Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 2515962, 8 pages
https://doi.org/10.1155/2022/2515962

Research Article

Modeling and Analysis of Group User Portrait through WeChat Mini Program

Guangmin Li,[1,2] **Wenjing Chen**,[3] **Xiaowei Yan**,[4] **and Li Wang**[1]

[1]*College of Computer Information Engineering, Hubei Normal University, 435002, China*
[2]*College of Arts and Science, Hubei Normal University, 435002, China*
[3]*High-Tech Development Promotion Center Huangshi, 435002, China*
[4]*School of Computer Science, China University of Geosciences, 430074, China*

Correspondence should be addressed to Wenjing Chen; 1269639837@qq.com and Xiaowei Yan; kkuma7@outlook.com

Received 26 June 2022; Accepted 9 August 2022; Published 2 September 2022

Academic Editor: Lingling Lv

Meeting users' preferences and increasing business revenue is an ongoing challenge in the mobile service application. In this paper, we address these challenges by mining mobile user behavior patterns and propose an approach to construct a group user portrait by analyzing access data collected from the users of the WeChat Mini Program. We extract the attributes of mobile users considering their geographic information, online duration, and age group. Using Z-score standardized processing and K-means clustering algorithm, we then model the user portraits through three dimensions including daily average duration, interaction intensity, and access frequency. Our analysis has two important features. Firstly, the significant log data used in our experiments was collected from the production environment ensuring that the results reflect the real attributes of WeChat Mini Program users' behavior. Secondly, we provide data-driven decision-making to help marketers enhance the quality of the product and improve user experience. The experimental results indicate that by distilling and analyzing the key factors from the log data, the characters of typical users can be properly profiled to help product owners better optimize the exact set of the features which need to sustain and further grow.

1. Introduction

In recent years, mobile Internet applications have been extensively developed and the activity profile of mobile users has been significantly changed. The corresponding available event log data includes a wealth of information regarding the user behavior which can be mined to obtain useful insights for commercial incorporations. Such insights can be also used to dramatically improve user's experience and unearth hidden revenue opportunities, either through enhanced performance, customized user interface, or targeted advertisements.

WeChat Mini Program represented by Tencent is dominating the mobile ecosystem in China. According to the WeChat Mini Program Official Report, the daily number of active users (DAU) exceeded 400 million, and the average monthly use time of the Mini Program was 64 minutes as of March 2020. The permeability of the active users of the Mini Program accounted for 78.9%. Competition in this market has also recently increased. Therefore, one of the main challenges is finding new ways to effectively improve the functions of the product and maintain a high customer retention rate.

User portrait, namely, user profiling, refers to acquiring, extracting, and representing the features of the user in the form of a rich semantic-based structure [1, 2]. User profiling includes basic demographic information (such as name, gender, and nationality) as well as dynamic behavior information (such as interests and preferences) [3]. User portrait is widely used in several research fields. For instance, Ontika et al. [4, 5] presented a machine learning method to realize the identity of lyrics authors through user portraits. Also, Xu et al. [6] collected Douban movie data and the content of users' comments to design a social media resource aggregation model. They then used this model to establish a mapping relationship between the user portrait and the resource portrait, providing a reference for resource aggregation. Zhang et al. [7] also extracted the characteristics of a group of the paid user group and provided a three-dimensional user attribute based on their viewing data to retain the core users.

In the traditional retail industry, Gu et al. [8] proposed a psychological modeling method to profile the big five personality traits of the users with their emotion-bearing tweets to accordingly customize their personalized services.

Although there are many research works on the user portrait, there are still research gaps especially on the behavioral data generated from WeChat Mini Program. To the best of our knowledge, there are no systematic studies on the user portrait to profile the target users and provide decision support for the companies.

In this paper, we derive insights on identifying potential users and the importance of understanding different users' motivations and concerns. This paper utilizes log datasets including mobile users' behavior and mine the deep-level representative implicit information and outlines the specific groups of the user portrait. The datasets are collected from a medium-sized application called EnglishMyName and collected from September 5, 2018, to November 24, 2019. The multidimensional log data consists of geographical distribution, online duration, and user term query.

There are three main contributions in this paper: (1) Inferring the users' behavior characteristics and static attributes using mobile clickstream data; (2) building and deploying a WeChat Mini Program application to collect simultaneously behavior data online in compliance with the user privacy; and (3) providing practical suggestions for product operators and service providers using clustering technology and user portrait features.

