Interest Portrait Construction and Dynamic Evolution Analysis Based on Multi-Source Behavioral Data

Shuaizhe Huang
School of Mathematics and Statistics, Anqing Normal University, Anqing, Anhui, China
**Corresponding Author*

Abstract: In the digital era, diverse and heterogeneous user behavior data are trapped in data silos. Traditional static or single-source interest profiling methods cannot support precise user modeling. This paper focuses on constructing user interest portraits and analyzing their dynamic evolution using multi-source heterogeneous behavioral data. It integrates e-commerce behavioral logs (from the Taobao platform) and movie rating interaction data (from the MovieLens dataset), constructs a unified behavior representation via data cleaning, alignment, and fusion, uses NLP to extract fine-grained interest tags, and introduces a time-series framework with a sliding window and attention mechanism to capture dynamic features and identify persistence, mutation, and periodic patterns of user interests.

Keywords: User Interest Profile; Multi-source Heterogeneous Data; Data Fusion; Dynamic Evolution; Natural Language Processing(NLP)

1. Introduction

1.1 Research Background and Significance

With digital transformation, user behaviors are scattered across platforms, generating massive heterogeneous data. Traditional user modeling uses single-source and static data, which fails to reflect dynamic interest changes and reduces the effectiveness of personalized recommendations and user retention. Multi-source data vary in semantics and structure, and user interests change over time and due to life events. A survey shows approximately 68% of consumers changed brand preferences during the epidemic, indicating the instability of user interests. Static models can't adapt to such dynamics. This study aims to solve temporal dynamics and data heterogeneity in user modeling, support precise profiling, and offer technical support for personalized services.

1.2 Research Objectives

This study aims to build dynamic user portraits using multi-source heterogeneous data to address the fragmentation of single-source analysis. It captures both explicit and implicit user demands for recommendation systems and precision marketing. Second, it applies time-series analysis to model dynamic interest changes and enhance the model's adaptability to environmental and contextual factors. Third, it conducts empirical validation with real-world data to verify the effectiveness of the framework and provide technical guidance for industrial applications.

1.3 Domestic and International Research Status

1.3.1 Abroad Research Status

Recent advances in AI, big data, and interdisciplinary research have spurred interest profiling and multi-source dynamic evolution analysis. The key research lies in multi-source data integration and multi-model fusion for behavioral interpretation. Meanwhile, relevant studies on mechanical fault diagnosis and reliability optimization help eliminate data differences and construct effective user interest portraits.

The development of large language models globally is rapid, and the introduction of RAG technology has solved the causal inference problem. Tengfei Ren et al. noted that RAG technology can analyze the root causes of aviation accidents and reasonably analyze narratives and incident reports [1]. This method can fully utilize AI technology to identify patterns and build models, which is an effective way to analyze user interests.

In international research, the data silos problem has drawn much attention. Moritz Lell et al. verified the feasibility of breaking enterprise data silos for genome prediction modeling, and

this provides a valuable reference for user behavior analysis with scattered multi-platform data (social media, e-commerce, education portals) [2]. Bingchen Li et al. used a distributed deep-learning method to solve the data silos problem in landslide displacement prediction without sharing raw data during model training [3]. In recent years, the study of temporal dynamics has attracted more and more attention. Zachary F Fisher et al. proposed a time-series estimation method to characterize individual developmental trajectories in a population framework, which supports research on interest evolution [4]. Users have a common pattern of behavior, but there are also differences in behavior. Longyun Chen et al. built a model to describe the evolution of spatial and temporal graphs, which can be used in the study of brain network evolution [5]. The research results indicate that interpretability is essential when tracking relational structure changes.

In the process of education, a large number of data sets have been formed. Luise von Keyserlingk et al. analyzed student learning behaviors using multi-source data and predicted STEM performance, finding that temporal clustering directly affects academic outcomes and supports model construction for user behavior analysis [6].

