A Study on the User Group Classification of Online Q&A Communities Based on Knowledge Sharing Behavior

Xiran Han [1, #], Qingyang Ma [1, #, *], Wenbo Liu [1], Zhihang Fu [1], Zhuoman Huang [1], Feng Hu [2]

[1] Beijing-Dublin International College at Bjut, Beijing University of Technology, Beijing, China, 100124

[2] Beijing University of Technology College of Economics and Management, Beijing University of Technology, Beijing, China, 100124

* Corresponding Author Email: qingyang.ma@emails.bjut.edu.cn

[#]These authors contributed equally.

Abstract. User classification is crucial for studying knowledge-sharing behavior in online communities. A well-designed classification system can reveal the distinct characteristics and needs of different user groups, enhance the specificity of information retrieval, and ultimately improve the efficiency of knowledge sharing, thereby bringing substantial social benefits to information exchange. This paper utilizes natural language processing techniques to develop an analytical framework based on three dimensions: information entropy, the relevance between questions and answers, and sentiment polarity. A novel user classification method is proposed, categorizing users into four types: routine, knowledge-driven, responsive, and indifferent. The knowledge-sharing value corresponding to each type is calculated and compared. This method not only addresses the limitations of traditional digital feature indices but also, by integrating management theory with quantitative analysis, provides fresh perspectives and strategies for the healthy development of online Q&A communities and the enhancement of knowledge-sharing efficiency.

Keywords: Online Q&A communities, user classification, knowledge sharing, natural language processing.

1. Introduction

Online Q&A communities are knowledge service platforms based on the Internet, where users pose questions, provide answers, and engage in discussions [1]. With the rapid advancement of science and technology and the widespread adoption of the Internet, the extensive penetration of social media has significantly contributed to the rise of online Q&A communities. In recent years, community Q&A websites have proliferated, with many emerging platforms. Websites such as Reddit, Quora, and Zhihu have become increasingly popular among users seeking information, sharing knowledge, and solving problems [2].

However, the knowledge shared in online Q&A communities, generated through the interaction of users and content, is often fragmented. The value of this collective knowledge sharing has been largely overlooked, leading to difficulties for users in filtering relevant information when searching for specific answers. As a result, the efficiency of knowledge sharing is reduced [3]. Therefore, scientifically classifying the user groups in online social media included Q&A communities, can improve the specificity of information retrieval, reveal the characteristics and needs of different user groups, and provide a basis for community management and optimization. This is of significant theoretical and practical importance for enhancing the quality and efficiency of knowledge sharing in online Q&A communities [4].

There has been some research on the classification of user groups in online Q&A communities based on knowledge sharing behavior both domestically and internationally. In early Internet-related research, users of online communities were primarily classified into active contributors, lurkers (who only browse but do not participate), and peripheral participants [5]. With the continuous development of research on the classification of Internet user groups, the list of categories has been expanded.

Zhang Jiantong and Chu Weichao, using Zhihu as a case study, employed social network analysis and factor analysis to classify users into five categories: core users, questioners, responders, lurkers, and active users. This not only revealed the status of knowledge sharing in the Zhihu community but also provided targeted strategies for community management [6].

At the same time, numerous studies have explored the impact of user classification on knowledge sharing. Angeletou et al. argued that an important goal of user classification research is to improve the efficiency of knowledge sharing. By monitoring community activities, we can better understand and predict whether a community is developing positively or negatively, thereby proposing effective measures to improve the efficiency of knowledge dissemination [7]. Pan Mengya et al. pointed out that user classification can promote the healthy development of online Q&A communities. By identifying professional responders with a high likelihood of providing answers to specific questions, it can help questioners obtain high-quality responses and shorten the waiting time for satisfactory answers, thus promoting the continuous and healthy development of the community [8]. Zhong Qing, in his research, also emphasized the importance of behavior classification of mobile Internet users based on preference tags, suggesting that precise content push can enhance content acceptance, thereby increasing commercial value [9].