The rest part of this paper is organized as follows: The related work is reviewed in Section 2. Section 3 introduces experimental preparation and data processing. Section 4 presents cluster analysis, compares four internal validation measures, and also profiles the groups of user portraits. Finally, the conclusions are drawn, and future research direction is discussed in Section 5.

2. Related Work

Alan Cooper, the Father of Interaction Design, introduced the concept of user portrait which is considered to be a fictional, specific, and concrete representation of the target user [9]. In the era of big data, massive data can reflect on user intentions, implying the user behavior patterns and interest preferences. Therefore, user portrait has been extensively investigated in recent years. Alan et al. [10] analyzed Twitter users from three perspectives, geographic location, gender, and belief, and found that the Internet users truly reflect the true population distribution in every area of the USA. Ruas et al. [11] identified different user behaviors through clustering methods based on the degree of interaction among Facebook users. They classified the users into three types: audience, participant, and content producer. Yu et al. [12, 13] also proposed variant regression algorithms to model mobile user gender and personality traits from their mobile phone sensory data. Based on the hotel review data, Shan et al. [14] constructed a user portrait from three dimensions, user information attribute, hotel information attribute, and user evaluation information attribute, and then provided the basis for merchants to understand customers' needs through precision marketing. Zeng et al. [15] conducted star fans' user portraits from social media topic data collected from Sina Weibo and dug out the targeted fans groups. This can help enterprises better tailor their marketing efforts. Liu et al. [16] concluded that the individual user portrait research is focused on specific users in a certain scene and labeled them with multifaceted features. It is suitable for distinguishing different users but not for exploring the user's behavior regulations in groups.

Furthermore, Akbari et al. [17] proposed that the group user portrait research can highlight users' behavior patterns and categorize them into different groups by clustering and association rules. The existing studies have modeled the user portrait based on the datasets mainly collected from social media platforms such as Twitter, Facebook, Weibo, and e-commerce platforms. However, few researchers paid attention to the user portrait modeling techniques on the WeChat Mini Program application. To bridge this gap, here, we explore geography distribution, online duration, and user query preference to mine the behavioral patterns of the mobile users.

Here, we use the K-means algorithm for clustering mainly due to its simplicity and efficiency [18–22]. Clustering here is used to categorize mobile visitors based on their behavioral data and formulate their corresponding relevant marketing strategies. Additionally, a considerable amount of the previous works on the behavioral attributes for user profiling inspired us to have a comprehensive group user profiling type [21, 23]. Considering the typical business scenarios, we further extract three behavioral features (i.e., daily average duration, interaction intensity, and access frequency) for user portrait inference to make clustering analysis results and decision more applicable.

3. Data Preprocessing and Data Analysis

In this section, we collect log items from the WeChat Mini Program app, EnglishMyName under a strict privacy policy. This app aims to serve Chinese users who pick a transliterated English name for their given name. The entire dataset contains 515, 684 items that occurred from September 5, 2018, to November 24, 2019, and are saved as CSV files. The files consist of user identification number, request time, operation type, and request content. The log list allows us to learn which page content has been accessed, at what time, where, and by whom.

The original request content is in an arbitrary format, but it must be machine-readable. To achieve it, we use data preprocessing which is critical for performing user behavior analysis before performing the data mining task. It includes data cleaning, data transformation, and data reduction.

The process of data cleaning is to remove noisy or irrelevant data. Specifically speaking, if a user just logs in to this app with no further action, the landing records are