In VR exhibition research, Shichun Z et al. collected multi-source behavioral data from 89 visitors. They used dynamic time warping and hierarchical clustering to identify three behavior patterns [7]. This validates multi-source data fusion and clustering's effectiveness for user behavior mining and provides a reference for interest portrait construction.

In the analysis process, innovation at the platform level is necessary. Dequan Zhang et al. built a multi-source data integration platform, which is important in industrial fermentation and shows data integration architecture advantages in decision-making [8]. These platforms don't focus on human behavior but highlight the importance of engineering principles in multi-channel data integration.

International research consensus states that integrating multi-source, multi-type and temporal data under security, causality and interpretability is crucial for accurate user interest modeling. Meanwhile, cross-field methods from education and mechanical diagnosis support robust portrait construction and digital ecosystem development.

1.3.2 Domestic Research Status

In recent years, domestic scholars have made breakthroughs in dynamic evolution and interest portrait construction research, and the integration of user modeling and spatial-temporal analysis has achieved ideal results. Xu Jianhui et al. have excellent performance in this area. Their personalized recommendation system can recommend academic resources to users in a personalized way, which is related to user portrait technology [9]. The approach in this study can describe user profiles in more dimensions and dynamically adapt to users' academic interests. This study further clarifies the use of interest portraits in digital education and can describe changes in user preferences through continuous data streams.

Chinese scholars applied multi-source data fusion to ecological and social research, refining the theoretical framework of behavioral interest modeling. Fan Hongmin et al. analyzed China's urban business environment spatiotemporally, focusing on macro-level changes and the behavioral responses of enterprises and policies [10] . While the study did not focus on personal interest portraits, their multi-source and longitudinal data analysis methods provide valuable references for large-scale interest dynamics research.

1.3.3 Literature Review and Comments

Existing studies have verified the importance of longitudinal data and time-series modeling in capturing dynamic user behaviors, while temporal graph learning, distributed deep learning, and multi-source fusion platforms provide technical support for interest evolution analysis and privacy-preserving modeling. Applications in user profiling and resource recommendation also demonstrate the practical value of interest modeling. However, current research still relies heavily on static portraits, ignores long-term evolution and contextual adaptability, and lacks sufficient discussion on ethical issues such as algorithm transparency. Future research should balance macro frameworks and micro behavioral patterns to support large-scale user analysis.

1.4 Research Content and Methods

This study uses multi-source heterogeneous data to investigate the dynamic evolution of user interest profiles, aiming to overcome insufficient dynamic modeling. It integrates e-commerce

behavioral logs (from the Taobao platform) and movie rating interaction data (from the MovieLens dataset) , and unifies cross-platform behavioral signals through fusion. NLP extracts fine-grained interest tags, and a time-series model with sliding window and attention mechanism captures temporal patterns to improve profiling accuracy for personalized services. A multi-stage method including multi-source data preprocessing, semantic space mapping, embedding alignment fusion, and NLP-based content analysis is adopted. The dynamic model tracks interest shifts, and experiments verify its effectiveness and robustness.

2. Theoretical Foundations and Technical Framework

2.1 Multi-source Heterogeneous Data

Multi-source heterogeneous data refers to data from different sources, which are different in structure, format, semantics and time. The data in the process of user interest profiling is heterogeneous, including unstructured data, semi-structured data and structured data. There are obvious differences in semantic granularity, syntactic representation and time resolution between different data sources. For example, the purchase history can reflect the user's preference, and the user's mouse movement can only reflect his curiosity. The integration of these signals is conducive to the formation of user portraits, which is a challenge in the process of construction. In the past five years, the amount of data has increased rapidly.

2.2 Core Concepts of User Interest Profiling

User interest profiling extracts preferences from digital footprints. Traditional single-source portraits are fragmented. A complete portrait consists of raw data, interest tags, and dynamic states. NLP enriches semantic tags, time-series modeling captures dynamics, and multi-source integration enhances comprehensiveness. Table 1 summarizes the characteristics of different data sources used in this process.

