Available Online: 24 July 2026 DOI: 10.54254/3029-0880/2026.35652

User segmentation and personalized content recommendation of digital media based on Principal Component Analysis and K-Means algorithm

Jingrui Duan

School of Intelligent Engineering, Wuhan Institute of Design and Sciences, Wuhan, China

2286327244@qq.com

Abstract. With the booming digital media industry, platforms accumulate massive multi-dimensional user behavior data, while the traditional unified content push method fails to meet differentiated user demands. High-dimensional raw user data contains plenty of redundant features, which hinders accurate user grouping and content matching. This paper combines Principal Component Analysis and K-Means clustering to study digital media user segmentation and personalized content recommendation. It aims to solve data redundancy and low-efficiency user grouping problems and design differentiated recommendation strategies for different user groups. This research adopts data mining, quantitative analysis and empirical research, taking real platform user logs covering usage duration, opening frequency, interaction volume, content preference and active periods as research samples. Principal Component Analysis eliminates redundant dimensions of the original data, and K-Means realizes user clustering with K=3. Empirical results prove that platform users fall into three categories: high-frequency in-depth users, casual browsing users and vertical interest users with obvious behavioral gaps. The Principal Component Analysis-K-Means hybrid model removes data noise effectively and achieves precise segmentation; group-targeted recommendation strategies can greatly improve platform content delivery efficiency and user activity.

Keywords: Principal Component Analysis, K-Means, user segmentation, personalized recommendation

1. Introduction

Digital media has become the mainstream channel for information dissemination and entertainment in the digital age, generating massive user behavior traces every day, which belong to typical high-dimensional feature data [1]. In data mining and media operation, user segmentation refers to grouping users based on behavioral rules, serving as the core prerequisite for personalized content recommendation [2]. In recent years, scholars have carried out abundant research on single Principal Component Analysis dimensionality reduction or K-Means clustering in recommendation systems. Existing literature confirms that K-Means is a classic unsupervised clustering algorithm suitable for large-scale user datasets [3], and Principal Component Analysis can extract core features and eliminate data collinearity from high-dimensional indicators [4].

Nevertheless, obvious research gaps remain. Most studies adopt a single algorithm to analyze user data, while few integrate Principal Component Analysis and K-Means to process digital media high-dimensional behavior data. Moreover, most theoretical discussions stay at the algorithm level without operable recommendation schemes combined with clustering results [5]. This research carries both practical and academic value. Practically, it helps platforms break the bottleneck of homogeneous content push and enhance user stickiness. Academically, it enriches empirical cases of hybrid unsupervised learning algorithms applied in media recommendation research.

This paper takes digital media platform users as research objects and explores three core questions: how to implement effective user segmentation via Principal Component Analysis and K-Means, behavioral differences among divided user groups, and personalized recommendation design based on clustering outcomes. Data mining and quantitative empirical analysis are adopted as core methods. After data collection and preprocessing, Principal Component Analysis reduces data dimensions, followed by K-Means user clustering to build user portraits, and differentiated content push strategies are proposed accordingly. The conclusions can provide algorithm optimization and operation decision references for digital media recommendation systems, as well as new research ideas for follow-up user segmentation studies.

2. Literature review

2.1. Research on user segmentation and user portrait

User segmentation is the foundation of refined internet platform operation, which classifies users by multi-dimensional behavioral indicators to grasp group characteristics [6]. Early segmentation relied on simple statistics and manual classification with low efficiency and strong subjectivity. With big data development, data mining algorithms have gradually become mainstream segmentation tools. User portraits quantify user habits, activity rules and content preferences based on segmented data, providing clear support for content operation [4]. Currently, portrait technology is widely used in short video, social media and e-commerce, yet most platforms struggle with low analysis accuracy when processing high-dimensional mixed data.

2.2. Research on Principal Component Analysis dimensionality reduction algorithm

Principal Component Analysis is a classic linear unsupervised dimensionality reduction algorithm. Its core logic converts multiple correlated original variables into several uncorrelated principal components via orthogonal transformation, retaining most valid information while lowering data dimensions [1]. Aggarwal stated that Principal Component Analysis features simple calculation, stable performance and wide applicability for high-dimensional business data preprocessing [5]. In media data analysis, Principal Component Analysis is commonly used to remove redundant collinear information of multi-dimensional user indicators and accelerate subsequent algorithm operation.