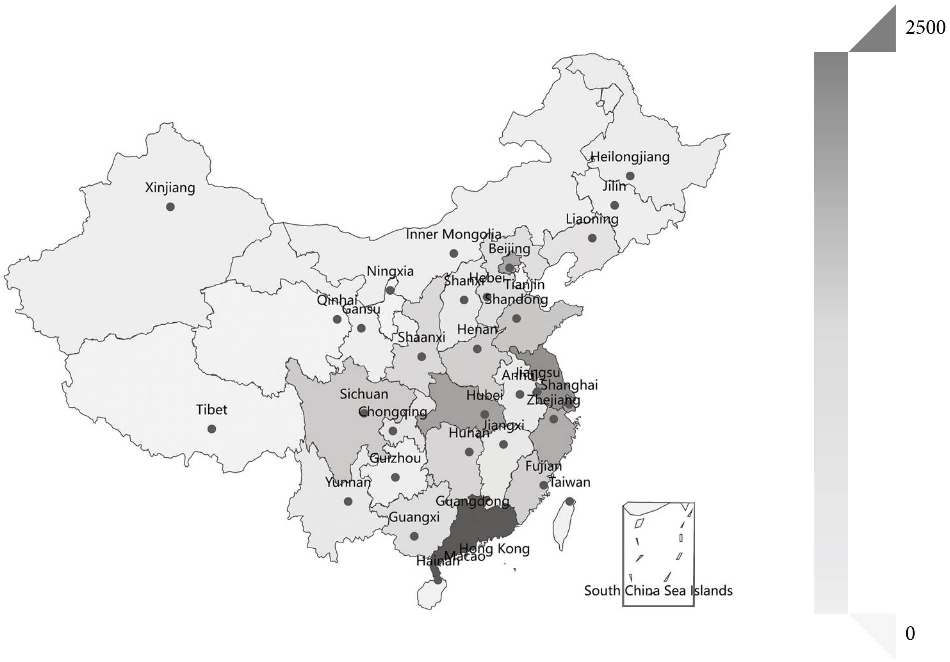

Figure 1: Distribution of the users' locations.

eliminated. During the development and testing phase, huge amounts of debug records are generated, and we eliminate them from the raw logs since they are irrelevant. Furthermore, automated programs like web crawlers are removed from the log files. On the other hand, the data format which is stored in chronological order must be changed into the tabular form that is at a relatively fine granularity. In the last phrase, the structured data is processed with the technique called Z-score normalization. Finally, we selected 19,383 records on online user behavior and used these resources to model and analyze the user portrait.

For one mobile application, each user action in one session shows differentiated performance. Inspired by the user information to profile [24–28], this paper illustrates mobile user behavior from multidimensional features such as location, online duration, and visitor's query term [29]. We then explore the behavioral patterns in each interactive session and categorize the users into different groups using the K -means clustering algorithm.

3.1. Geographic Distribution of User. In Figure 1, deeper color means a larger number of users. Through the terminal IP analysis of the province to which the user belongs, it is seen that the major users are from Guangdong province followed by Zhejiang and Shanghai. The number of habitats in these southeastern coastal cities is much higher than that of the inland cities. In the inland cities, users are highly concentrated in economically developed regions such as Beijing, Hubei, and Sichuan. The overall visitor of this application is mainly distributed in the provinces with high information levels and leading cultural development. Analyzing the geographical distribution of the users can help the decision-maker to evaluate the effect of targeted advertising and formulating the delivery strategies.

3.2. Analysis of User Online Duration. Analyzing the distributions of the frequencies over each operation during a day can improve the quality of experience (QoE) and save operational costs for service providers. Besides, it optimizes the allocation of system resources. For instance, operators can reduce network bandwidth during periods of low workload. We take a day as one periodical unit to count the distribution of requests from the users within one day. As shown in Figure 2, by analyzing the request frequency of all users' login time within the period in days, it is seen that the total number of visitors is low overnight before the beginning of a rapid ascent commencing at 4 a.m. Then, it remains active over the day, reaching its initial peak around 11 : 00 a.m., a tiny peak at 4 : 00 p.m., and an evening high between 8 and 10 p.m. It can be found that there is a peak in the morning, noon, and evening. The active periods are concentrated in (9 : 00 a.m. to 11 : 00 a.m.), (14 : 00 p.m. to 16 : 00 p.m.) and (20 : 00 p.m. to 23 : 00 p.m.), with the number of users in (20 : 00 p.m. to 23 : 00 p.m.) being the most active in the evening. This finding is in line with the "2018-2019 Mini Program Industry Growth Research Report" that indicates the active peak time is from 9 a.m. to 11 a.m. and from 2 p.m. to 4 p.m. The peak of users' activity mainly occurs during the idle period when they go off work or school. Combining with more multidimensional data, analyzing and verifying more thorough users' occupation distribution for users are our future work.

3.3. Age Group Estimation. Understanding the correlation between the customers' demographics (e.g., age and gender) and behavior is essential for marketing operators. Age group estimation plays a key role that prompts companies to target potential customers in the right place at the right time and enhance their service [30]. As shown in Figure 3, the birth