Table 1. Characteristics of Multi-Source Behavioral Data in Interest Profiling

Data Source	Data Type	Temporal Granularity	Primary Interest Signal
Social Media	Text, likes, shares	High (real-time)	Socially endorsed topics, trending interests
Search Logs	Queries, clicks	High	Active information needs, emerging curiosity
E-commerce Browsing	Page views, cart adds	Medium	Product-related preferences, intent to purchase
Transaction Records	Purchase history	Low (batched)	Confirmed preferences, consumption patterns
Video Streaming	Watch duration, skips	Medium-High	Content genre affinity, attention span

2.3 Data Fusion Methodologies for Behavioral Data

Data fusion unifies heterogeneous signals into a consistent representation using semantic alignment, entity resolution, and time normalization. Transformer and BERT-based embedding technologies support cross-modal fusion. Dynamic fusion considers recent and historical behaviors to enhance adaptability.

During the fusion of multi-source heterogeneous behavioral data, directly aggregating raw logs may introduce cross-platform data leakage risks. To mitigate this issue, a federated transfer learning framework can be implemented: each data source performs local extraction of interest labels and time-series feature encoding, while the central node aggregates only low-dimensional interest vectors or model gradients. This approach prevents the external transmission of original behavior sequences. Concurrently, differential privacy noise is added to statistical metrics during the aggregation process (e.g., category interest intensity, behavioral frequency within time windows), ensuring that the final published user profile group statistics satisfy $\varepsilon -$ differential privacy. This method effectively mitigates privacy risks without significantly compromising the modeling accuracy of interest evolution.

2.4 Fundamentals of Dynamic Evolution Modeling

User interest modeling needs to capture temporal dynamics and behavioral patterns. Static profiles can't reflect preference volatility, while dynamic models based on time-series analysis view interest trajectories as stochastic processes. By using sliding window and attention mechanisms, the model weights behaviors and captures user-system interactions. It assumes user preferences are relatively stable but change with external contexts and lifestyles, allowing for adaptive representation of complex behavioral

changes.

$$I_t=\alpha \cdot Attention(\beta_{t-k:t})+(1-\alpha) \quad (1)$$

In this formulation, I_t denotes the interest state at time t, $\beta_{t-k:t}$represents the behavior sequence in a sliding window, and α in [0, 1] controls the balance between new behaviors and historical interest states. The attention mechanism helps the model focus on salient actions to detect emergent interests or shifts.

2.5 Overview of Supporting Technologies: NLP and Time Series Analysis

NLP processes unstructured texts to extract sentiment, entities, and topics. Transformer-based models enable fine-grained semantic understanding. Time-series analysis captures temporal patterns and uses attention mechanisms for weight decay adjustment. The combination of these supports accurate and dynamic profiling.

3. Interest Portrait Construction Based on Multi-source Behavioral Data

3.1 Data Collection and Preprocessing

3.1.1 Sources and Types of Behavioral Data

First, the Taobao User Behavior Dataset [11] from Alibaba's Tianchi Platform contains 100 million behavioral records from approximately 988,000 anonymous users over ten days (November 24 to December 3, 2017). Each record includes user ID, product ID, product category ID, behavior type (click, favorite, add-to-cart, purchase), and timestamp, enabling multi-behavior interest modeling.

Second, the MovieLens dataset [12] is employed to validate cross-domain generalization. Specifically, we use the MovieLens 1M version, which consists of 1,000,209 anonymous ratings from 6,040 users on 3,706 movies. Each rating is associated with a timestamp (precise to seconds) and movie genre information. As noted by Harper and Konstan [12], the MovieLens datasets have been widely used as benchmarks for evaluating recommender systems because of their rich temporal information and standardized format. This dual-dataset setup allows us to evaluate our method in both e-commerce (Taobao) and movie recommendation (MovieLens) scenarios.

