Academic Education Publishing House -AEPH-

Building a User Profiling Model for Product Complaints in Open Innovation Communities: A Digital Transformation Perspective

Shizhuo Pan*
School of Management, Zhengzhou University, Zhengzhou, China
**Corresponding Author*

Abstract: Open innovation communities serve as crucial platforms for enterprises undergoing digital transformation, facilitating the integration of internal and external innovation resources to enhance efficiency. The increasing diversity and value of user feedback highlight the need for research on transforming dispersed product complaints into valuable innovation resources. This study develops a user profiling model for product complaints within open innovation communities, using the MIUI community as a case study. User complaints are categorized into four themes: positioning deviation, strap material, activity detection, and health monitoring. This approach aims to uncover diverse needs and preferences hidden in user complaints, providing insights for enterprises to better align product innovation with user expectations.

Keywords: Digital Transformation; Open Innovation Community; User Complaint Profiling; K-Modes

1. Introduction

In recent years, a new wave of technological revolution and industrial transformation has accelerated, with digital technologies such as big data, cloud computing, mobile internet, blockchain, and artificial intelligence becoming more active than ever. These technologies are reshaping the global economic landscape and industrial structure, and have also driven changes in enterprise management and optimization. Digital transformation in enterprises is a process that supports the use of digital technologies, products, and platforms to comprehensively transform, adjust, and optimize business management activities, ultimately affecting financial performance and market value [1].Against this backdrop, open innovation has gradually replaced traditional "closed" innovation [2]. Open innovation communities, empowered by mobile internet technology, are developing rapidly, and many enterprises regard them as an important source of product development innovation, such as Huawei's Product Definition Community, Haier's Idea Platform, Dell's IdeaStorm Community, Xiaomi's MIUI Community, and other open innovation community platforms. User participation in product innovation has become key for enterprises to obtain external knowledge and innovation resources [3].Currently, an increasing number of users are freely expressing their opinions or needs within communities, participating in discussions about ideas or tasks initiated by enterprises [4]. Among these, there are also negative complaints and grievances from users during the product usage process. Although this can have a negative impact on reputation, it provides valuable external perspectives for developers, helping to identify and solve problems encountered by users during product usage, thereby optimizing product functions and deepening understanding of market demands.

However, the surge in user feedback leads to information overload. Clustering and characterizing product complaint users with similar features to customize more personalized solutions for them is crucial for enhancing user satisfaction and fostering brand loyalty. At the same time, how to effectively connect user demands with corporate innovation and co-create value through platforms such as open innovation communities has become a serious challenge that enterprises must face to maintain a competitive advantage in the process of deepening digital transformation. Based on this, this paper uses unsupervised sentiment analysis technology to retain negative post data of product complaints, extracts basic attributes, content features, and influence features of product complaint users to establish a user profiling label system, constructs user profiling labels, and builds a user profiling model. The final output of the profiling results distinguishes different types of product complaint users, and proposes targeted product improvement suggestions, aiming to provide guidance and direction for corporate product improvement and innovation practices.

2. Research Overview

2.1 Open Innovation Communities

Open innovation communities are online discussion forums where users of a specific brand, unrestricted by geography, share their experiences with products or services, propose improvements and solutions, and engage in social interactions. These communities have become crucial platforms for businesses to leverage external resources to understand customer needs and achieve success in product innovation [5].Currently, scholars have conducted extensive research on open innovation communities, focusing primarily on user participation and value co-creation.