However, there are still some shortcomings in the existing literature on the classification of user groups in online Q&A communities. For instance, previous studies have largely focused on macro-level statistical analyses and failed to adequately reveal the micro-level characteristics and evolutionary processes of user behavior. Moreover, the influence of psychological and social factors on the differential behaviors of different user groups has not been fully explored, resulting in insufficient understanding of the intrinsic motivational mechanisms for users to engage in knowledge sharing. Additionally, the operability and accuracy of the user classification models used are often limited. In response, this paper employs natural language processing techniques to classify users based on language features, addressing the limitations of previous studies that relied on digital feature indices such as follower count and click volume for user classification. By constructing an analytical framework encompassing three dimensions—information entropy, question-answer relevance, and sentiment polarity, this study integrates management theory with quantitative analysis to propose more practical user classification methods and incentive mechanism optimization strategies, aiming to promote the healthy development of online Q&A communities and enhance knowledge-sharing efficiency.

2. Research Methods and Data Analysis and Processing

2.1. Study design

This study employs web crawler technology to obtain and collect personal information and answer information data of the user's homepage from Zhihu. Following data collection, we utilize quantitative methods to analyze the users' answer text based on three dimensions: information entropy, question-answer correlation, and emotional polarity. This comprehensive analysis yields a multidimensional dataset for each user, which is subsequently categorized and examined using clustering algorithms. Additionally, the user quality coefficient was calculated from the user's personal information. To investigate the relationships between this quality coefficient and three control variables pertinent to Zhihu’s official certification and further explore the characteristics of each class of user, the regression analysis was carried out. Ultimately, the findings from this analysis will inform recommendations and strategies for enhancing the online question-and-answer community.

2.2. Data Collection

Zhihu, as a representative online Q&A community in the Simplified Chinese Internet, had over 220 million users by 2023, with 99 million active users. This demonstrates the significant role that online Q&A communities, especially for younger demographics, play in knowledge acquisition in contemporary society.

Therefore, this paper utilizes Python web scraping techniques to obtain user data from the Zhihu platform. The basic approach for data acquisition is to select popular and comprehensive topic tags from the community and then gather relevant information from users who participated in answering within these topics. The data is primarily categorized into two levels: the personal information of users and all the answers listed on their user profiles.

Personal-related information includes metrics such as the number of followers, the number of likes, the number of saves, the number of public edits, the number of answers recorded by Zhihu, the presence of certification information, and the status of being a Zhihu judge. For the last two attributes, binary coding is employed, with '1' denoting "yes" and '0' indicating "no." The answer-related information includes the user's answer content, answer release time, question title, question content, and question tag.

In order to facilitate the collection and aggregation of information, the selected topic should be group-oriented, comprehensive, and rich in the answers covered, while also being relevant to the knowledge-sharing behavior. After comprehensively considering the hot topics in various categories of Zhihu, this study focuses on the topic of 'university', which boasts approximately 40 billion views and 28.62 million discussions. To ensure that the volume of collected data is appropriate, users are filtered based on the number of answers they have provided, ensuring that the total number of answers falls within a specified range. Finally, the user's personal information is quantitatively analyzed and processed after the list of the user's answer information system.

As of 2024, a total of 869 users, each contributing between 20 and 250 answers, were selected under the subject of "university" following this screening process. Ultimately, the data collection yielded over 87,000 records of user-related data and corresponding answer content.

2.3. Three-dimensional text processing of answers

2.3.1 Three-dimensional introduction

This paper analyzes the user's response behavior by performing natural language processing techniques to assess the textual content of user answers across three dimensions: the information entropy of the user's answer, the relevance of the user's answer to the question, and the emotional polarity of the user's answer. For each user, the text of each answer is processed and quantized into numerical values along these three dimensions. Subsequently, the average of all the answer values of the user in each dimension is calculated to obtain the data of the user in the three dimensions.

2.3.2 Information entropy calculation

In terms of the selection of dimensional information entropy, this study employs one-dimensional information entropy to process the user's answer text. This approach is justified by the observation that one-dimensional information typically generates higher entropy values than high-dimensional information entropy, thereby enhancing the user distinction.