To represent heterogeneous behavioral streams mathematically, a unified event schema is defined. Let U denote the set of users, and for each user μ in U, their behavioral sequence is modeled as a time-ordered list of events $\varepsilon_u=\{e_1,e_2,\ldots,e_n\}$. Each event $e_i=\{t_i,s_i,c_i,d_i\}$ consists of timestamp t_i , source type $s_i \in \{browse,social,search,transaction\}$. , content payload c_i, and metadata d_i.

3.1.2 Data Cleaning and Normalization Techniques

Multi-source data has quality problems such as missing values, duplicates, and inconsistent encoding. A preprocessing pipeline is used, including time interpolation, mode filling, outlier detection, text normalization, user ID standardization, and cross-platform entity resolution. Feature normalization unifies semantic and numerical representation through min-max scaling, embedding, one-hot encoding, and z-score standardization. Timestamps are unified for dynamic modeling. These steps ensure data quality and consistency for user interest profiling.

3.1.3 Feature Extraction from Textual and Numerical Streams

In constructing user interest portraits, the fundamental task is to extract user behavior characteristics from text and data. Natural language processing technology is used for text data to understand user interests and semantic intentions. The TF-IDF weighting method analyzes keyword significance in text and compares keywords with the global corpus. For numerical data, standardization is done to obtain a standardized vector and analyze user behavior for a unified representation space.

We $T=\{t_1,t_2,\ldots,t_n\}$ set a sequence of text inputs and $N=\{n_1,n_2,\ldots,n_m\}$ a series of numerical indicators. The TF-IDF weight formula for a term in text input is $TF-IDF(w,t_i)=tf(w,t_i) \cdot \log\left(\frac{|D|}{|\{d \in D: w \in d\}|}\right)$, where D is the document set. When extracting numerical features, min-max normalization is used:

$$n_j^{'}=\frac{n_j-\min(N)}{\max(N)-\min(N)} \quad (2)$$

The extracted features are important for dynamic modeling and interest tag generation.

3.2 Multi-Model Data Fusion Strategy

3.2.1 Alignment of Heterogeneous Data Semantics

Semantic alignment is crucial for unified interest portrait construction and multi-source data integration. User behaviors vary across platforms and often lack contextual consistency. Polysemy and synonymy are major challenges,

tackled by embedding user-generated content and taxonomy labels into a shared vector space and using cosine similarity for mapping. Temporal features and interaction frequency are added to enhance contextual authenticity, and behaviors in the same time window are grouped to improve interest inference. The unified semantic space enables dynamic model updating for subsequent modeling.

3.2.2 Weighted Fusion Model Based on Behavior Reliability

In heterogeneous behavior integration, this study ensures semantic consistency and builds a data reliability mechanism. Different behavioral signals have different weights, with intentional actions (e.g., purchases, reviews) more reliable than accidental clicks. A scoring system based on intentionality (action depth), consistency, and recency (exponential decay) is proposed. Then, a weighted fusion model aggregates interest tags using reliability-adjusted scores. For a given interest category i, its fused strength S_i is calculated as $S_i = \sum_{b \in B_i} w_b \cdot r_b$ described, where B_i relevant variables denote behavior i set, w_b normalized weight from behavior type, and r_b reliability score from intentionality, consistency, and recency. This ensures high-intent, consistent, and recent behaviors strongly influence the final profile, aligning with real-world user psychology. Figure 1 illustrates the weighted proportion assigned to each behavior type within this fusion framework.