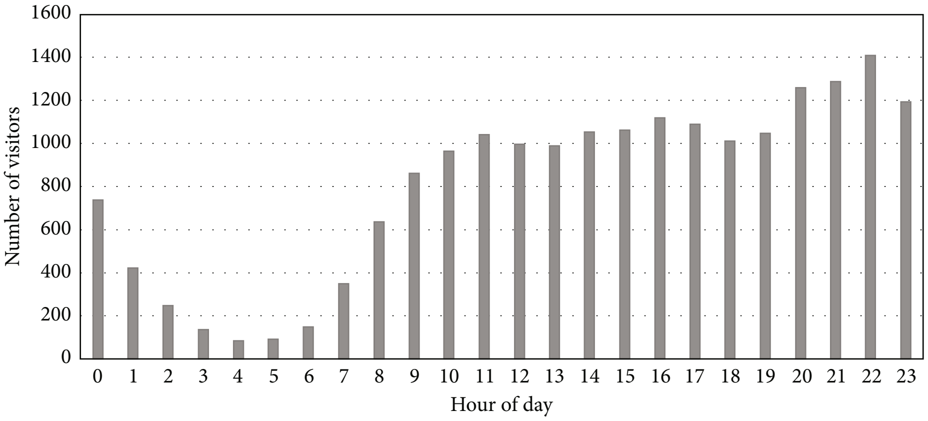

Figure 2: Distribution of the hourly number of online users.

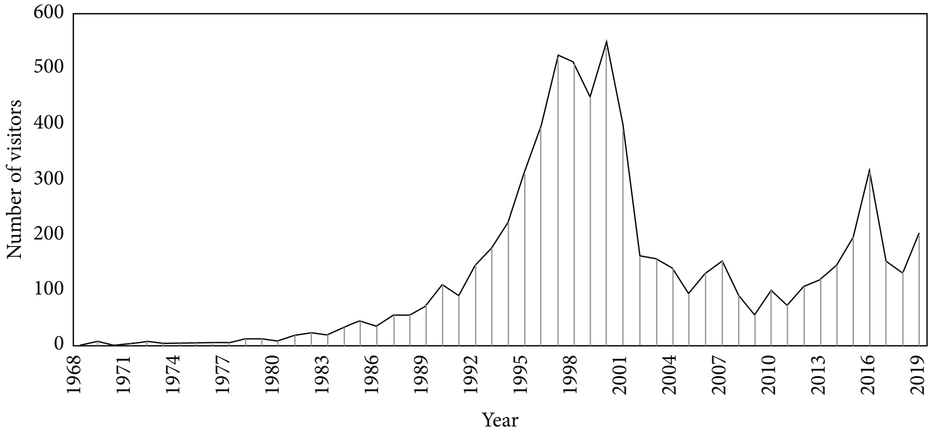

Figure 3: Distribution of the users' birth years.

year selected by users is mostly between the late 1990s and early 2000s. It can be known from the calculation that the age of the user group should be between 18 and 23 years old as of the year 2019, and most of them are university students. Also, we find the number of visitors increases steadily since 2011 until there is a local peak in 2016. For the reason that the population is statistically aged from 3 to 8, we can reasonably assume that they have little ability and accessibility to the Internet and most likely their parents used this application for them.

4. *K*-means Clustering Analysis

In this section, we describe our unsupervised method to build a user behavior model from clickstream log data. Clustering is the process of organizing data into classes that are internally cohesive and well-separated. As a representative of the distance-based unsupervised clustering algorithms, K-means [31] assigns each data point to the cluster which has the closest centroid [32] and produces tighter partitions than the hierarchical clustering. Usually, N samples or observations are divided into K clusters, and the K value is specified by the elbow rule and silhouette score.

Clustering based on original statistical data is feasible in theory. In practice, it has very rarely done so—not only will the redundant and nonbusiness-related features increase the computational complexity, but also such results are not much practical significance.

Users in distinct life stages have different intentions and perform different activities in the process of obtaining information, and the valuable behavioral features within each group are often hidden in the original statistical data. According to the characteristics of clickstream log data and specific business, we construct such indicative features: average interval between user operations(s), average number of operations per session, and average duration of each session(s).

(i) *Average interval between user operations*. It means how much time on interval between two operations in one session

(ii) *Average number of operations per session*. It counts how many times the users interact with this app in one session. The session is defined as a period wherein a user is actively engaged with an app

(iii) *Average duration of each session*. It shows how long users every session last

