	1.30	1.38	0.63	0.35	1.00			
Funny	2.14	1.96	0.84	1.20	0.36	0.83	1.00		
Music	1.51	1.35	0.95	1.01	0.50	1.22	1.39	1.00	
Facial attarction	0.81	0.75	1.03	0.88	0.49	1.21	0.96	0.92	1.00
Film&TV	1.88	1.85	0.79	1.22	0.48	1.16	1.87	2.05	0.79
Emotion	0.93	0.87	0.74	0.64	0.51	1.14	1.19	1.01	0.78
Sport	1.11	1.15	0.55	0.95	0.36	0.90	1.11	1.36	0.80
Pets	0.72	0.80	0.89	0.77	0.38	0.99	0.84	0.92	0.99
Car	0.60	0.66	0.37	0.88	0.35	0.46	0.72	0.77	0.52

To facilitate the subsequent part of the study, Food labels were intercepted as a comparison for the subsequent part of the validation. Based on the results of the data analysis in the previous section, Food tag was selected as a representative for the study. We collected the last 10 likes of videos from users who liked food videos (excluding short videos of food and social hotspot), a total of 6 bloggers and 120 samples of user likes, as illustrated in the Fig. 2 and Fig. 3.

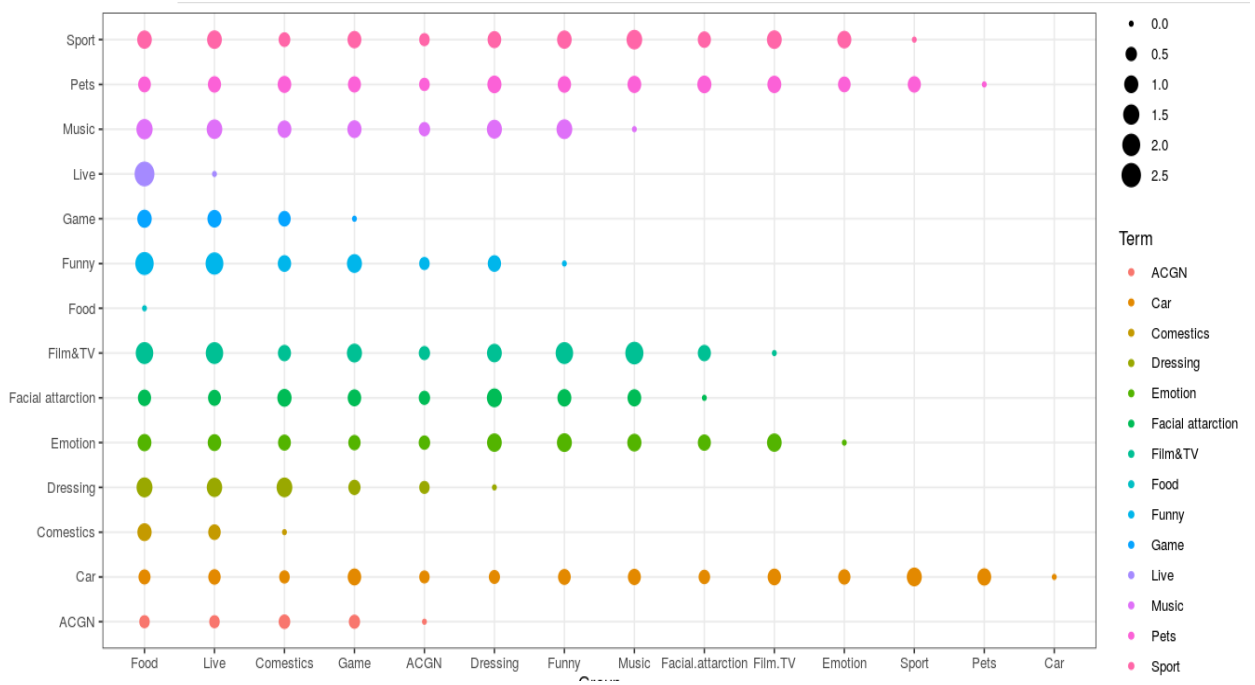

Figure 1. Relevant coefficients' heatmap.