Firstly, from the perspective of user participation: (1) In terms of user role classification, You Qijun et al. [6] used the RFM model, considering the last post date, average hourly post volume, and points as indicators to build a user classification model for creative crowdsourcing communities, categorizing users into five types: star users, genius users, loyal users, regular users, and dormant users; Shan Xiaohong et al. [7], on the basis of constructing a user-knowledge bipartite network, proposed a link prediction method based on the exponential random graph model, identifying 20 leading users in different fields of Huawei's product definition community. (2)Regarding the acquisition of user needs, Ruan Pingnan et al. [8] achieved the extraction of demand keywords by calculating word weights and used a quadrant chart to clarify the current distribution of demand points in the Xiaomi community; Zhao Jinghua et al. [9], considering the importance of various product attributes in selected reviews as influential factors, proposed a user need identification framework that integrates RF-BERT and UGC. (3) In terms of idea identification, Xue Shui Jing et al. [10] took the Open IDEO open innovation community as an example, constructing a valuable idea identification and evaluation model from three dimensions: user characteristics, user participation, and idea content features, selecting the best-performing machine learning algorithm for classification to improve the efficiency of idea identification; Liu Jiayu et al. [11] moved away from the single text features in previous research on idea adoption, integrating diverse heterogeneous data based on graph models to build a prediction model for idea adoption in open innovation communities. Secondly, from the perspective of value co-creation, Yan Jianyuan et al. [12] categorized customers into regular and special based on the frequency of their interactions with the community and, based on the Construal Level Theory, analyzed the differences in customers' motivations to participate in value co-creation in product innovation communities; Zhang Ning et al. [13], based on the Signaling Theory, explored the factors influencing the adoption of user concepts in open innovation communities from different levels, and pointed out that the entropy of creative information and the intensity of emotion have a positive impact on the adoption of ideas at the level where consumers use corporate products to create value.

2.2 User Profiling

The concept of user profiling was first introduced by the father of interaction design, A. Cooper [14], who defined it as a virtual reflection of real users, constructed based on massive amounts of actual user data to form a target user model. It is a tagged user model abstracted based on characteristics such as natural attributes, social attributes, living habits, and behaviors, with different types of user groups having distinct tags, allowing for group differentiation through these tags. The steps in user profiling modeling generally include data processing, tag establishment, algorithmic aggregation, and profile presentation [15]. Specifically, in the context of virtual communities, scholars primarily use web crawlers to automatically collect user data, extract user tags, and construct profiling models based on clustering algorithms [16], topic modeling [17], and social network analysis [18], to output user profiles. For instance, Xia Lixin et al. [19], in the field of health communities, refined user behavior through the RFM model and identified four types of groups: "observers," "promoters," "high-yield," and "experienced," revealing the different needs of specific groups for health information; Chen Tianyuan et al. [20], in the context of university library services, integrated multi-source data to build a user profiling identification and recommendation analysis framework, capturing user behavioral characteristics through RFCLS tags and using a random forest model for fine-grained classification of user needs, ensuring that library service recommendations accurately meet the diverse information needs of users; Yang Yang et al. [21] focused on Bilibili's educational video barrages, integrating barrage text content and user behavioral timing information to construct a more comprehensive user profile, revealing users' learning patterns and information needs, helping barrage video platforms better understand user requirements and preferences.

3. Research Design

This paper constructs user personas for product complaints based on user data from open innovation communities, proposing three stages: data processing, model construction, and persona generation. In the data processing stage, user data from open innovation communities are crawled using Python programs. After data preprocessing, sentiment analysis is employed to identify comments containing negative emotions, from which user complaint information is screened out. Secondly, in the user persona model construction stage, three dimensions—basic attributes, content features, and influence characteristics—are extracted to establish a demand persona labeling system. The BERTopic model [22] is utilized to determine product attribute topics, and user influence characteristics are quantified through influence indices. Finally, in the persona generation stage, the K-Modes clustering algorithm is adopted to perform clustering analysis on the processed dataset, outputting user personas for product complaints. The specifics are illustrated in Figure 1:

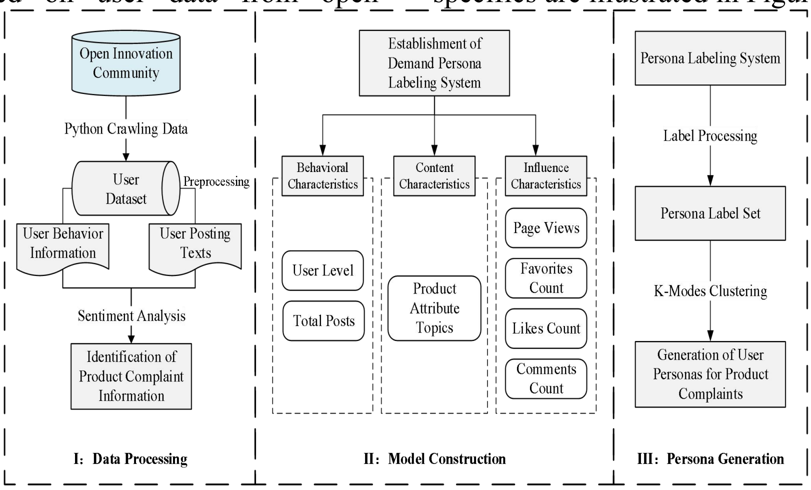

Figure 1. Research Framework

3.1 Data Processing

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

3.1.1 Data collection and preprocessing

Data is collected from open innovation communities using Python. The preprocessing steps for post texts are as follows: (1) Cleaning: Remove URLs, punctuation, emoticons, etc., and eliminate duplicate or promotional posts that are not valid. (2) Tokenization: Use the Jieba library to tokenize the text. (3) Stopword removal: Integrate the stopword list from the Machine Intelligence Laboratory of Sichuan University, Harbin Institute of Technology's stopword dictionary, and Baidu's stopword list to build a custom stopword list, which is continuously updated and refined by manually adding domain-specific words to filter out common but irrelevant terms in the text.

3.1.2 Product complaint information identification

Sentiment analysis is the process of analyzing, processing, and extracting subjective texts with emotional tones using natural language processing and text mining techniques [23]. Posts with negative sentiment polarity often indicate user dissatisfaction or complaints about products, suggesting product deficiencies and the need for product improvements to enhance user satisfaction. To effectively identify product complaint information with a negative emotional tendency, this paper employs an unsupervised dictionary-based sentiment analysis method, using the BosonNLP sentiment dictionary [24], which is built on social media sentiment-annotated data, and integrates the negation and degree adverb dictionaries provided by CNKI HowNet to extract posts with negative emotions from numerous texts.

The specific technical steps are as follows: First, read the BosonNLP sentiment dictionary to obtain sentiment words and their corresponding sentiment scores, and obtain negation words and degree adverbs from CNKI HowNet, integrating them into the sentiment analysis model. Secondly, by iterating through the tokens in each post, calculate the overall sentiment score of each post text, with the calculation formula shown as Equation (1). Texts with a final sentiment score greater than 0 are classified as positive sentiment, while texts with a sentiment score less than 0 are classified as negative sentiment.

$$finalSentiScore = \sum(sentimentScore \times (-1)^{num\ of\ notWords} \times degreeMultiplier) \quad (1)$$

In this context, represents the sentiment value of the sentiment word. denotes the number of negation words preceding the sentiment word; if there are any, the sentiment value's sign is reversed.is the intensity value of the degree adverb, which is used to adjust the sentiment value of the sentiment word.

Finally, based on the returned sentiment scores, posts with a negative emotional tendency are retained, and product complaint users are filtered from these negatively inclined texts using the following criteria: (1) The post contains ten or more characters; (2) The post uses adjectives or adverbs; (3) The post explicitly mentions product attributes or functions.

3.2 Model Construction

Profile labels serve as the foundation for establishing a profiling model, representing pre-set identifiers that summarize corresponding characteristics. The specific construction process of the user profile label system in this section is as follows: Firstly, building upon the research indicators of existing open innovation community user profiles, the study also draws on research from other related virtual communities [18];Secondly, element extraction is based on the characteristics of the selected open innovation community and the data features of the users, which involves extracting elements from the user profiles, post texts, and interactive data of product complaint users; Finally, based on the actual situation of product complaint users, irrelevant attribute elements are eliminated, and basic attributes, content features, and influence features that align with the characteristics of the research subjects are retained. Ultimately, a user profiling feature indicator system is established, as shown in Table 1.

Table 1. User Persona Tagging System

Dimension	Tag Name	Tag Description
Basic Attributes	User Level	User level advancement is achieved through the accumulation of activity points.
	Total Posts	The total number of posts made by the user during the data collection period
Content Features	Product Attribute Topics	The BERTopic model is employed to determine the most probable topic category for each post.
Influence Features	Views	The total number of views on the user's posts during the data collection period.
	Likes	The total number of likes received on the user's posts during the data collection period.
	Comments	The total number of comments received on the user's posts during the data collection period.
	Favorites	The total number of favorites received about a user's posts during the data collection period.