4.1. Data Normalization. Data normalization is one of the preprocessing in data mining and is especially needed for

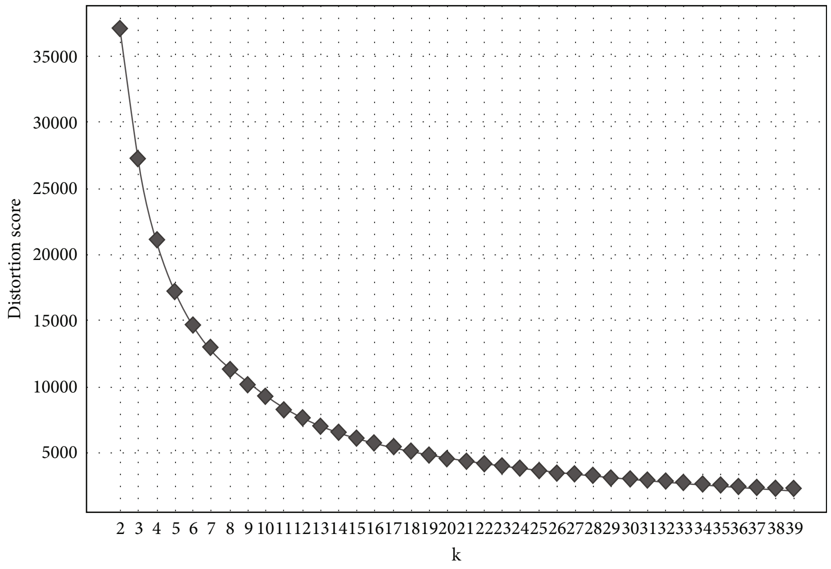

Figure 4: Distortion score of elbow for K-means clustering.

distance metrics, such as Euclidian distance which is sensitive to differences in the magnitude or scales of the features. The distance-based K-means algorithm will give a higher weighting to the variable with higher magnitude, so we use data normalization for all variables to overcome the bias.

The importance of normalization is that it can generate high-quality clusters and improve the performance of clustering algorithms [33]. There are different normalization techniques such as min-max, Z-score, and decimal scaling. In this paper, we adopt Z-score normalization as it handles outliers well, and do not have a predetermined range. The calculation formula is as follows:

$$x_{\text{norm}} = \frac{x - x_{\text{mean}}}{x_{\text{std}}}. \tag{1}$$

4.2. Identify K Value. Before performing K-means clustering on the above data, we need to get the optimal value of K. The most prevalent methods for determining the right cluster numbers are the elbow technique and the silhouette method. The elbow method measures the sum of squared error (SSE) for different K value to choose the most appropriate number of clusters. The SSE will decrease as the K value rises. The lower the SSE, the fewer the samples in each cluster are and the more homogenous they are. We choose K around the point where the degree of distortion lowers the most. The silhouette method estimates the cohesion and the separation [34]. For clustering results, we want modest differences inside clusters and huge disparities across clusters; therefore, the optimal cluster number is determined by the highest score of the index, and the calculation formula is as follows:

$$s(i) = \frac{b(i) - a(i)}{\max\,(a(i), b(i))}, \tag{2}$$

where $b(i)$ represents the mean distance between point i and all sample points in the nearest cluster and $a(i)$ represents the mean distance between point i and other sample points in its cluster.

To identify the optimal K value, we evaluate the performance of the elbow method and the silhouette method on the mobile user log dataset. We use the K-means algorithm with K ranging from 2 to 40. Figures 4 and 5 demonstrate the experimental results. We also consider the specific business on our real-world dataset to be much more realistic and actionable; hence, we subdivide the data sets into 8 groups to keep clustering accuracy. As shown in Figure 4, at $k = 8$, the line graph begins to flatten significantly, and the sum of squared distance (SSE) is 11322.847 when $k = 8$. In Figure 5, the silhouette score is 0.49 at $k = 8$. The structures of clustering results are illustrated in Figure 6. Distinct clusters are clearly shown in 3 dimensions and the clustering algorithm groups data points into nonoverlapping subgroups in a clear and distinct way.

4.3. Analysis of Group Portrait. By qualitatively analyzing the clustering results, we divide the potent visitors into eight groups. The result is consistent with the Pareto principle, generally known as the 80/20 rule. In other words, 20% of users devote 80% of the app's network volume.

Persona 1: As can be seen from the Table 1, this persona contains more operations per session on average, with 77 operations per session, despite the short interval between operations (an average of 12.737 s). From the high number of actions and short intervals, we can presume users may be interested in the main functionalities of this app, but the specific content does not address their genuine demands. The conductor can conduct a questionnaire to capture the essence the users' wants and desires, based on the survey results, taking efforts to improve the content supply to satisfy the needs of diverse users.