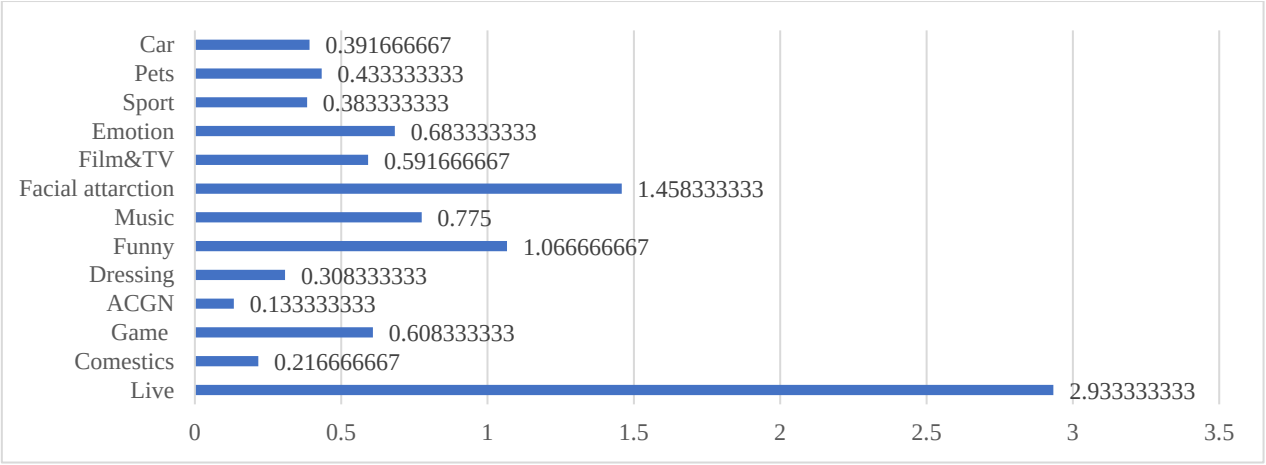

Figure 2. Frequency of Different Types of Videos.

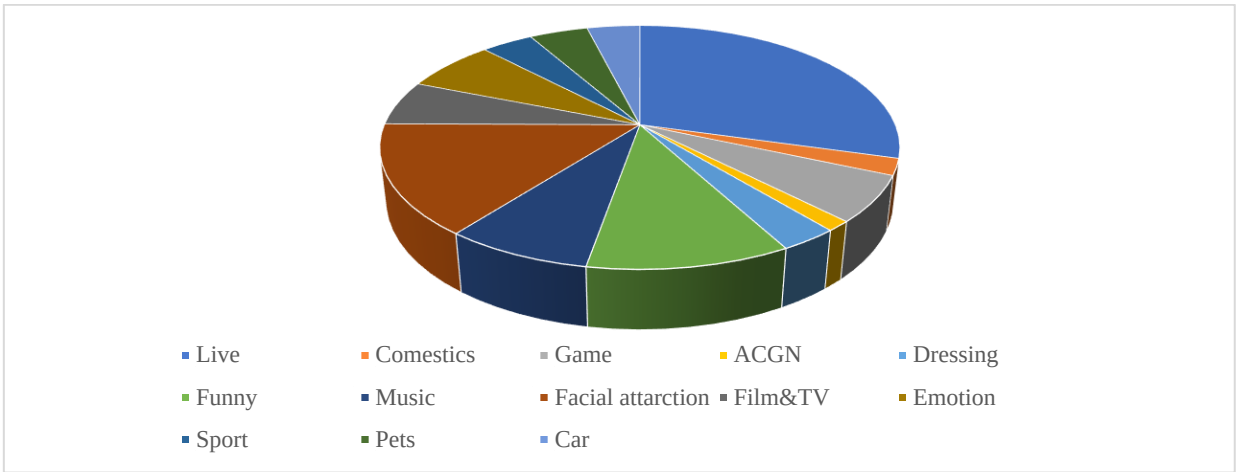

Figure. 3. Pie Chart of Different Types of Videos.

