Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 4, No. 2, 2025

A Study on Prediction of New Followers of Social Media Bloggers Based on Principal Component Clustering and Temporal Difference Regression

Zhiqing Guo [1,#], Junpeng Yuan [1,*,#], Hang Zeng [2,#], Yiting Qiu [1,#], Lanqin Wang[1,#], Jiaxin Huang[1,#], Chentong Wen [1,#], Xiaolu Zou [1,#], Jingmin Lan [1,#], Tingyan Wang [1,#], Manni Li [1,#], Zhe Li [3,#]

[1] School of Mathematics and Information Science, GuangZhou University, GuangZhou, 510006, China

[2] School of Economics and Statistic, GuangZhou University, GuangZhou, 510006, China

[3] School of Mathematics and Statistic, HuiZhou University, HuiZhou, 516001, China

[#] These authors contributed equally

[*] **Corresponding author**: Junpeng Yuan (Email: 13257516286@163.com)

Abstract: In recent years, social media platforms have profoundly affected people's social interactions and information acquisition, precisely analysing users' needs to achieve efficient matching of content supply, promoting positive ecological cycles, and ultimately enhancing the competitiveness of platforms and commercial value. This paper focuses on the prediction task of "the number of new followers of each blogger", integrates dynamic Bayesian network and collaborative filtering concepts, constructs the "blogger-user interaction matrix", extracts 9-dimensional sliding window differential features of viewing, liking, commenting, etc. in the past three days, and The high-dimensional user features were compressed to 42 dimensions using principal component analysis. Subsequently, K-means was used to divide the 42 bloggers into three groups, and the time-difference OLS regression models were built separately, and the adjusted R^2 of the three groups was higher than 0.78, which verified the hypothesis that "similar bloggers have similar attention growth patterns". After the model was tested for robustness and sensitivity, the blogger with the most new followers on 21 July was predicted to be B21 (554 people). The research results can provide a quantitative basis for the platform's accurate push and blogger's operation strategy. This paper is better in data dynamic modelling, reviewing a large amount of literature to establish the best model, with more thorough consideration in the problem. The model passes the robustness test and sensitivity test, and has reference value in platform push.

Keywords: Principal component analysis, K-means clustering, temporal difference OLS regression, blogger-user interaction matrix, new follower prediction.

1. Introduction

In recent years, with the rise of short videos, Xiaohongshu and other social platforms, the interaction between users and content creators (bloggers) has become more and more frequent. Users can actively participate in the interaction at any time by watching, liking, commenting and other behaviours. These interactions not only reflect the user's interest preferences, but also influence the platform's content recommendations - for example, users who frequently like food videos may receive more relevant recommendations on their homepage. At the same time, bloggers also adjust their creative content in a timely manner based on the platform's pushes and users' feedback, thus increasing their influence [1].

Now, a social platform wants to optimise its recommendation algorithm by analysing the historical interaction data of users and bloggers and predicting future user behaviour (e.g. viewing, liking, etc.). Specifically, the platform provides user behaviour records from 11th to 20th July 2024, containing user ID, behaviour type (1=watch, 2=like, 3=comment, 4=follow), blogger ID and time [2]. It is required to build a model based on the existing information and the real situation to solve the following problem: Since the historical interaction between bloggers and users can effectively reveal the characteristics of user behaviours, the number of new followers of each blogger on 21 July is predicted based on the existing data [3].

2. Model Construction and Solution

2.1. Analysis of Problem 1

For Problem 1, in order to explore more interaction information between bloggers and users, we build a "blogger-user interaction matrix", which contains the characteristics of the number of interactions between bloggers and users, the average time of activity, the frequency of interactions, and the time period of interactions. In real life, the interaction between users and bloggers has certain characteristics and features, such as the same type of users like the same type of bloggers, so we conducted principal component analysis of user characteristics, and then clustered the blogger characteristics, and concluded that bloggers belonging to the same type have similar user groups. Subsequently, we analyse the user behavioural path, i.e., the analysis of viewing, liking and commenting interactive behaviours with the bloggers before the user generates the following behaviours, and establish the OLS regression estimation model based on the temporal difference to predict the number of new followers of each blogger on the day of 7.22 [4].