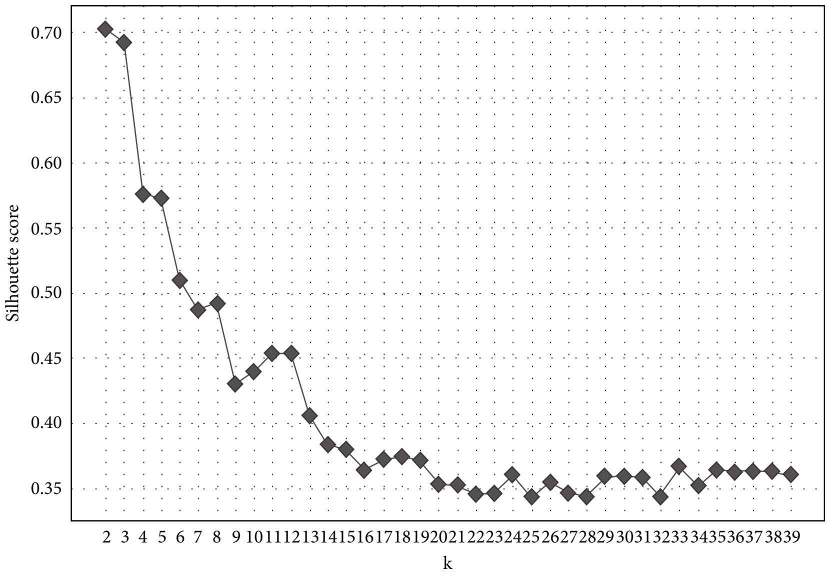

Figure 5: Silhouette score for K-means clustering.

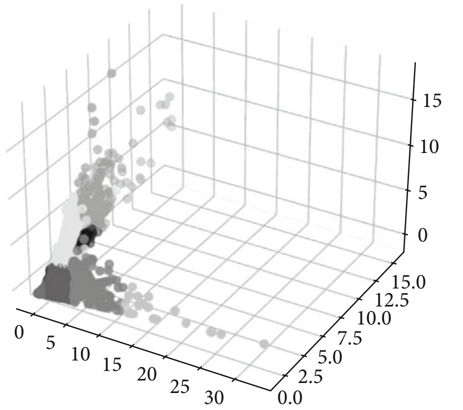

Figure 6: Clustering effect when k is equal to 8.

Persona 2: It accounts for 20.92% of all the analyzed users. The average interval between user actions is 15.071 seconds, with each session having 32 operations on average and lasting 424.124 seconds or nearly 7 minutes. The small number of operations and short duration time imply that this persona does not have a strong desire for continuous usage, which might mean that the content does not meet user demands or that the novice guiding instruction is too complicated. The conductor should perform an investigation to discover the possible causes behind user churn and focus efforts on enhancing capabilities and periodically sending updates to users in order to create a powerful chance to deepen users' comprehension of the product and reestablish bonds with them.

Persona 3: This group of users, on average, spends around 9 minutes in one session, and their average access frequency is 11 times per session. It is assumed that this type of user has an intention to use. To build customer loyalty, the conductor may utilize collaborative recommendations to deliver more information depending on users' interests.

Persona 4: It is the largest cluster with about 65.11% of users with the mean interval between two operations of 14.343 s, each session contains only 9 operations on average, and the mean session duration time is about 1 minute. Based on the statistics, it is reasonable to conclude that this user group does not touch the app's module thoroughly. The conductor can provide eye-catching content to them to pique their attention and encourage the users to return.

Persona 5: Although there are fewer operations in this session, the average session duration is greater, at 18 minutes, and the gap between operations is longer. The session contains fewer operations, but the average session duration is longer, at 18 minutes, and the interval between operations and operations is longer. This signifies that the page's content is popular among users. To raise users' dependency on the product, strengthen their sense of belonging to the platform, and improve user retention rate, the conductor needs completely comprehend their attractive content and perform targeted pushing.

Persona 6: This persona has the most operations in each session (173) and the longest session duration (43 minutes on average). According to the data, this persona has a high level of loyalty and trust in the platform and is the valuable customer who should be the conductor's first focus. Persona 6 is essential for achieving a virtuous cycle of development. To develop a closer relationship with this persona, incentivize consumers with tailored service, points, and unique offers.

Persona 7: This persona has the fewest users, the longest average gap between operations (570.866 s), and the smallest average number of operations each session (just 4). We believe this persona has achieved its objectives with only a few activities. Therefore, for those users, we can employ material incentives to stimulate them to share the product with their friends.

Table 1: The mean value of the behavior features for various users.