3.3 Profile Generation

Portrait generation mainly involves the aggregation of labels, and the aggregation mainly relies on clustering algorithms to achieve. The traditional K-means algorithm is based on the Euclidean distance, which is only applicable to continuous data, and has some deviation for discrete data or mixed data with both. K-Modes algorithm, as an extended version of K-Means algorithm in dealing with categorical data, has the core advantage of being able to efficiently deal with the dataset containing categorical attributes (e.g., labels in the user's portrait), and it is optimized in terms of initial clustering center selection, convergence speed, and dataset size adaptation, compared with the K-Means algorithm. Means algorithm, compared with the K-Means algorithm, it is optimized in terms of initial clustering center selection, algorithm convergence speed and dataset size adaptability, with higher clustering performance and stronger applicability, which can be better adapted to the characteristics of the user portrait dataset, whereby the algorithm is used for clustering of the processed portrait label dataset mentioned above in order to generate the product complaint user portraits.

Steps of the k-modes Algorithm [25]:

(1) Random Selection of Initial Centroids:

Randomly select k initial centroids from the dataset.

(2) Assignment of Data Points to Clusters:

For each sample point in the dataset, the Hamming Distance (HD) [26] between the sample point and the k centroids is calculated, i.e., the number of different characters in the corresponding positions of two strings of the same length, and the dissimilarity calculation function is defined as:

$$d_H(X,Y)=\sum_{i=1}^{n}\delta(X_i,Y_i) \qquad (2)$$

$$\delta(X_i,Y_i)=\begin{cases}0\ (X_i=Y_i)\\1\ (X_i\neq Y_i)\end{cases} \qquad (3)$$

For two user sample datasets X and Y, each sample dataset comprises a series of feature indicators, including user level, total number of posts, product attribute topics, as well as the number of likes, comments, and views.

Wherein: $d_H(X,Y)$ represents the Hamming distance between samples X and Y. X_i and Y_i are the values of the i-th feature indicator for user sample datasets X and Y, respectively. n denotes the number of indicators included in the sample. $\delta(X_i,Y_i)$ signifies the dissimilarity function, which is utilized to determine whether the user profiles of the two samples are identical in the i-th feature indicator.

If the values of X_i and Y_i are the same, then $\delta(X_i,Y_i)=0$ indicates no difference in that feature indicator; If they differ, $\delta(X_i,Y_i)=1$ indicates a difference in that feature indicator.

Assign each data point to the cluster of the nearest centroid.

(3) Update of Centroids:

After assigning all data points to clusters, update the centroids of each cluster. The new centroid for a cluster is determined by the mode of each attribute within the cluster, i.e., the most frequent value of each attribute among the data points in the cluster.

4. Iteration Until Convergence:

Repeat steps 2 and 3 until the total distance (sum of the distances between data points and their respective centroids) no longer decreases. The algorithm then returns the final clustering result.

4. Case Study

4.1 Data Processing

4.1.1 Data collection and preprocessing of MIUI community's "Wristband Circle"

Python programs were written to crawl user data from the MIUI Community's "Wristband Circle" from January 5,2024, to August 30, 2024, obtaining information such as user growth levels and total posts, as well as capturing likes, comments, and views for each post text. The preprocessing steps for post texts are as follows: First, filter out duplicate posts, community administrator posts, promotional posts, etc., retaining 15,829 valid post texts; second, perform Jieba tokenization on the cleaned texts; finally, introduce the stopword list from the Machine Intelligence Laboratory of Sichuan University, the Harbin Institute of Technology's stopword dictionary, and Baidu's stopword list, and establish a custom stopword list, adding high-frequency but meaningless words like "wristband" and "Xiaomi" to improve the accuracy of tokenization.

4.1.2 Identification of product complaint information in MIUI Community's "Wristband Circle"

The emotional score of a user's post is determined by the emotional words in the text and their corresponding emotional intensity values. Negation words can change the emotional tendency of the emotional words, while degree adverbs can either strengthen or weaken the emotional intensity. This paper uses an emotion analysis method based on the BosonNLP dictionary, which takes into account the impact of emotional words, negation words, and degree adverbs, returning an emotional score value to represent the emotional tendency of the text. By calculating the emotional scores of the 15,829 valid post texts extracted from the MIUI Community's "Wristband Circle," an emotional binary classification result of the user posts was achieved. Ultimately, 8,139 valid negative post data were retained.

4.2 Construction of User Profiling Model in MIUI Community's "Wristband Circle"

Based on the user profiling indicator system, user profile labels are established. This section refines user profile labels from three dimensions: basic attributes, which describe user levels and total posts, can be directly obtained without processing; the remaining content features and influence features require processing to obtain the corresponding user label attributes.