In this study, the user's answer texts undergo a preprocessing phase in which all non-Chinese characters, as well as spaces, line breaks, and tabs, were removed from the text through regular matching. Subsequently, the jieba library of Python was utilized for text segmentation, dividing the cleaned answer text into a list of individual words. The frequency of each word's occurrence within the text is then calculated to derive the word frequency distribution.

To compute the unary entropy value of the answer text, the relative frequency of each word is multiplied by its corresponding self-information value, and the negative values are summed. This process culminates in the determination of the unary entropy for the analyzed texts.

$$H = -\sum_i P(\omega_i) \cdot log_2 P(\omega_i) \quad (1)$$

where $P(\omega_i)$ represents the relative frequency of the $i-th$ word, which is calculated by dividing the word frequency by the total number of words. The greater the information entropy, the higher the uncertainty of the text and the greater the amount of information it carries.

2.3.3 Calculation of the relevance of the answer to the question

In the dimension pertaining to the correlation between answers and questions, this study quantifies the correlation between the question and answer by calculating the cosine similarity between the question word vector and the answer word vector.

For the problem word vector, it is derived from the integration of three fields. The crawled question label field serves as a pre-existing word vector, which is directly incorporated into the question word vector. Additionally, the question title and question content fields are concatenated into a single text, from which keywords are extracted using the TextRank algorithm and added to the question word vector. For the answer word vector, the TextRank algorithm is similarly employed to extract the keywords from the answer text, with a maximum limit of ten keywords to define the answer word vector.

The main computational process of the TextRank algorithm is as follows: first, the input text is segmented into words and stop words as well as irrelevant terms are removed. Next, a co-occurrence graph is constructed, representing the text words as an undirected graph $G = (V, E)$, where V denotes the set of words and E represents the co-linear relationships between them. An edge is established between two words if they co-occur within a fixed-size sliding window, and the weight of this edge is assigned based on the frequency of co-occurrences. Finally, node weights are computed by assigning an initial score to each word, which is subsequently updated through an iterative propagation process.

$$S(V_i) = (1-d) + d \sum_{V_j \epsilon In(V_i)} \frac{\omega_{ji}}{\Sigma_{V_k \epsilon Out(V_j)} \omega_{jk}} S(V_j) \tag{2}$$

After multiple iterations, the formula gradually converges, allowing the nodes to be ranked according to their scores from highest to lowest. The top-scoring words are selected as keywords.

Subsequently, vocabulary is constructed based on the question word vector and the answer word vector. One-Hot encoding is applied to both vectors, and the cosine similarity of the encoded word vectors is then calculated to measure the relevance between the question and the answer.

One-Hot Encoding:

$$one\ hot\ question[i] = f(x) = \begin{cases} 1, & if\ i = question\ code[j]\ for\ \ some\ j \\ 0, & otherwise \end{cases} \tag{3}$$

$$one\ hot\ answer[i] = \begin{cases} 1, & if\ i = answer\ code[j]\ for\ some\ j \\ 0, & otherwise \end{cases} \tag{4}$$

Cosine similarity calculation:

$$cosine\ similarity = \frac{\sum_{i=1}^{n} one\ hot\ question[i] \times one\ hot\ answer[i]}{\sqrt{\sum_{i=1}^{n} one\ hot\ question[i]^2} \times \sqrt{\sum_{i=1}^{n} one\ hot\ answer[i]^2}} \tag{5}$$

2.3.4 Affective Polarity Calculations

In the dimension of sentiment polarity, this study calculates the sum of the scores of the key sentiment words in the user's answer text using a sentiment dictionary. This total score is then divided by the total number of words in the answer text to obtain the sentiment polarity of the text.