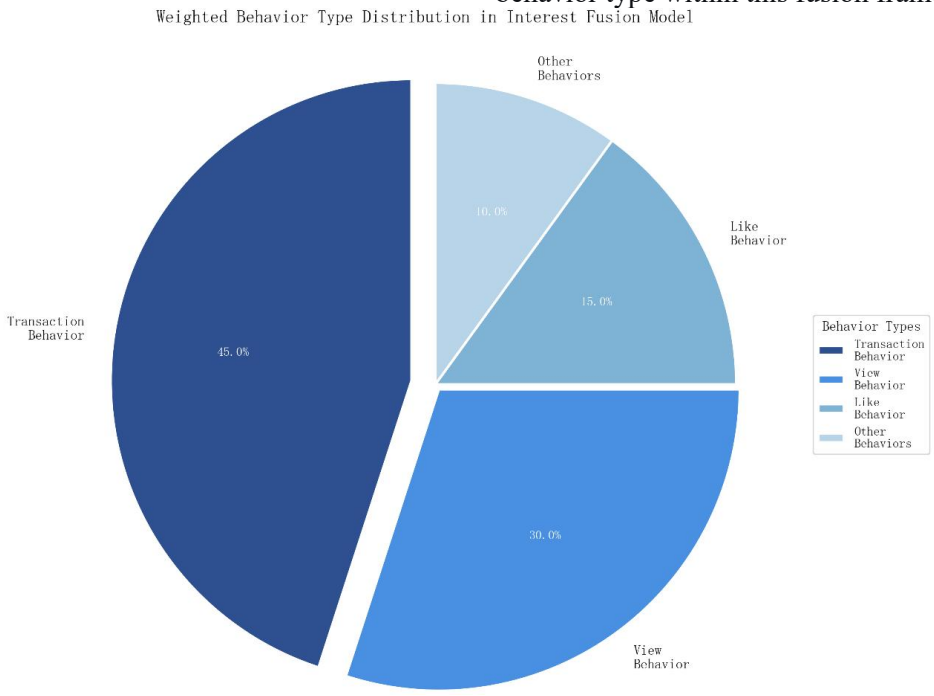

Figure 1. Weighted Proportion of User Behavior Types in the Interest Fusion Framework

3.3 Static Interest Profiling Using NLP Enhanced Methods

3.3.1 Topic Modeling and Keyword Extraction from Textual Behaviors

Static interest profiling extracts user interests from noisy, unstructured textual data. It combines BERT keyword extraction and LDA topic modeling to mine effective information. BERT is better than TF-IDF in semantic representation and polysemy resolution, and its combination with LDA ensures accurate topic extraction. In the behavioral-only baseline, user interest portraits are constructed solely from structured behavioral logs without utilizing any textual semantics. Meanwhile, the NLP-enhanced static portrait extends the behavioral baseline by incorporating textual semantics from product titles.

3.3.2 Validation and Evaluation of Constructed Interest Portraits

The effectiveness of the proposed interest portrait is confirmed by internal evaluations-using interest vector clustering and Jensen-Shannon divergence on a 500,000-user sample from the Taobao dataset-which demonstrate strong coherence, interpretability, and diversity, establishing the method as reliable and feasible for subsequent dynamic evolution modeling. Table 2 provides a performance comparison of the proposed models against baselines.

Table 2. Performance Comparison of Interest Portrait Models

Model Type	Taobao (AUC)	MovieLens (AUC)	Avg Improvement
Behavioral-only baseline	0.721	0.708	-
NLP-enhanced static portrait	0.782	0.769	+8.3%
Full dynamic model (for reference)	0.811	0.798	+9.6%

Note: The MovieLens results are averaged over 5-fold cross-validation following the standard protocol of Harper and Konstan [12].

4. Dynamic Evolution Analysis of User Interests

4.1 Temporal Granularity and Interest Drift Definition

4.1.1 Time Window Selection Strategies

A dual time window (10 days for short-term mutation, 30 days for long-term preference) is used. Stability and volatility are quantified to distinguish user groups. Young users are more fluctuating, while elderly users are more stable (Bobzien et al., 2025) [13]. Concept drift is detected via statistical and semantic methods. Table 3 presents age-specific usage patterns that inform our window selection strategy.