For content features, BERTopic is used to determine the product attribute themes within the content features. First, the model parameters are continuously optimized until the output is the most clearly defined themes. Secondly, after manual review, it was found that some themes have a high similarity between the theme words and the text content. Therefore, referring to the relevant research on the classification of product attribute themes in the wristband field, and combining expert opinions, similar themes were screened and merged, ultimately resulting in 10 product attribute themes and theme words, as shown in Table 2.

For influence features, the Influence Index is used to measure, as shown in Equation (3) [27].

$$I = \omega_1 I_p + \omega_2 I_l + \omega_3 I_c + \omega_4 I_v \tag{3}$$

Here, $\omega_1, \omega_2, \omega_3, \omega_4$ are the weight values for I_1, I_2, I_3 and I_4 respectively.

The basic principle of weight assignment is as follows: if one of the four actions of browsing, liking, commenting, or collecting occurs less frequently, this indicates that the user must invest more time or have a strong interest in the post to perform the action. Therefore, the action that occurs less frequently during browsing, liking, commenting or favoriting is assigned a higher weight [28].

Table 2. Extraction of Negative Posting Topics Related to Product Attributes for Smart Bracelets

Topic Number	Topic Name	Topic Keywords
Topic0	Health Monitoring	Heart Rate, Sleep, Stress, Monitoring
Topic1	Temperature Control	Heating, Battery Consumption, High Temperature, Temperature, Hot
Topic2	Screen Operation	Freeze, Reboot, Lag, Black Screen, Malfunction
Topic3	Screen Display	Photos, Display, Brightness, Outdoor, Screen-On
Topic4	Strap Material	Strap, Contact, Allergy, Cleaning, Crack
Topic5	Connection Status	Mobile Phone, Bluetooth, Pairing, Disconnection, Reboot
Topic6	Battery Life	Battery Life, Days, Charging, Battery, Power Drain
Topic7	Location Inaccuracy	Location, GPS, Trajectory, Position, Inaccurate
Topic8	Vehicle Integration	NFC, Wearable, Unlock, Automobile, BYD
Topic9	Sports Monitoring	Heart Rate Broadcasting, Running, Cycling, Cadence, Distance

The final categorization of user characteristics and quantified values for product complaints in the Wristband Circle is presented in Table 3.

Based on the aforementioned user profiling labels, the unsupervised K-Modes clustering algorithm is utilized for clustering. To achieve the optimal clustering outcome, it is essential to first determine the optimal number of clusters. Following the elbow method criterion, the sum of squared errors (SSE) for clusters ranging from 1 to 10 were calculated, and the resulting curve is depicted in Figure 2. The point where the curve's rate of decrease significantly diminishes is taken as the approximate value for the optimal K value; thus, the optimal number of clusters is determined to be 4.

Table 3. Classification of User Persona Characteristics and Quantification Values

Dimension	Tag Name	Quantification of Characteristic Tags
Basic Attributes	User Level	Users with levels below 3 are classified as beginner users; those with levels 4-7 are

		classified as intermediate users; and those with level 7 are classified as advanced users.
	Total Posts	Users with a total number of posts below 30% are labeled as low-output users; those with a total number of posts between 30% and 70% are labeled as medium-output users; and those with a total number of posts above 70% are labeled as high-output users.
Content Characteristic	Product Attribute Topics	The user product attribute topic for each post is determined using BERTopic and labeled sequentially as Topic0-Topic9.
Influence Features	Views	Users with a weighted value below 30% are labeled as having high influence; those with a weighted value between 30% and 70% are labeled as having general influence; and those with a weighted value above 70% are labeled as having low influence.
	Likes	
	Comments	
	Favorites	

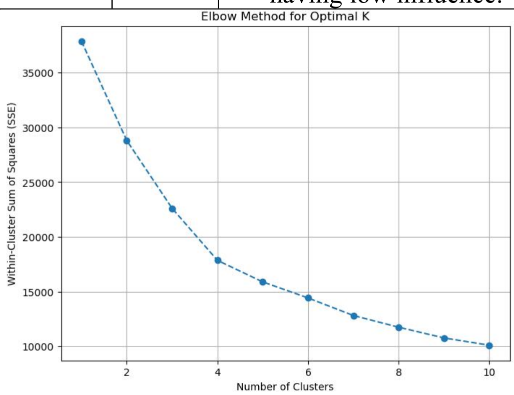

Figure 2. Sum of Squared Errors (SSE) Change Curve

4.3 Generation of User Personas for Product Complaints in the "Bracelet Circle" of the MIUI Community

The optimal number of clusters was determined to be K=4. Statistical analysis was conducted on the clustering results, with the cluster centers regarded as typical representatives of each user group. User personas for each category were output according to the product attribute topics in the content characteristics, as shown in Table 4.