For the selection of sentiment dictionaries, this study utilizes four specific dictionaries to compute user sentiment values: the sentiment word dictionary, the negation word dictionary, the degree adverb dictionary, and the stop word dictionary. The sentiment word dictionary employed is the BosonNLP Sentiment Dictionary, which includes a vast array of sentiment words from the internet and provides scientifically assessed sentiment scores, making it particularly suitable for analyzing Chinese text in an online environment.

Initially, the response text undergoes preprocessing to remove non-Chinese symbols and stop words. The sentiment polarity is then calculated using the following formula:

$$score = \frac{1}{T}\sum_{i=1}^{n} W_i \cdot s_i \tag{6}$$

Where s_i represents the sentiment value of the $i-th$ sentiment word, W_i is the weight factor of the $i-th$ sentiment word, which depends on the number of negation words and degree adverbs preceding it, and T denotes the total word count.

The weight factor W_i is defined as:

$$W_i = (-1)^N \times \prod_{j=1}^{M} d_j \tag{7}$$

where N indicates the number of negation words preceding the current sentiment word, and d_j represents the weight adjustment factor for each degree adverb, with the product taken over all degree adverbs j.

2.3.5 Classify users through cluster analysis

After obtaining numerical indicators of user response behavior across different dimensions, this study employs the $k-means$ algorithm for three-dimensional cluster analysis. Users are categorized into four classes based on their performance indicators.

The clustering algorithm begins by randomly selecting K data points as initial centroids, followed by iterative calculations of the Euclidean distance between data points and centroids. Data points are assigned to the nearest cluster, and the centroid positions for each cluster are recalculated. This process of "assigning clusters" and "updating centroids" is repeated until the centroids stabilize or a predetermined number of iterations is reached.

The Euclidean distance is calculated as follows:

$$d(x_i, \mu_k) = \sqrt{\sum_{j=1}^{n}\left(x_{ij} - \mu_{kj}\right)^2} \tag{8}$$

The centroid update and objective function are defined as:

$$\mu_k = \frac{1}{|C_k|}\sum_{x_i \in C_k} x_i\,,\ J = \sum_{k=1}^{K}\sum_{x_i \in C_k}\|x_i - \mu_k\|^2 \tag{9}$$

Through the k-means algorithm, Figure 1 was obtained, where data points formed four clusters across three dimensions, with a relatively even distribution of data points within each cluster. To ensure the independence of the three dimensions, this study computed the correlation among the three sets of data, yielding a correlation coefficient matrix. The correlation coefficients between any two dimensions were all less than 0.3, indicating minimal correlation among the selected dimensions, thereby validating the appropriateness of the dimension selection.

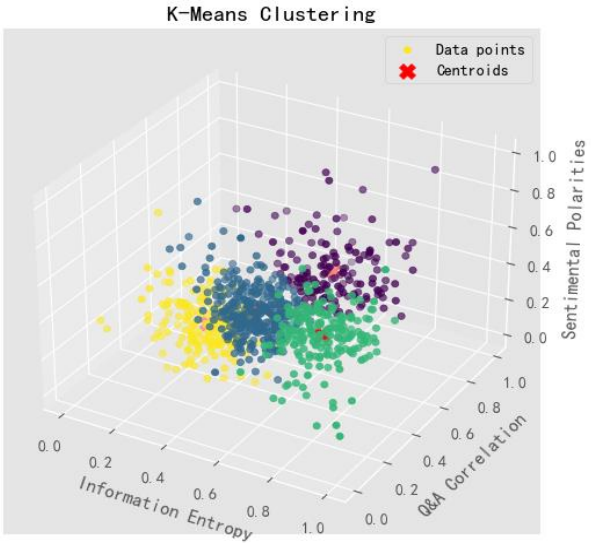

Figure 1. Results of Cluster Analysis

2.4. User mass coefficient calculation

In order to better derive the contributions of different categories of users to the Creative Commons community, this paper classifies the public information data of users' homepages. Previous studies have shown that the online prestige and communication effect of knowledge disseminators promote each other, and the similarity of actions has a positive impact on the online prestige, while the richness of knowledge content has a significant effect on the communication effect [10]. Therefore, this paper calculates a comprehensive quality coefficient based on the interaction behavior data of users and other users, which is used to determine the knowledge-sharing value of users. Additionally, since Zhihu has established corresponding honors and reward mechanisms for users with high knowledge-sharing value, this study also includes the official honors and incentive evaluations from Zhihu as control variables, alongside the quality coefficient, as indicators for assessing user value.

