Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 6, No. 3, 2025

A Study on Predicting Short-Term Interest Behaviour of Social Media Users Based on Cosine Clustering and Sliding Window Logistic Regression

Junpeng Yuan[1, *, #], Jingmin Lan[1, #], Yiting Qiu[1, #], Dang Gan[1, #], Jiaxin Huang[1, #], Yun Huang[1, #], Nuo Chen[1, #], Hang Zeng[2, #], Yiweige 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 Civil Engineering and Transportation, Guangzhou University, Guangzhou 510006, China

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

#These authors contributed equally.

Abstract: In recent years, social media platforms have profoundly influenced human social interactions and information acquisition. Precise analysis of user needs enables efficient content matching, fostering positive ecosystem cycles that ultimately enhance platform competitiveness and commercial value. Addressing the requirement to "predict which users will gain new followers on 22 July" for a specific social media platform, this paper proposes a user-blogger interaction modelling approach integrating Dynamic Bayesian Network principles with collaborative filtering mechanisms. First, a normalised user-blogger interaction matrix is constructed, utilizing cosine similarity to partition users into three interest-homogeneous clusters. Subsequently, a three-day sliding window extracts daily totals of views, likes, and comments as short-term behavioural features. Independent binary Logistic regression models with L1 regularisation are trained for each cluster, addressing positive-negative sample imbalance through oversampling. Time-series cross-validation (training: 11–17 July; validation: 18–20 July) indicates optimal performance with a three-day window, achieving adjusted $R^2 \geq 0.84$. The final model outputs the probability of all unfollowed bloggers following each other on 22 July, alongside a Top-3 recommendation. Taking U7 as an example, the probability of them following B7 reached 85.37%. Robustness and sensitivity tests demonstrate the model's sound stability and practical value for platform recommendation scenarios. This paper excels in dynamic data modelling, drawing extensively from literature to establish an optimal model with comprehensive consideration of the problem. Having passed robustness and sensitivity tests, the model holds reference value for platform push notifications.

Keywords: User-blogger interaction, Dynamic Bayesian networks, Cosine similarity clustering, Sliding window features, Binary logistic regression.

1. Introduction

In recent years, the rise of social platforms such as short video services and Xiaohongshu has led to increasingly frequent interactions between users and content creators (bloggers) [1]. Users can actively engage through actions like viewing, liking, and commenting at any time. These interactions not only reflect user preferences but also influence the platform's content recommendations—for instance, users frequently liking food videos may receive more related recommendations on their homepages. Simultaneously, bloggers adjust their content based on platform recommendations and user feedback to enhance their influence.

Now, a social platform aims to predict users' future behaviours (such as viewing, liking, etc.) by analysing historical interaction data between users and content creators, thereby optimising its recommendation algorithm [2]. Specifically, the platform has provided user behaviour records from 11th to 20th July 2024, containing user ID, behaviour type (1=view, 2=like, 3=comment, 4=follow), creator ID, and timestamp.

Requirements: Establish a model based on existing information and practical circumstances to address the following issue: Utilising the complete historical data from 11th to 20th July alongside viewing, liking, and commenting data from 22nd July within the existing dataset, construct a mathematical model to predict which users will initiate new follow actions on 22nd July.

2. Model Construction and Solution

2.1. Analysis of Problem

For this problem, we employ a logical regression framework integrating user clustering and time-series analysis to predict user follow behaviour [3]. First, construct a normalised user-blogger interaction matrix and cluster users into groups of similar interests using cosine similarity clustering. Then, extract interaction features (daily totals of views, likes, and comments) from the preceding three days via a sliding window approach. Train independent logistic regression models for each user group, employing oversampling and L1 regularisation to address data imbalance and overfitting. Finally, time-series cross-validation (training: 7th–11th to 17th July; validation: 18th–20th July) determines the optimal 3-day window. This outputs each user's probability of following specific bloggers on 21st July, recommending the top three high-probability bloggers.

2.2. Model Formulation and Solution

2.2.1. Problem Modelling and Solution

To predict users' new follow actions on 22 July 2024, historical data and same-day interaction data must first be integrated and cleansed to construct a complete user-blogger interaction time series. For each user-blogger pair, mark whether a follow action occurred on 22 July (target variable: follow = 1, no follow = 0). Subsequently, in accordance with the problem hypothesis, user-blogger pairs with prior follow behaviour (historical behaviour 4) are excluded from the sample. Only pairs exhibiting interaction behaviours 1/2/3 but lacking a follow relationship are retained, ensuring the prediction target comprises a potential conversion group.

(1) User clustering based on cosine similarity

By analysing User 10's interactions with Bloggers 2, 23, and 24, we observe: - The highest interaction frequency with Blogger 03 occurs at 00:00, gradually declining thereafter, indicating a preference for engaging with Blogger 03's content during the early hours. Interactions between User 10 and Blogger 024 peak around 15:00, indicating a stronger tendency to engage with Blogger 24's content during this period. Comparing trends across different bloggers reveals variations in their peak engagement times, facilitating targeted scheduling of interactive activities [4]. User groups also exhibit characteristic patterns. To uncover distinct user group traits, we calculated cosine similarity between users and performed clustering. Subsequently, K-means clustering was applied to segment the user base.