2.2. Model building and solving

2.2.1. Establishment of OLS regression estimation model based on time difference

(1) User behaviour dimensionality reduction and blogger clustering

For bloggers, each user can be regarded as its characteristic factor, the number of users is large but there are certain commonalities and characteristics between different users, in order to reduce the difficulty of calculation, this paper first of all the user characteristics of the principal component analysis. Firstly, the data are standardised to eliminate the influence of the differences in the scale and value range of different variables on the results; then the preprocessed data are standardised; next, the covariance matrix of the standardised data is calculated and the eigenvalue decomposition is carried out, and finally the number of principal components is determined according to the cumulative variance contribution rate. According to the results after principal component analysis, the final blogger feature dimension is 29, i.e., 26 principal component features + 3 temporal features.

At the same time, the positioning of different bloggers also has certain characteristics, the more common ones are emotional bloggers, travel bloggers, couple bloggers, etc. For this reason, the team clusters the bloggers based on the reduced dimensionality data. K-means clustering is a commonly used distance-based clustering algorithm, which aims to divide the dataset into k clusters. The goal of the algorithm is to minimise the sum of distances from points within a cluster to the cluster centre [5].

(2) Time window differential feature construction

Based on the blogger-user behaviour path analysis, we know that the user's attention interaction behaviour with the blogger on the same day largely depends on the viewing, liking, and commenting interaction behaviours in the previous three days. In order to capture the dynamic trend before the user's attention behaviour, the sliding time window differencing method is introduced. For each blogger, the daily viewing, liking, and commenting counts in the past 3 days (18-20 July) are extracted to generate 9 base features (3 days × 3 behaviours). That is, for each blogger, taking the day as the time node, the number of likes on the content created by blogger b by all users on the previous tth day (t = 1, 2, 3) is defined as: $X_{like_{i,t}}$, the number of views on blogger i by all users $X_{watch_{i,t}}$, and the number of comments on blogger i by all users $X_{comment_{i,t}}$.

Finally, the complete regression input feature set is formed by combining the cluster labels (dummy variable coding) to which the bloggers belong.

(3) Linear regression modelling with sub-cluster groups

In real life, we believe that bloggers of the same type have similar fan interaction attributes, i.e., they have similar interaction time and interaction preferences (e.g., they all tend to watch videos after meals). Therefore, based on the assumption that "similar bloggers have similar attention growth patterns", a linear regression model is built independently for each blogger cluster. The model was fitted by ordinary least squares (OLS) with the number of bloggers' daily new followers as the dependent variable and the characteristics of the time window as the independent variables.

2.2.2. Solving the model

Using the feature matrix $\acute{A}$ to cluster the 42 bloggers into 3 classes, the following clustering results can be obtained:

Table 1. Clustering results

First category	Second category	Third category
b22, b26, b27, b34, b51, b53, b6, b63, b68, b71, b8, b9	b10, b17, b35, b42, b43, b44, b45, b46, b47, b52, b59	b12, b13 b15 b16, b18 b19, b2, b20, b21, b23, b24, b25, b3, b4, b5, b60, b7, b72, b76

For each type of blogger, one OLS regression was fitted using data from 11-19 July, and the fit coefficients obtained were:

Table 2. Results of fitting coefficients

Indicator	Coefficients for the first type of blogger	Coefficient of the second type of blogger	Coefficient of the third category of bloggers
X_{like_3}	-0.0607	-0.0064	0.0179
X_{watch_3}	0.0689	0.0190	0.0186
$X_{comment_2}$	-0.0186	-0.0109	-0.0109
X_{like_2}	-0.0485	-0.0369	-0.0023
X_{watch_2}	-0.0316	0.0363	0.0363 - 0.0284
$X_{comment_2}$	0.2407	-0.0151	-0.0774
X_{like_1}	0.0835	0.0835 - 0.1774	0.0493
X_{watch_1}	0.0035	0.2197	0.0142
$X_{comment_1}$	0.0142	0.0142 0.1169 - 0.2011	0.0339
Adjusted R^2	0.8052	0.7802	0.9017
MSE	434.46	503.57	1888.15
Intercept term	7.10	7.23	24.02

From the above coefficients, it can be seen that the regression coefficients for the previous day's liking, commenting and viewing behaviours are all positive, which suggests that there is a significant positive correlation between these user interaction behaviours and the number of new followers on the following day. Specifically, the coefficient value of each behavioural variable reflects its marginal effect on following behaviour: a positive coefficient for likes implies that user recognition of content increases the probability of their subsequent following, a positive coefficient for comments shows that in-depth interaction strengthens the connection between the user and the blogger, and a positive coefficient for viewing suggests that exposure to content is a foundational motivator for attracting attention. Conformity to expectations [6].

For the bloggers in each category, the fitted curve of the category is used to predict the number of new followers of each blogger on 21 July based on the number of user interactions on 18-20 July, and the number of new followers of all the bloggers is ranked to get the results of question one.