The quality coefficient is derived from four parameters: the number of followers F, the number of likes L, the number of saves S, and the number of public edits E. To avoid mathematical errors associated with logarithmic calculations, the number '1' should first be added to each of these four indicators before taking the logarithm.

$$F' = log(F+1), L' = log(L+1), S' = log(S+1), E' = log(E+1) \tag{10}$$

Subsequently, these four processed indicators are normalized using the following standardization formula to unify their scales to the range [0, 1]:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{11}$$

Then, the normalized values of the four parameters are summed up to obtain a comprehensive score. This comprehensive score is then divided by the user's active duration (i.e., crawl time, minus the time the user first posted an answer, and converted to the number of years with decimals) to reflect the user's quality output per unit time. This method effectively accounts for user behavior characteristics while mitigating the bias introduced by user activity duration, enhancing the objectivity and comparability of the results.

$$Q = \frac{F'' + L'' + S'' + E''}{T} \tag{12}$$

Subsequently, this study visualizes the distribution of quality coefficients for all users, as shown in Figure 2, which reflects the distribution after removing interference. Analyzing the histogram of the quality coefficient distribution in Figure 2, along with the fitted normal distribution curve, shows that the quality coefficients approximate a normal distribution. This observation suggests that the algorithm effectively mitigates differences in magnitude between different indicators during data processing, while also reducing the impact of outliers on the overall distribution. As a result, this ensures the validity and robustness of the results obtained.

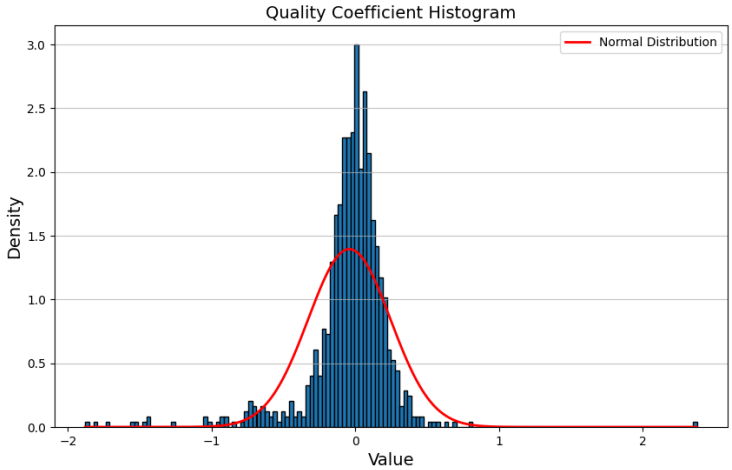

Figure 2. Distribution of Mass Coefficients

2.5. Regression analysis

To better ascertain the specific impact factors of the aforementioned four categories of users on knowledge-sharing platforms, this study further conducts a regression analysis focusing on these user categories and their corresponding quality coefficients. The regression analysis primarily aims to compare the overall quality levels among the different user categories identified through the clustering algorithm. Additionally, it allows for the examination of the relationships between three control variables and the overall quality coefficients of the users.

The control variables comprise three types of data: the number of answers recorded by Zhihu, the presence of certification information, and the status of being a Zhihu judge. The latter two variables are represented as 1 for "yes" and 0 for "no."

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 Z_1 + \beta_5 Z_2 + \beta_6 Z_3 \tag{13}$$

where the dependent variable y represents the quality coefficient, and the independent variables x_1, x_2, x_3 correspond to the control variables: the number of answers recorded by Zhihu, the presence of certification information, and the status of being a Zhihu judge, respectively. The independent variables Z_1, Z_2, Z_3 indicate the user class identifiers, with values as outlined in the following Table1:

Table.1. User Classification Correspondence in Regression Analysis

User in Class	Z1	Z2	Z3
0	0	0	0
1	1	0	0
2	0	1	0
3	0	0	1

Thus, by observing the coefficients of Z_1, Z_2, Z_3, one can determine the relative quality of users in classes 1, 2, and 3 compared to users in class 0.