The above data is expressed as follows: among those who like the food tag, the average number of likes per 10 recent likes is 2.933 for life videos, almost 3; the average number of likes for facial

attraction videos is 1.458; and the average number of likes for funny videos is 1.067. According to the above analysis of the relevant coefficients, the relevant coefficients coefficient between the food and life categories is 2.62 and has the highest percentage. From the real collected data, the life tag does appear more frequently together with the food tag, i.e., those two tags have the highest relevant coefficients coefficient. Whereas funny videos, as calculated by the algorithm, should be the second most frequent type of tag, the facial attraction tag appears more often than the funny videos. This indicates that for some of the tags, a larger sample size is needed to correct the data; and that for specific tags, there are more additional reaction that affect their frequency of occurrence.

5. Conclusions

In summary, this paper investigates the Douyin recommendation algorithm based on the formula mentioned in Analysis on the "Douyin (Douyin) Mania" Phenomenon Based on Recommendation Algorithms. According to the data of questionnaire, the relevant coefficients between food and life was the highest. In addition, according to the videos of some food blog fans, we observed their preferences for the last 10 times to analyze their preferences. In addition to the higher frequency of food and life labels, appearance and funny labels also appeared more frequently. Moreover, users with food labels can also recommend some videos related to life, appearance and funny when Douyin recommends them, so as to improve the viscosity of users. In the future, Douyin should try to control advertising and innovative video content in order to make more profits in its development.

Nevertheless, the results and statements of this paper do have some shortcoming and drawbacks. Primarily, this paper analyzes the video types of likes manually when analyzing the relevant coefficients types. Although this method is relatively simple and reliable, it is subjective to define the video types, which inevitably leads to some deviations. Secondly, only 10 videos were sampled during the analysis. However, since some users like the similar videos at the same time, and there are few samples, it may not reflect the actual user interests, which may affect the judgment and cause errors. In the future, Douyin should try to control advertising and innovative video content in order to make more profits in its development.

The important marketing strategy of Douyin is that people of almost all ages can use it. Therefore, Douyin video should cover all ages and allow users to participate. In order to enable larger numbers of users, most of the initial recommendations will have their interests with you. Whereas, as more and more people use the application for a long time, Douyin also needs to not only rely too much on the recommendation algorithm, but to recommend more videos with large relevant coefficients coefficient of interest. In this case, it can prevent users from aesthetic fatigue and improve users' viscosity. Overall, these results offer a guideline for Douyin estimation mechanism.

References

[1] CNNIC. 49th China Internet Development Statistics Report[R]. 2021

[2] L. Wei, "Development status and countermeasures of short video APP," Jiangxi Normal University, 2017.

[3] Peng Xunwen. What algorithms do we need in the mobile Internet era [J]. China Newspaper Industry, 2021(05): 46-47.

[4] Wen Fengming, Xie Xuefang. Operation Logic and Ethical Concerns of Short Video Recommendation Algorithm - Based on the Perspective of Action Network Theory [J]. Journal of Southwest University for Nationalities (Humanities and Social Sciences), 2022, 43(02):160-169.

[5] Yulun Ma, Yue hu.Business Model Innovation and Experimentation in Transforming Economies: Byte Dance and TikTok [J]. Management and Organization Review .2021,17(2): 382–388

[6] Yu Wei, Xu Dehua. Overview and Prospect of Recommendation Algorithm [J]. Technology and innovation, 2019(04): 50-52.

[7] Matthias Ratsch et al. Personalized Clothing Recommendation Based on User Emotional Analysis [J]. Discrete Dynamics in Nature and Society, Volume 2020, pp. 1-8.

[8] Ge Qingkun. Research on Short Video Recommendation Mechanism [D]. University of Dalian for Science and Technology, 2020.

[9] Zhong Cheng. Study on Diffusion Law of Buffet Short Video in Scene Vision [D]. South-central University for Nationalities, 2019.

[10] Liu Mengxuan. Survey on the influence of algorithm recommendation on college students' use of buffeting sounds [D]. Nanjing University,2021.

[11] Huang Baiyu. The Reasons for Douyin's Success from the Perspective of Business Model, Algorithm and Functions [J]. Advances in Economics, Business and Management Research, 2021, volume 166.

[12] Z. Zhao. Analysis on the "Douyin (Tiktok) Mania" Phenomenon Based on Recommendation Algorithms [J] E3S Web of Conferences 235, 03029, 2021.