Table 3. Age-Specific Digital Platform Usage in Germany (SOEP-IS 2023)

User Segment	Daily Social Media Time (minutes)	Passive Users (%)	Purchase Based on Recommendation (%)
Young Adults (18–24)	199	~50%	70%
Professionals (25–34)	~150	~55%	70%
Seniors (65–74)	43	~75%	>50%

Source: Bobzien et al. (2025, Socius, Vol. 11) [13]

4.2 Time Series Modeling for Interest Trajectories

Interest vectors are constructed and updated dynamically. ARIMA and LSTM are compared for prediction, and LSTM performs better in nonlinear scenarios such as entertainment and fashion. Additionally, change-point detection and regime switching analysis identify persistence, mutation, and periodic patterns.

As reported by Ding and Li, in their study on personalized recommendation based on predictive analysis of user interests, they conducted five rounds of user simulation tests. Their proposed method achieved a 47.63% reduction in MAE (Mean Absolute Error) and a 44.61% reduction in RMSE (Root Mean Square Error) compared to baseline methods [14]. These results demonstrate the effectiveness of time-series modeling for capturing user interest dynamics, providing a valuable reference for the dynamic modeling framework adopted in this paper. To better illustrate the adaptability of different time-series methods in interest prediction, Table 4 provides a theoretical comparison between ARIMA and LSTM models, including their core characteristics and typical application scenarios in modeling user interest evolution.

Table 4. Theoretical Comparison of Time Series Models for Interest Prediction (Based on Ding & Li [14])

Model	Key Characteristics	Application in Interest Prediction
ARIMA	Linear, interpretable	Captures overall trend of interest evolution
LSTM	Nonlinear, captures long-term dependencies	Models complex interest fluctuation patterns

4.3 Experimental Results of Interest Evolution

4.3.1 Interest Evolution Heatmap

To visually demonstrate the dynamic nature of user interests, we randomly selected a representative user from the Taobao dataset and constructed an interest evolution heatmap following the methodology described in Section 3.3. The resulting heatmap, shown in Figure 2, visualizes the fluctuations in category-level interest intensity over time.

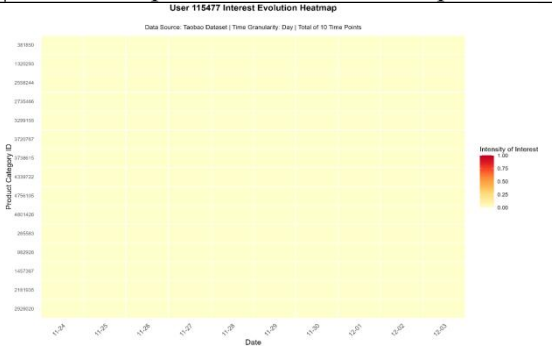

Figure 2. Interest Evolution Heatmap of a Representative User from Taobao Dataset

Table 5. Quantitative Metrics of Interest Evolution

Metric	Value
Observation period	10 days
Number of time points	10
Number of active categories	15
Average adjacent cosine similarity	0.742
Standard deviation of similarity	0.156
Detected interest shift points	1

5. Conclusion

This study creates a new multi-source dynamic framework to overcome limitations of traditional static user interest models. It integrates e-commerce behavioral logs (from the Taobao platform) and movie rating interaction data (from the MovieLens dataset) into a unified system with semantic enrichment and alignment to capture both explicit and implicit user demands. The proposed time-aware interest evolution model, enhanced by natural language processing for in-depth analysis of unstructured text, effectively characterizes the recurrence and abrupt changes of user interests. It offers theoretical and methodological innovations for user modeling and aligns with industrial trends. However, the framework is constrained by strict data privacy regulations, limited ability to capture long-term interests and conceptual drift, and insufficient modeling of data source reliability, which may cause noise and limit full performance.