The results of the regression analysis for the aforementioned variables are as follows:

$$R - squared = 0.140 \tag{14}$$

$$Adj.R - squared = 0.134 \tag{15}$$

$$F - statistic = 23.37 \tag{16}$$

$$Prob(F - statistic) = 1.15 \times 10^{-25} \tag{17}$$

After sorting and summarizing, Table 2 is obtained.

Table.2. Results of Regression Analysis

Variable	Coefficient	Standard error	t	P > |t|
x_1	0.009	0.005	1.777	0.076
x_2	0.044	0.028	1.562	0.119
x_3	0.144	0.027	5.370	0.000
Z_1	0.123	0.025	4.996	0.000
Z_2	-0.102	0.024	-4.239	0.000
Z_3	-0.005	0.027	-0.202	0.840
Intercept	-0.073	0.016	-4.658	0.000

From the above results, it is evident that there is a significant positive correlation between being a Zhihu judge and a high-quality coefficient, while the number of answers recorded by Zhihu and the presence of certification information do not exhibit a significant relationship with the quality coefficient.

Moreover, users in class 1 demonstrate significantly higher quality than those in class 0, while the quality of users in class 3 is comparable to that of class 0 users. Conversely, users in class 2 exhibit a significantly lower quality than those in class 0.

3. Results

During the clustering process, we marked the centroid of each cluster. Since k-means clustering is based on mean maintenance, the coordinates of the cluster center represent the average values of the users in that cluster across three dimensions. The following results were obtained and recorded in Table3:

Table.3. Three-Dimensional Performance of the Four User Types

Class	Information Entropy	Q&A Correlation	Sentimental Polarities
0	0.498	0.338	0.413
1	0.771	0.309	0.408
2	0.334	0.315	0.294
3	0.665	0.568	0.536

As shown in the table, the users in cluster 0 have answer content characterized by average levels of information entropy, question-answer relevance, and sentiment polarity. Based on these characteristics, they are labeled as ‘Routine Answerers’.

Users in cluster 1 have the highest information entropy in their responses, with weak relevance to the question and a neutral sentiment. This suggests that their answers primarily involve rational explanations, knowledge statements, and information sharing, leading to their classification as ‘Knowledge-driven Answerers’.

Cluster 2 users have the lowest information entropy, weakest question-answer relevance, and negative sentiment. These users tend to provide simple, short answers with a negative emotional tone and little engagement with the question. They can therefore be classified as ‘Indifferent Answerers’.

Users in cluster 3, on the other hand, are characterized by high information density, strong relevance to the question, and a positive emotional tone. These users are typically proactive, friendly, and helpful in answering questions or providing detailed content. Consequently, they are labeled as ‘Responsive Answerers’.

After identifying the four user categories, we further associated the personal value scores of each user type with their corresponding categories, as shown in Table 4:

Table.4. Personal Value of the Four User Types

User Type	Quality Coefficient	Certification Information	Recorded Answers	Zhihu Judges
Routine Answerers	-0.0480	0.1079	0.1143	0.1333
Knowledge-driven Answerers	0.0993	0.1378	0.2704	0.2857
Indifferent Answerers	-0.1701	0.0392	0.0000	0.0245
Responsive Answerers	-0.0406	0.2727	0.7013	0.1364

Based on these results, we can summarize that, on the online Q&A platform, Knowledge-driven Answerers have a higher quality score than Routine and Responsive Answerers, with Indifferent Answerers scoring the lowest. Among these, Routine Answerers have slightly higher quality than Responsive Answerers.