From Table 4, it can be observed that the user persona of Type 1 primarily focuses on high-level users who have a high total number of posts and primarily concern themselves with the product attribute theme of positioning deviation. These users exhibit high influence characteristics. Therefore, Type 1 users can be summarized as "High-Influence Positioning Deviation Focusers." The user persona of Type 2 is also dominated by high-level users with a high total number of posts, but they show a greater interest in the product attribute theme of wristband material, exhibiting moderate influence characteristics. This indicates that Type 2 users can be categorized as "Moderate-Influence Wristband Material Focusers." Type 3 users consist of medium-level users with a medium total number of posts, primarily focusing on the product attribute theme of sports monitoring and demonstrating high influence characteristics. Consequently, Type 3 users can be described as "High-Influence Sports Monitoring Focusers." Finally, Type 4 users are also composed of high-level users with a medium total number of posts, who focus on the product attribute theme of health monitoring and exhibit high influence characteristics. This makes Type 4 users "High-Influence Health Monitoring Focusers."

Table 4. Results of User Personas in the Bracelet Circle

Category		Category1	Category2	Category3	Category4
Basic Attributes	User Level	High	High	Medium	High
	Total Posts	High Output	High Output	Medium Output	Medium Output
Content Character-istic	Product Attribute Themes	Positioning Deviation	Strap Material	motion detection	Health Surveillance
Influence Character-istics	Weighted Values of Views, Likes, Co-mments, and Favorites	High Influence	Moderate Influence	High Influence	High Influence

5. Conclusion

In the wave of digital transformation, enterprises are facing unprecedented challenges and opportunities. Based on this background and a review of previous user personas, this paper conducts an empirical study on user complaint data from the "Bracelet Circle" of the Xiaomi Community by constructing a user persona model for product complaint users in an open innovation community. The aim is to reveal the implicit different demand characteristics of product complaint users and provide practical guidance for managing product complaint users on open innovation community platforms and formulating strategies for improving next-generation products. The research process is divided into three stages: data processing, user persona model construction, and product complaint user persona generation. Ultimately, four user groups with distinct characteristics are successfully identified, revealing users' focus points and influence characteristics regarding different product attributes. This provides data support for precise marketing and personalized services in the community, contributing to optimizing user experience and enhancing product satisfaction.

This paper still has certain limitations. Future research can build on this study by further exploring the long-term behavioral patterns of different user groups and their acceptance and feedback effects on product improvement suggestions to achieve deeper user demand mining and value co-creation. Meanwhile, with the development of data science and artificial intelligence technologies, more advanced algorithms and models can be explored to improve the accuracy and practicality of user personas, providing more scientific decision support for product innovation and user management in open innovation communities. This will enable enterprises to seize opportunities in the tide of digital transformation.