References

[1] Tengfei Ren, Zhipeng Zhang, Bo Jia, Shiwen Zhang. Retrieval-Augmented Generation-aided causal identification of aviation accidents: A large language model methodology[J]. Expert Systems With Applications,2025,278

[2] Moritz Lell, Abhishek Gogna, Vincent Kloesgen, Ulrike Avenhaus, Jost Dörnte, Wera Maria Eckhoff, Tobias Eschholz, Mario Gils, Martin Kirchhoff, Michael Koch, Sonja Kollers, Nina Pfeiffer, Matthias Rapp, Valentin Wimmer, Markus Wolf, Jochen Reif, Yusheng Zhao. Breaking down data silos across companies to train genome-wide predictions: A feasibility study in wheat [J]. Plant biotechnology journal,2025,23(7):

[3] Bingchen Li, Changdong Li, Yong Liu, Jie Tan, Pengfei Feng, Wenmin Yao. Harnessing Distributed Deep Learning for Landslide Displacement Prediction: A Multi-Model Collaborative Approach Amidst Data Silos[J]. Journal of Earth Science,2024,35(5):

[4] Zachary F Fisher, Younghoon Kim, Vladas Pipiras, Christopher Crawford, Daniel J Petrie, Michael D Hunter, Charles F Geier. Structured Estimation of Heterogeneous Time Series [J]. Multivariate behavioral research,2024,59(6):

[5] Longyun Chen, Chen Qiao, Kai Ren,Gang Qu, Vince D Calhoun, Julia M Stephen, Tony W Wilson, Yu Ping Wang. Explainable spatio-temporal graph evolution learning with applications to dynamic brain network analysis during development [J]. NeuroImage,2024,298

[6] Luise von Keyserlingk, Fani Lauermann, Qiujie Li, Renzhe Yu, Charlott Rubach,Richard Arum,Jutta Heckhausen. Students' study activities before and after exam deadlines as predictors of performance in STEM courses: A multi-source data analysis[J]. Learning and Individual Differences,2025,117

[7] Shichun Z, Wang L, Xing Z. A Study on Visitors' Tour Patterns and Perceptual Experiences in VR Exhibitions Through the Integration of Multi-Source Behavioral Data [J]. International Journal of Human–ComputerInteraction, 2025, 41 (24): 15893-15914.

[8] Dequan Zhang, Wei Jiang, Jincheng Lou, Xuanzhou Han, Jianye Xia. Biofuser: a multi-source data fusion platform for fusing the data of fermentation process devices[J]. Frontiers in Digital Health,2024,6

[9] Jianhui Xu, Mustafa Man, Ily Amalina Ahmad Sabri, Guoyi Li, Chao Yang, Mingxue Jin. Research on Personalized Recommendation of High-Quality Academic Resources based on user Portrait[J]. International Journal of Advanced Computer Science and Applications (IJACSA),2022,13(10):

[10] Fan Hongmin, Liu Chang, Wang Qing song. Measurement, Spatial Differences, and the Dynamic Evolution of China's Urban Business Environment Levels[J]. Journal of Urban Planning and Development,2024,150(2):

[11] Alibaba Group. User Behavior Data from Taobao for Recommendation[DS/OL]. Tianchi Big Data Dataset, 2018. https://tianchi.aliyun.com/dataset/649.

[12] Harper, F. M., & Konstan, J. A. (2015). The MovieLens Datasets: History and Context. ACM Transactions on Interactive Intelligent Systems, 5(4), 1-19.

[13] Bobzien L, Verwiebe R, Kalleitner F. Visualizing age-specific digital platform usage in Germany[J]. Socius: Sociological Research for a Dynamic World, 2025, 11: 1-4. https://journals.sagepub.com/doi/10.1177/23780231251319360

[14] Ding Hao, Li Shuqing. Personalized Recommendation Based on Predictive Analysis of User's Interests[J]. Data Analysis and Knowledge Discovery, 2019, 3(11): 43-51 https://doi.org/10.11925/infotech.2096-3467.2019.0370