The possible reason for this result lies in the fact that users with higher information entropy, more neutral sentiment, and greater objectivity in their answers tend to share more information, which better aligns with the needs of knowledge sharing. These users are more likely to attract traffic and attention, thereby increasing their follower count. Hence, Knowledge-driven Answerers have a higher

quality score. Correspondingly, their average number of followers is the highest, and their answer record rate and certification details are also slightly higher than those of Routine Answerers.

Responsive Answerers, while focused on answering the questions with kind help and support, tend to provide a large volume of information. However, their quality score is almost the same as that of Routine Answerers. This may be because their responses are highly targeted to the questioner, relatively specific and hardcore, and do not offer broad help to other viewers, leading to similar levels of likes, saves, and follows as Routine Answerers, with relatively modest traffic. Control variables show that these users have a much higher answer record rate than the other three categories, confirming their Responsive Answerer traits. Additionally, they have the highest certification rate among the four types, indicating that Responsive Answerers are often senior professionals or experts in their respective fields.

Finally, Indifferent Answerers exhibit the lowest scores across all three dimensions, indicating that their answers are short, indifferent, and lack relevance to the questions, often containing negative emotional expressions. As a result, their quality score is the lowest, along with the lowest levels of certification, recorded answers, and followers. It is clear that this user group is least favored by both the community and Zhihu's official platform.

This result reveals the distribution characteristics and potential issues within the Zhihu platform user base. The community should consider developing relevant policies and algorithm improvements to incentivize and address these issues effectively (Table 5).

Table.5. Final Results of User Classification

User Classifications	Cluster Identifier	Knowledge Sharing Quality	characteristics
Routine Answerers	0	Medium	Generic response
Knowledge-driven Answerers	1	High	Informative responses, knowledge sharing
Indifferent Answerers	2	Low	Concise and indifferent responses, expressing negative emotions
Responsive Answerers	3	Medium	Friendly and in-depth responses, typically from professionals

4. Improvement Strategies

As the largest online Q&A community in the Chinese Internet space, Zhihu's role in knowledge sharing is unquestionable. However, with the changing Internet environment and culture, along with the increasing number of users, certain negative effects have emerged, such as one-sided emotional outbursts and the proliferation of short, ineffective answers, which have somewhat undermined the community environment and knowledge sharing efficiency. This section, based on the earlier user classification results and the evaluation of user knowledge-sharing value, proposes a series of improvement strategies for the Zhihu community, starting from the four user categories, in order to enhance knowledge-sharing efficiency and improve the Q&A community.

For Knowledge-driven Answerers, the platform should encourage these high-quality users to participate more actively in answering questions and increase their influence within the community. Although Zhihu's current incentive and support mechanisms are relatively well-established, there are still some biases in identifying and recognizing high-quality answers. Therefore, Zhihu can incorporate factors such as information entropy, question-answer correlation, and sentiment polarity into the platform's recommendation and incentive algorithms. By prioritizing answers that meet the criteria of Knowledge-driven Answerers, these answers can be pushed to a larger audience, thereby increasing their exposure and impact.

For Responsive Answerers, the platform should provide more exposure and push for their answers. These users already have higher certification rates and more recorded answers, indicating that Zhihu has recognized their identity and affirmed the quality of their answers. Therefore, the platform could

further improve its recommendation algorithms to give more exposure to these high-quality answers with relatively low traffic, ensuring that their valuable contributions reach a broader audience.

For Indifferent Answerers, the platform should guide them toward more rational expressions. Zhihu can optimize its answer review mechanism by enhancing sentiment polarity detection algorithms to closely monitor and reduce answers with strong emotional tones. This would help users incorporate more knowledge-based and constructive information into their responses, reducing the likelihood of meaningless emotional disputes.

For Routine Answerers, fostering a more friendly community culture to promote knowledge dissemination is crucial. The platform can increase activities such as upvote guides and knowledge-sharing initiatives, encouraging ordinary users to read and learn from the answering styles, structure, and logic of Knowledge-driven Answerers, thus improving the overall content quality of the platform.