Cluster Id	Average interval between user operations (s)	Average number of operations per session	Average duration of each session (s)	Number of users in the cluster	The proportion of users
1	12.737	77	881.229	1024	5.28%
2	15.071	32	423.124	4054	20.92%
3	73.242	11	576.579	887	4.58%
4	14.345	9	100.069	12621	65.11%
5	235.898	8	1087.636	122	0.63%
6	17.497	173	2639.856	181	0.93%
7	570.866	4	957.968	21	0.11%
8	50.088	49	1938.294	473	2.44%

Persona 8: This type of user has a higher access frequency and is interested in the features of the app and willing to take time to indulge in it. For such great potential users, the conductor should focus on delving into the public character of the users and, on that basis, adopt suitable product development strategies to raise users' degree of satisfaction and promote the conversion of such active users into the valuable user.

5. Conclusions

The amount of time spent on using mobile apps (especially for the WeChat Mini Program) has been significantly increased in recent years. User behavior patterns should be thoroughly modeled to provide the operators with deeper insights and expand their targeted customer market. To tackle this important problem, in this work, we visually categorized users with real-world data into different user portrait groups. In particular, we presented a classical clustering analysis quantitatively and qualitatively, where we combine the static user information, e.g., their region, birth year, and gender with their dynamic attributes, such as users' online duration, interaction intensity, and access frequency to find correlations between WeChat Mini Program users. We further compared various internal evaluation measures for the most appropriate clustering results. To the best of our knowledge, this is the first research work to model Mini Program user behavior patterns that provides actionable insights for practitioners. The findings of how to label the targeted mobile users in this paper enhance future products and help companies keep their competitive advantage.

Further research directions include applying more sophisticated algorithms to identify social users from log data; Deng et al. [35] proposed for frequent pattern mining user identification algorithm, extending more comprehensive key attributes to explore user portraits. Consequently, we plot precisely user portraits and then instantly prompt decision-makers to regulate market strategy.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this article.

Authors' Contributions

Guangmin Li, Wenjing Chen, and Xiaowei Yan contributed equally to this work.

Acknowledgments

This work is partly supported by the Natural Science Foundation of Hubei Province of China (Grant 2020CFB828); the Teaching Research Project of Hubei Normal University (Grant XJ202001); and the Outstanding Young and Middle-aged Technological Innovation Team of Hubei Province of China (Grant T201430).

References

[1] X. Zhou, Y. Xu, Y. Li, A. Josang, and C. Cox, "The state-of-the-art in personalized recommender systems for social networking," *Artificial Intelligence Review*, vol. 37, no. 2, pp. 119–132, 2012.

[2] X. Tao, Y. Li, and N. Zhong, "A personalized ontology model for web information gathering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 4, pp. 496–511, 2011.

[3] S. Kanoje, S. Girase, and D. Mukhopadhyay, "User profiling trends, techniques and applications," https://arxiv.org/abs/1503.07474.

[4] N. N. Ontika, M. F. Kabir, A. Islam, E. Ahmed, and M. N. Huda, "A computational approach to author identification from Bengali song lyrics," in *Proceedings of International Joint Conference on Computational Intelligence, Algorithms for Intelligent Systems*, M. Uddin and J. Bansal, Eds., pp. 359–369, Springer, Singapore, 2020.

[5] L. Zhang, C. Xu, Y. Gao, Y. Han, X. Du, and Z. Tian, "Improved Dota2 lineup recommendation model based on a bidirectional LSTM," *Tsinghua Science and Technology*, vol. 25, no. 6, pp. 712–720, 2020.

[6] H. Xu, H. Zhang, M. Wei, and H. Yin, "Research on the construction of social media user portrait and resource aggregation model," *Library and Information Service*, vol. 63, no. 9, pp. 109–115, 2019.

[7] L. Zhang, X. Zhang, H. Lu, and L. Zhang, "Research on user persona of knowledge online Live's paid-up members," *Library and Information Service*, vol. 63, no. 5, pp. 84–91, 2019.

[8] H. Gu, J. Wang, Z. Wang, B. Zhuang, and F. Su, "Modeling of user portrait through social media," in *2018 IEEE international conference on multimedia and expo (ICME)*, pp. 1–6, San Diego, CA, USA, 2018.

[9] A. Cooper, "The inmates are running the asylum," in *Software-Ergonomie'99*, U. Arend, E. Eberleh, and K. Pitschke, Eds., vol. 53 of Berichte des German Chapter of the ACM, p. 17, Vieweg, Teubner Verlag, Wiesbaden, 1999.

[10] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and J. Rosenquist, "Understanding the demographics of Twitter users," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, pp. 554–557, 2011.

[11] P. H. B. Ruas, A. D. Machado, M. C. Silva et al., "Identification and characterisation of Facebook user profiles considering interaction aspects," *Behaviour & Information Technology*, vol. 38, no. 8, pp. 858–872, 2019.