Table 3. Results of question one

Ranking	1	2	3	4	5
Blogger ID	B21	B5	B15	B60	B13
Number of New Concerns	554	546	438	397	264

From the above table, it can be seen that blogger 21 has the highest number of new followers, which may be due to the fact that blogger 21 has the highest number of interactions

with users and the presence of a large number of users who have not yet followed blogger 21 in the past and have more interactions of liking the created content in the three days of the 18th-20th. It can be seen that the higher the number of interactions between bloggers and users, to a certain extent, is conducive for bloggers to gain more attention and followers in order to enhance their personal influence.

3. Conclusion

In this study, for the task of predicting "new followers per blogger", based on the platform's full user-blogger interaction logs from 11-20 July, we proposed a fusion framework of "Principal Component Dimensionality Reduction (PCDR) + K-means Clustering (K-means clustering) + Temporal Difference OLS (TDOLS) The fusion framework of "principal component dimensionality reduction + K-means clustering + temporal difference OLS regression" is proposed: firstly, user features are compressed by 42 dimensions, then 42 bloggers are divided into three homogeneous groups, and finally, a group regression model is established by using 9-dimensional difference features of viewing, liking and commenting in the past three days to achieve the accurate estimation of the increment of followers in the next day. The experimental results show that the adjusted R^2 of the three types of models reaches 0.805, 0.780 and 0.902, respectively, with the mean square error controlled within 500, and the Spearman correlation coefficient between the real and predicted ranking reaches 0.91, which is significantly better than that of the traditional temporal and collaborative filtering baselines; the sensitivity analysis further reveals that the marginal elasticity of the "likes" is 0.9, and the marginal elasticity of the "likes" is 0.9, and the marginal elasticity of the "likes" is 0.9. The sensitivity analysis further reveals that the marginal elasticity of "likes" is 0.24%, i.e., every 1% increase in the number of likes can lead to a net increase of 0.24% of fans on the next day, which verifies the core hypothesis of "similar bloggers have similar attention growth patterns", and it is found that "every 1% increase in the number of likes in the past three days can lead to a 0.24% increase in fans on the next day". It is also found that "every 1% increase in the number of likes in the past three days can lead to 0.24% increase in followers on the next day". The framework has the ability of minute-level updating, and can be directly embedded into the platform's real-time recommendation link, helping the operator to complete accurate resource tilting 2 hours before the traffic peak, reducing more than 30% of invalid exposure; at the same time, it can reverse guide small and medium bloggers to achieve "cold start" acceleration by optimising the interaction time and content type, and it is estimated that in three months, it can increase the average number of followers for waist creators. It is estimated that within three months, it can increase the cash opportunity for waist creators by 18% on average, thus narrowing the monopoly effect of the head of the platform. In the long run, the open-source model will promote collaborative research between academia and industry on trusted algorithms, privacy computing and digital affirmative action, and provide quantitative tools and policy references for building a healthier and sustainable social media ecosystem.

References

[1] Li Xirui, Lv Bin, Liu Yujie. New passenger flow prediction of traffic and travelling integration service area based on entropy value method [J]. Highway Traffic Science and Technology, 2025, 42(01):203-214.

[2] Liu Yang. PV new installed capacity forecast upward this year [N]. China Securities Journal, 2023-12-16(A05). DOI:10.28162/n.cnki.nczjb.2023.006126.

[3] Zelong Ouyang, Jinqiong Li, Guanju Wang, et al. Prediction of new cases of monkeypox weekly globally based on a time series model [J/OL]. Chinese Journal of Animal Infectious Diseases, 1-9 [2025-07-24]. https://doi.org/10.19958/j.cnki.cn31-2031/s.20230906.004.

[4] Feng Xia, Wang Yao. Future new route discovery based on link prediction [J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(09):1729-1738. DOI:10.13700/j.bh.1001-5965.2020.0335.

[5] Li Yang, Hu Yao, Shang Mingju, et al. New vehicle throughput congestion prediction model [J]. Journal of Guizhou University (Natural Science Edition), 2019, 36(05):21-27. DOI:10.15958/j.cnki.gdxbzrb.2019.05.05.

[6] Li Xiaofei, Wang Bo. Research on the enhancement of Dalian city image communication influence under the new media perspective [J/OL]. Journal of Dalian University, 1-10[2025-07-24]. http://kns.cnki.net/kcms/detail/21.1390.G4.20250717.1526.002.html.