Additionally, Zhihu could strengthen the regular analysis and feedback of user behavior data. By mining data and periodically analyzing the answering characteristics and performance of different user types, the platform can assess the effectiveness of its incentive mechanisms and recommendation algorithms, ensuring that it continues to develop towards a high-quality knowledge-sharing model.

5. Conclusion and Outlook

This study combines the characteristics of knowledge-sharing behavior in online Q&A communities, using Zhihu as an empirical case. The analysis of user responses is conducted from three dimensions: information entropy, question-answer correlation, and sentiment polarity. By integrating natural language processing and clustering analysis, users are classified into four categories: Controversial Answerers, Routine Answerers, Knowledge-driven Answerers, and Responsive Answerers.

The results of the study indicate that Knowledge-driven Answerers outperform both Routine and Responsive Answerers and also have higher quality than Indifferent Answerers. Knowledge-driven Answerers, due to the richness of their answers and neutral sentiment, fulfill the requirements of knowledge sharing, thereby attracting more attention and followers. As a result, they have the highest quality score and the most followers. Responsive Answerers show a quality score similar to that of Routine Answerers. They focus more on specific issues, providing helpful or detailed responses, and share a substantial amount of information. Indifferent Answerers, however, provide short and emotionally negative responses, resulting in the lowest quality score and difficulty in gaining user approval.

Based on the aforementioned results, this paper proposes a series of algorithmic improvements for the community, taking into account the characteristics of the four user types. These suggestions aim to enhance the knowledge-sharing environment of the community by addressing aspects such as message pushing, user incentives, and speech guidance, thereby fostering a stronger knowledge-sharing atmosphere.

During the data collection process, the study was limited by a lack of better entry points and only selected a single topic, University', from the Zhihu platform. As a result, the findings may not fully represent or cover the overall user situation of the Zhihu platform. Moreover, the paper proposed a series of algorithmic improvements to enhance the knowledge-sharing efficiency of the Q&A platform. Therefore, future research on this issue could expand the scope of the study, select more precise dimensions for analyzing answer texts, and explore the practical impact and effects of algorithmic improvements.

References

[1] Lv, M., Li, Q. Study on the influencing factors of knowledge interaction behavior of social Q&A users. Information Science [J]. 2022, 40(06), 141-148+193.

[2] Hadfi, R., Moustafa, A., Yoshino, K., et al. Best-answer prediction in Q&A sites using user information. Cornell University Library [J]. 2022, arXiv.org, Ithaca.

[3] Sher, N., Rafaeli, S. Associative linking for collaborative thinking: Self-organization of content in online Q&A communities via user-generated links. PloS One [J]. 2024, 19(3), e0300179.

[4] Nasirian, S., Nogara, G., Giordano, S. Not My Fault: Studying the Necessity of the User Classification & Employment of Fine-Level User-Based Moderation Interventions in Social Networks[C], 2024.

[5] Shahbaznezhad, H., Dolan, R., Rashidirad, M. The role of social media content format and platform in users' engagement behavior. Journal of Interactive Marketing, [J]. 2021, 53(1), 47-65.

[6] Zhang, J.T., & Chu, W.C. User classification in social Q&A communities from the perspective of knowledge sharing: A case study of Zhihu. Shanghai Journal of Management Science[J]. 2020, 42(05), 30-37.

[7] Angeletou, S., Rowe, M., Alani, H. Modelling and analysis of user behaviour in online communities. In L. Aroyo et al. (Eds.), Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) [C].2011, 35-50.

[8] Pan, M.Y., Shen, W., Dai, W., et al. Research on the discovery of question-answerers in social Q&A communities. Library and Information Work [J]. 2020, 64(18), 76-88.

[9] Zhong, Q. A study on user behavior classification in mobile Internet based on preference tags. Mobile Communications [J]. 2016, 40(09), 93-96.

[10] Chen, X.H., Hu, P., Ma, Y.T. Network reputation, action similarity, and knowledge dissemination effects of knowledge disseminators in Q&A communities. Journal of Management [J]. 2021, 18(05), 741-750.