References

[1] Furr N, Shipilov A. 2019. "Digital doesn't have to be disruptive: The best results can come from adaptation rather than reinvention". Harvard Business Review, 97(4): 94-104.
[2] Martinez-Torres R. "Analysis of open innovation communities from the perspective of Social Network Analysis". Intangible Capital, 2013, 9(1). DOI: 10.3926/ic.352.
[3] Wan W H, Liu L J. "Impact of internal and external collaboration between employees and users on innovation performance of platform enterprises: Based on the perspective of value co-creation". Nankai Business Review, 2021, 24(2): 72-84.
[4] Schroder A, Holzle K. "Virtual communities for innovation: Influence factors and impact on company innovation". Creativity and Innovation Management, 2010, 19(3): 257-268.
[5] Di Gangi P M, Wasko M. "Open innovation through online communities". Knowledge Management and Organizational Learning, 2009: 199-213.
[6] You Q J, Ma Y B. "Research on user role classification in creative crowdsourcing communities - Based on the RFM model". Productivity Research, 2022, (12): 125-128+144.
[7] Shan X H, Wang C W, Liu X Y, Han S X, Yang J. "Identification of lead users in open innovation communities-A knowledge-based view". Data Analysis and Knowledge Discovery, 2021, 5(09): 85-96.
[8] Ruan P N, Zhao Y Q. "Research on user demand identification based on open innovation communities - Taking Xiaomi community as an example". Soft Science, 2017, 31(12): 20-24.
[9] Zhao J H, Xie W Y, Lv X T, Zhao J L. "User demand identification and its development trend prediction based on RF-BERT and UGC". Information Science, 2024, 1-22.
[10] Xue S J, Guo W, Lian W T, Zhang J. "Research on creative identification methods in open innovation communities based on machine learning". Packaging Engineering, 2023, 44(08): 72-80.
[11] Liu J Y, Li H, Shen W, Zhu L L, Li S Y. "Research on the prediction of creative adoption in open innovation communities integrating multi-source heterogeneous online comments". Journal of the China Society for Scientific and Technical Information, 2024, 43(01): 48-60.
[12] Yan J Y, Qiao Y F, Qin F. "Research on value co-creation behaviors of different levels of customers in product innovation communities-Taking the MIUI community as an example". Management Review, 2019, 31(02): 58-70.
[13] Zhang N, Zhao W F, Pang Z L, Zhang X. "Research on influencing factors of creative adoption in enterprise open innovation communities - A value co-creation perspective". Science & Technology Progress and Policy, 2021, 38(16): 91-100.
[14] Cooper A. "The inmates are running the

asylum: Why high-tech products drive us crazy and how to restore the sanity". 2nd ed. New York: Sams Publishing, 2004: 176-193.
[15] Xiong R, Donath J. "PeopleGarden: Creating Data Portraits for Users". ACM Symposium on User Interface Software & Technology. ACM, 1999. DOI: 10.1145/320719.322581.
[16] Zhou L X, Wang S. "Research on survival portraits of multiple concurrent cutting-edge disruptive soft technology limit competitions". Library and Information Service, 2024, 68(04): 97-108.
[17] Zhang L, Shen K, Zhang N. "Research on user portraits of institutional accounts in social Q&A communities from the perspective of activity theory". Library and Information Knowledge, 2024, 41(03): 93-104.
[18] Li D. "Text Mining Model for Virtual Community User Portrait Based on Social Network Analysis". Strojarski fakultet u Slavonskom Brodu; Fakultet elektrotehnike, računarstva i informacijskih tehnologija Osijek; Građevinski i arhitektonski fakultet Osijek, 2019(4).
[19] Xia L X, Hu P, Liu K H, Zhai S S. "Research on user portraits of online health communities integrating information recommendation scene elements". Library and Information Knowledge, 2023, 40(03): 116-128.
[20] Chen T Y, Mei X. "Empirical research on user portrait identification and recommendation based on multi-source data fusion". Information Theory and Practice, 2024, 47(04): 171-180.
[21] Yang Y, Yu W J. "Research on user portraits integrating content features and behavior features of barrage - Taking B-station teaching videos as an example". Information Science, 2022, 40(12): 161-169.
[22] Roman E, Joanne Y. "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts". Frontiers in Sociology, 2022, 7886498-886498.
[23] Rangel B G R. "An Emotional Analysis of False Information in Social Media and News Articles". ACM Transactions on Internet Technology, 2020, 20(2).
[24] Min K, Ma C, Zhao T, et al. "BosonNLP: An Ensemble Approach for Word Segmentation and POS Tagging". Springer International Publishing. Springer International Publishing, 2015.
[25] Khan S S, Ahmad A. "Cluster center initialization algorithm for K-modes clustering". Pattern Recognition Letters, 2004, 25(11): 1293-1302.
[26] HUANG Z X. Extensions to the K-means Algorithm for Clustering Large Data Sets with Categorical Values. Data Mining and Knowledge Discovery, 1998, 2(3): 283-304.
[27] Ye J X, Xiong H X, Yi M, Liu M. "A social network group recommendation method integrating influence propagation". Journal of the China Society for Scientific and Technical Information, 2022, 41(4): 364-374.
[28] Gao J C, Peng L H, Zhang Y F, Jiang X; Hong C. "Construction and empirical study of user portrait model for health anxiety in online medical communities". Library and Information Service, 2023, 67(16): 124-134.