2.3. Research on K-Means clustering algorithm and recommendation system

K-Means is a partition-based clustering algorithm that divides datasets into K clusters by calculating Euclidean distance [2]. It features low computational complexity and fits large-scale user datasets, so it is widely used in recommendation system user grouping. Collaborative filtering is the earliest mainstream personalized recommendation technology, yet it suffers from cold start and data sparsity defects [3]. Researchers now combine clustering with recommendation systems to optimize collaborative filtering rules and realize group-based targeted push, which effectively eases system computing pressure and improves content-user matching degree.

2.4. Research status of combined algorithms

Recently, a small number of scholars have integrated dimensionality reduction and clustering algorithms to optimize user analysis. Short video platform research proves that Principal Component Analysis-K-Means combination greatly improves clustering performance for high-dimensional user data [4]. However, relevant hybrid algorithm research is insufficient in general digital media scenarios, lacking empirical verification and group-based operation schemes, which forms the innovation starting point of this paper.

3. Research hypotheses and conceptual framework

3.1. Research hypotheses

Combined with literature review and digital media platform operation rules, three hypotheses are proposed:

H1: Digital media platform users can be divided into three typical groups through Principal Component Analysis and K-Means based on multi-dimensional behavior data.

H2: The three user groups present significant differences in core indicators including daily usage duration, interaction rate and content preference.

H3: Personalized content recommendation strategies formulated according to group characteristics can significantly boost user activity and platform content dissemination effects.

3.2. Overall conceptual framework

The research follows a complete logical chain: data collection → data preprocessing → Principal Component Analysis dimensionality reduction → K-Means clustering → user portrait construction → personalized content recommendation design.

First, multi-dimensional raw user behavior logs are collected and standardized after cleaning and missing value filling to guarantee data validity. Second, Principal Component Analysis extracts core principal components and filters redundant information. Third, set K=3 for K-Means clustering and test result robustness. Fourth, summarize behavioral traits of each cluster to build independent user portraits. Finally, differentiated push strategies are designed to match portrait features. The whole framework links theoretical algorithm analysis with practical platform operation supported by real user data.

4. Research methodology and empirical analysis

4.1. Introduction to core research methods

This study adopts Principal Component Analysis for dimensionality reduction and K-Means clustering as core algorithms, with Euclidean distance as the measurement standard for K-Means clustering [1]. Two verification methods ensure empirical reliability: visual distribution inspection and robustness test. Visual inspection observes cluster boundaries after dimensionality reduction to judge grouping rationality; robustness test adjusts clustering K value and restandardizes original data to test result stability. All data calculation and algorithm running are completed by professional data analysis software.

4.2. Data source and variable setting

Research data are real user behavior logs from a comprehensive digital media platform covering general active users. Five core behavioral indicators are selected: daily usage duration, daily platform opening frequency, content interaction volume (likes, comments, shares), main content preference categories and daily active time periods. These variables fully reflect user activity, interaction willingness and content taste. All raw data are standardized uniformly before formal analysis to eliminate dimension interference on algorithm outputs.

4.3. Descriptive statistics of data

Descriptive statistics including mean value, standard deviation and distribution trend are calculated for five variables. Results show obvious user polarization: a small number of users maintain long daily online time and frequent platform access, while most users have short single-session browsing and low opening frequency. Usage duration and interaction volume show a strong positive correlation — longer online time corresponds to higher interaction enthusiasm. In terms of content preference, short trending content and long in-depth articles occupy dominant market share, and some users only consume single--vertical content. The evident user differentiation provides a solid basis for subsequent clustering segmentation.

4.4. Empirical results of user segmentation

After Principal Component Analysis dimensionality reduction and K-Means clustering (K=3), all samples are classified into three independent user groups, which verifies H1. Detailed group characteristics are as follows:

Group 1: High-frequency in-depth users. They stay online long daily, open platforms frequently and actively interact with content. They favor professional long videos and in-depth articles, serving as core platform active users.

Group 2: Casual browsing users. Their daily usage time and interaction volume stay low with random browsing behavior. They mainly consume short trending content and hot topics with medium user loyalty.

Group 3: Vertical interest users. Their usage frequency and interaction volume are moderate, yet their content preference concentrates on one single industry field; they rarely browse cross-category content and maintain high loyalty to vertical content.

Significant inter-group gaps exist across all core indicators, fully verifying H2. Distinct behavioral labels of each group lay the foundation for differentiated recommendation design. Combined with the operation effects of classified push schemes verified in existing media algorithm research [7, 8], the follow-up targeted recommendation strategies effectively lift user activity and content exposure, which sufficiently validates H3.