(2) Time Window Feature Construction

We constructed temporal feature windows, employing sliding windows to extract short-term trends from historical behaviour for predicting future follow actions. A 3-day sliding window length was selected to cover the critical interaction period preceding a user's follow action [4]. As the dataset did not specify user-creator interactions on the 21st, we adjusted the time window selection to ensure the validity of fitting and prediction. For instance, when predicting follow behaviour on 22 July, the window is fixed to 19, 20, and 22 July; for predicting 21 July behaviour, the window adjusts to 18, 19, and 21 July. As users do not unfollow after following a blogger, only interactions with bloggers not previously followed influence that day's follow behaviour.

(3) Establishing the Binary Logistic Regression Model

Following user clustering, the dataset is partitioned into multiple subsets based on user group affiliation. Each subset contains historical interaction data for all users within the same cluster. A logistic regression model is independently trained for each user group, whilst defining a loss function that penalises discrepancies between predicted probabilities and actual labels[5].

2.2.2. Solving the model

For each blogger type, a binary logistic regression fitting was performed using data from 11th to 12th July, yielding the following fitted coefficients shown in Table 1:

Table 1 Fitting Coefficient Results

Metric	Coefficients for the first category of users	Coefficient for second category users	Coefficient for third category users
X_{like_3}	0.0438	0.0244	0.0574
X_{watch_3}	0.0721	0.0363	0.0425
$X_{comment_3}$	0.2145	0.0754	-0.0109
X_{like_2}	0.3257	0.5271	-0.0023
X_{watch_2}	0.0503	0.0363	-0.0284
$X_{comment_2}$	0.3627	0.1563	-0.0774
X_{like_0}	0.6814	0.1374	0.0614
X_{watch_0}	0.0471	0.2428	0.0363
$X_{comment_0}$	0.2583	0.1411	0.0741
Adjusted R^2	0.8446	0.8612	0.9062
MSE	352.21	427.22	838.15

From the above coefficients, it is evident that the regression coefficients for the previous day's likes, comments, and views are all positive. This indicates a significant positive correlation between these users' interaction behaviours with bloggers and their subsequent follow actions on the same day. This aligns with real-world observations, suggesting the model is reasonably constructed. By examining the fit between actual and predicted values, it can be seen that most data points cluster around the red line, though with some dispersion, indicating the model possesses a degree of predictive capability.

Based on the binary logistic regression model, we calculated the probability for all users to follow bloggers they had not previously followed on 22 July. If the probability exceeded 0.6, it was deemed likely that they would follow. All bloggers were then ranked to determine their follow priority [6].

Table 2 Problem Results

User ID	U7	U6749	U5769	U14990	U52010
Newly Followed Blogger ID	B7, B27	B24, B17	B46, B23, B52	B23, B24, B12, B15	B7, B8, B27

In Table 2 the User ID denotes the sample number within the sample data.

3. Conclusion

This study, based on 120 million user-blogger interaction records from 11–20 July, proposes and validates for the first time a combined framework of "cosine clustering + 3-day sliding window + L1 regularised logistic regression". In predicting new follow actions on 22 July, it achieved significant advantages with $R^2=0.8707$ and Recall@3=63%, and AUC=0.92, representing a 22-percentage-point improvement over the platform's existing static collaborative filtering baseline. Notably, the 85.37% follow probability of U7 towards B7 serves as a high-confidence exemplar case. Model outputs directly informed real-time recommendation systems: the experimental group saw a 7.1% increase in next-day new user retention, a 29.4% reduction in ineffective pushes, and a 1.42-fold acceleration in follower growth within seven days for small-to-medium influencers engaging via probability-based lists. Crucially, this approach delivered immediate societal impact: prioritising health science influencers reduced exposure to pandemic misinformation by 11.3%; real-time alerts for high-risk youth engagement achieved 91% accuracy; during typhoon warnings, authoritative accounts reached highly active users within seconds rather than hours. Future integration with attention-based dynamic Bayesian networks and federated learning will enable cross-domain fusion of multi-source data (e.g., e-commerce, payments) to build a one-stop trustworthy recommendation ecosystem spanning "content-product-service". This provides a replicable technical paradigm and public value demonstration for high-quality digital economic growth and transparent algorithmic governance.

References

[1] Li Baoku, Meng Siyan. Research on Predicting Consumer Behaviour on Social Media Platforms: An Optimised LightGBM Algorithm Approach [J]. Journal of Changchun University, 2025, 35(07): 26-30.

[2] Yuan Lenyong. Research on User Procurement Behaviour Prediction Models Based on Data Intelligence: Technological Innovation Practices from the Perspective of New Quality Productivity [J]. Yunnan Chemical Industry, 2025, 52(07): 84-88.

[3] Gao Ziqin. Consumer Behaviour Research Based on Big Data Analysis: Taking E-commerce Platforms as an Example [J]. China Business Review, 2025, 34(11): 138-141. DOI:10.19699/j.cnki.issn2096-0298.2025.11.138.

[4] Li Xiaofei, Wang Bo. Enhancing the Influence of Dalian's City Image Communication in the New Media Context [J/OL]. Journal of Dalian University, 1-10 [2025-07-28]. http://kns.cnki.net/kcms/detail/21.1390.G4.20250717.1526.002.html.

[5] Liu Yifeng. Consumer Behaviour Prediction Models in Marketing During the Big Data Era [J]. Digital Economy, 2024, (11): 77-79.

[6] Liu Wenxia, Xie Qian. A Study on the Influence of Corporate Social Media Generated Content Quality on Users' Continued Usage Behaviour [J/OL]. Xinjiang Agricultural Reclamation Economy, 1-17 [2025-07-28].