[12] Z. Yu, E. Xu, H. Du, B. Guo, and L. Yao, "Inferring user profile attributes from multidimensional mobile phone sensory data," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5152–5162, 2019.

[13] L. Zhang, Z. Huang, W. Liu, Z. Guo, and Z. Zhang, "Weather radar echo prediction method based on convolution neural network and long short-term memory networks for sustainable e-agriculture," *Journal of Cleaner Production*, vol. 298, article 126776, 2021.

[14] X. Shan, X. Zhang, and X. Liu, "Research on user portrait based on online review: taking Ctrip hotel as an example," *Information Studies: Theory & Application*, vol. 41, no. 4, pp. 99–104 +149, 2018.

[15] H. Zeng and S. Wu, "User image and precision marketing on account of big data in Weibo," *Modern Economic Information*, vol. 16, pp. 306–308, 2016.

[16] L. Liu, S. Wang, and Z. Hu, "A literature review on community profiling," *Library and Information Service*, vol. 63, no. 23, pp. 122–130, 2019.

[17] M. Akbari and T.-S. Chua, "Leveraging behavioral factorization and prior knowledge for community discovery and profiling," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 71–79, Cambridge, United Kingdom, 2017.

[18] P. Berkhin, "A survey of clustering data mining techniques," in *Grouping Multidimensional Data: Recent Advances in Clustering*, J. Kogan, C. Nicholas, and M. Teboulle, Eds., pp. 25–71, Springer, Berlin, Heidelberg, 2006.

[19] J. Blömer, C. Lammersen, M. Schmidt, and C. Sohler, "Theoretical analysis of the k-means algorithm—a survey," in *Algorithm Engineering*, L. Kliemann and P. Sanders, Eds., vol. 9220 of Lecture Notes in Computer Science, , pp. 81–116, Springer, Cham, 2016.

[20] L. Bottou and Y. Bengio, "Convergence properties of the k-means algorithms," *Advances in Neural Information Processing Systems*, vol. 7, 1994.

[21] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[22] X. Wu, V. Kumar, J. Ross Quinlan et al., "Top 10 algorithms in data mining," *Knowledge and Information Systems*, vol. 14, pp. 1–37, 2008.

[23] S. Pandey, M. Aly, A. Bagherjeiran et al., "Learning to target: what works for behavioral targeting," in *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1805–1814, Glasgow, Scotland, UK, 2011.

[24] R. Saia, L. Boratto, S. Carta, and G. Fenu, "Binary sieves: toward a semantic approach to user segmentation for behavioral targeting," *Future Generation Computer Systems*, vol. 64, pp. 186–197, 2016.

[25] K. Tsiptsis and A. Chorianopoulos, *Data Mining techniques in CRM: inside customer segmentation*, John Wiley & Sons, 2010.

[26] P. Khanthaapha, L. Pipanmaekaporn, and S. Kamonsantiroj, "Topic-based user profile model for POI recommendations," in *Proceedings of the 2nd International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence*, pp. 143–147, Phuket, Thailand, 2018.

[27] S. Mohamed and A. I. Abdelmoty, "Spatio-semantic user profiles in location-based social networks," *International Journal of Data Science and Analytics*, vol. 4, no. 2, pp. 127–142, 2017.

[28] S. Zhao, F. Xu, Z. Luo, S. Li, and G. Pan, "Demographic attributes prediction through app usage behaviors on smartphones," in *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pp. 870–877, Singapore, Singapore, 2018.

[29] C. Wu, F. Wu, J. Liu, S. He, Y. Huang, and X. Xie, "Neural demographic prediction using search query," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 654–662, Melbourne VIC, Australia, 2019.

[30] Z. Qin, Y. Wang, H. Cheng, Y. Zhou, Z. Sheng, and V. C. M. Leung, "Demographic information prediction: a portrait of smartphone application users," *IEEE Transactions on Emerging Topics in Computing*, vol. 6, no. 3, pp. 432–444, 2018.

[31] S. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[32] V. R. Patel and R. G. Mehta, "Impact of outlier removal and normalization approach in modified k-means clustering algorithm," *International Journal of Computer Science Issues (IJCSI)*, vol. 8, no. 5, p. 331, 2011.

[33] M. Souto, I. G. Costa, D. Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC Bioinformatics*, vol. 9, no. 1, 2008.

[34] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[35] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A user identification algorithm based on user behavior analysis in social networks," *IEEE Access*, vol. 7, pp. 47114–47123, 2019.