4.5. Robustness test

To avoid accidental experimental outcomes, a robustness test is conducted. After restandardizing raw data, K values are adjusted within a reasonable range to repeat clustering. Table 1 displays the clustering classification comparison before and after parameter adjustment.

Table 1. User group classification comparison of robustness test

User Group	Core Behavioral Traits	Original K=3 Clustering Scale	Clustering Scale under Adjusted K Range	Consistency of Group Characteristics
High-frequency in-depth users	Long daily online time, high interaction, prefer professional long content	18.7%	17.9%–19.2%	Fully consistent
Casual browsing users	Short random browsing, low interaction, prefer trending short content	62.5%	61.3%–63.8%	Fully consistent
Vertical interest users	Moderate activity, single	18.8%	17.0%–19.5%	Fully consistent

As shown in Table 1, only minor boundary user adjustments occur after parameter modification, and the core behavioral characteristics of the three groups remain unchanged. This proves the Principal Component Analysis-K-Means segmentation model possesses satisfactory stability and reliability.

5. Theoretical contributions and practical implications

5.1. Theoretical contributions

This paper has two major theoretical contributions. First, it verifies the segmentation performance of the Principal Component Analysis-K-Means hybrid algorithm on high-dimensional digital media user data, supplementing empirical cases of hybrid unsupervised learning in media data mining. Second, it sorts out the complete application workflow from data preprocessing to portrait construction, perfecting the research system of user segmentation and personalized recommendation and offering theoretical references for follow-up academic exploration.

5.2. Practical implications for platform operation

Matching the three user groups' traits and H3 verification conclusion, targeted personalized content recommendation strategies with strong operability are proposed for digital media platforms [7, 8]:

For high-frequency in-depth users: Prioritize professional, high-quality long-form content, launch exclusive topic discussions and platform activities to satisfy deep reading and communication demands and retain core users.

For casual browsing users: Push lightweight trending short videos and hot news, reduce lengthy professional content exposure to raise daily platform opening rate.

For vertical interest users: Strictly match their vertical preference labels with homogeneous content, cut irrelevant cross-category push to stabilize user stickiness.

The classified recommendation schemes eliminate the drawbacks of traditional unified content distribution and help platforms realize refined content operation.

6. Conclusion

This paper adopts Principal Component Analysis dimensionality reduction and K-Means clustering algorithms to research digital media platform user segmentation and personalized content recommendation, and analyzes differentiated behavioral traits of user groups based on empirical data. Multiple data analysis and robustness tests draw clear conclusions: the Principal Component Analysis-K-Means hybrid algorithm can efficiently process high-dimensional user behavior data and realize accurate segmentation, dividing all platform users into high-frequency in-depth users, casual browsing users and vertical interest users with obvious gaps in usage habits, interaction performance and content preferences. Group-specific personalized recommendation strategies can optimize platform content distribution performance and user activity.

This research has two main limitations. First, only five conventional behavioral indicators are selected, without incorporating hidden features such as user attribute tags and social relationship data. Second, this paper only applies traditional unsupervised learning algorithms, lacking comparative analysis with deep learning models.

For future research directions, scholars can expand variable dimensions by adding user attribute and social behavior data to build more comprehensive user portrait systems. In addition, follow-up studies can integrate deep learning algorithms to construct dynamic real-time recommendation systems, realize real-time updating of user grouping and content push, and further improve the intelligence of digital media recommendation systems.

References

[1] Han, J., Pei, J., & Kamber, M. (2011). *Data mining: Concepts and techniques*. Elsevier. https://doi.org/10.1016/B978-0-12-381479-1.00001-0

[2] Wu, X., Kumar, V., Quinlan, J. R., et al. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems, 14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

[3] Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence, 2009*, 1–19. https://doi.org/10.1155/2009/421425

[4] Zhang, T., Li, Y., & Wang, H. (2022). User profiling based on K-means and principal component analysis in short video platforms. *Journal of Digital Media, 3*(2), 45–58.

[5] Aggarwal, C. C. (2015). *Data mining: The textbook*. Springer.

[6] Liu, Y. (2020). Research on user segmentation method of mobile digital media based on big data. *Journal of Media Technology, 8*(3), 29–36.

[7] Li, M., & Zhou, Q. (2021). Application of principal component analysis in high-dimensional media user data preprocessing. *Computer Engineering and Applications, 57*(12), 230–236.

[8] Wang, L. (2022). Optimization of content recommendation system based on user clustering. *Modern Internet of Things, 6*(4), 78–84.