2022

Predicting the Way and the Degree of Users' Content Contribution in the Social Question and Answer Community

Yuting Jiang
Aalto University, yuting.jiang@aalto.fi

PREDICTING THE WAY AND THE DEGREE OF USERS' CONTENT CONTRIBUTION IN SOCIAL QUESTION AND ANSWER COMMUNITY

Research paper

Yuting, Jiang, Aalto University School of Business, Espoo, Finland, yuting.jiang@aalto.fi

Abstract

Most predictions of user behavior occur after a user has participated in the community for a while, and those who have just registered are easily overlooked because their community characteristics have not yet been revealed. However, users are easy to be lost in the early stage. Based on the theory of social capital, this paper proposes a new approach to predict the willingness, mode, and degree of content contribution of the newly registered user based on users' information disclosure behavior aiming at reducing the churn rate of newly registered users. We crawled the data of 4 million users in the Zhihu community and deeply studied the relationship between the disclosure behavior of different types of information and the content contribution degree of users through statistical analysis methods and machine learning algorithms. The result shows that if a user discloses personal information, the probability of his in-depth response contribution and in-depth questioning contribution will increase correspondingly, and different types of information disclosure will lead to a different probability of an increase. Furthermore, In addition, users' disclosure of different types of information will lead to differences in their preference for the way they contribute content.

Keywords: Social Question and Answer Community, Self-Disclosure Behavior, Content Contribution behavior, Predicting Analysis.

1 Introduction

People are increasingly seeking answers and knowledge in the social question and answer (SQA) community. In order to ensure question response rate and response quality, the question-answering matching mechanism is used to invite specific users to answer the corresponding questions in the SQA community. Identifying potential answerers will increase the probability that a question to be answered (Le and Chirag, 2018). Since questions are also a key component of the SQA community, high-quality questions can guarantee the quality of answers to a certain extent (Agichtein, Castillo, Donato, Gionis, and Mishne, 2008), so the questioner is also a key component of the SQA community. Identifying these content contributors (Agichtein et al., 2008), including questioners and answerers, are very important to the SQA community.

However, a relatively small number of content contributors (Lampe, Wash, Velasquez, and Ozkaya, 2010) and the growing quantity of registered users bring greater challenges in identifying content contributors. For example, the number of registered users in Zhihu which is the most popular SQA website in China has exceeded 200 million in 2018 (Techweb, 2018), but such a large group of new users without activity data is rarely studied. Analysis of previous studies on identifying potential answerers mainly draws on the active data of users, such as the theme similarity or content similarity (Le and Chirag, 2018). However, there is still a lack of relevant research on users who have no active data at present, such as new coming users.

User churn is most likely to occur in the early use of users (Milošević, Živić, and Andjelković, 2017). Online communities provide personalized recommendations by predicting new registered users' willingness to contribute to content, ways of contributing, and the extent of contributing content, which

can effectively reduce the loss rate of early users. And this group of users may disclose some personal information when they register even if they have no activity data. Therefore, we propose a new method to identify potential content contributors based on users ' information disclosure behavior.

When users register on social networking sites for the first time, they can freely choose whether to disclose some personal information. Without a separate privacy setting, most of the information a user discloses can be seen by other users on social networking sites (Van Gool, Van Ouytsel, Ponnet, and Walrave, 2015). Personal disclosure behavior of users is proved to be related to the content contribution behavior. Studies have shown that users' concern about privacy will significantly affect users' willingness to contribute knowledge and users' willingness to seek knowledge (LIU, 2013). Secondly, the disclosure of personal information will enhance users' perceived authentication and promote their knowledge contribution behavior (Meng and Agarwal, 2007). Since these studies all show that users' disclosure of personal information has a significant impact on users' content contribution behavior, we believe that it is a feasible attempt to take users' information disclosure behavior as a predictive factor of users' content contribution behavior.

Previous studies are mostly based on the data of the questionnaire survey, which has certain limitations. First, the small sample size will have a certain impact on the measurement. Secondly, the researchers did not verify the accuracy of the information provided by users on social websites. In addition, these studies only show that information disclosure has a significant impact on content contribution behavior. However, the difference of users' disclosure behavior to different types of personal information, and the relationship between the way and the extent of users' contribution to content and users' disclosure behavior have not been studied.

To fill the gap, we construct a prediction model of user content contribution degree based on the user's personal information disclosure behavior. Then, the empirical analysis was conducted on the data of 4 million users of Zhihu, the largest social question and answer (SQA) community in China. Combined with statistical analysis and machine learning method, the relationship between the user's personal information disclosure behavior and the way and degree of content contribution is deeply discussed in this article, aiming at excavating the early potential content contribution of users and improving the effectiveness of question and answer matching mechanism. We explored the following three research questions.

1. To study what user disclosure and the characteristics of personal information disclosed by users.

2. To explore the difference of performance on information disclosure behavior of users with different content contribution levels

3. To predict the way and the level of user's content contribution according to their self-disclosure behavior.

2 Theoretical Background

2.1 Self-disclosure behavior of social media users

Self-disclosure is the process of making yourself known to others (Jourard and Lasakow, 1958). In the context of social networking sites, self-disclosure can be defined as the amount of information shared in users' personal data (Krasnova and Veltri, 2011). It's worth noting that if there is no separate privacy setting, most of the information disclosed by users can be seen by other users (Van Gool et al., 2015). Personal information that can be disclosed on social networking sites includes name, address, date and year of birth, contact information, and photos (Nosko, Wood, and Molema, 2010).

Privacy and security issues related to personal information disclosure are constantly mentioned (Bansal, Zahedi, and Gefen, 2016; Benndorf, Kübler, and Normann, 2015; Benson, Saridakis, and Tennakoon, 2015). Nevertheless, the benefits of disclosing personal information on social networks have also been extensively studied. Disclosure serves two purposes. First, the information disclosed can be used by the platform or other users to identify and authenticate you, or by the other users to discover previously undetected features (Gross, Acquisti, and Heinz, 2005). For example, Batenburg et al. (2017) found that

sharing personal information with other employees on social media contributes to the establishment of a professional image and the supplement of professional traits, thereby enhancing their likeability and respect (Batenburg and Bartels, 2017). In addition, personal information disclosure is of great significance to the development of online relationships. For teenagers, the establishment of peer relationships and the exploration of self-identity are the reasons for their willingness to disclose personal information on social networks. The benefits of disclosure of personal information far outweigh the perceived risk of this behavior (Bryce and Fraser, 2014). In online learning communities, personal information disclosure (personal data and photos) is beneficial for some online learners to interact with each other (Kear, Chetwynd, and Jefferis, 2014).

Information disclosure behavior of online users may be affected by many factors, for example, emotions (Wang et al., 2011) affect users' information disclosure behavior. Some studies predicted users' self-disclosure behavior. Xie et al. (2015) used hierarchical regression analysis to predict data from Teens and Privacy Management Survey. It was found that the higher the frequency of social network use, the more social network friends, and the higher the degree of social network information disclosure. In addition, gender and age are the two most important demographic factors predicting self-disclosure behavior in social networks (Xie and Kang, 2015).

There are also some inquiries about the degree of user privacy disclosure. Kisekka (2013) explored the three categories of users with the lowest degree of privacy disclosure on Facebook: women, users accessing accounts using smart phones, and users with multiple accounts (Kisekka et al., 2013). Gender differences in self-disclosure were also studied (Barak and Gluck-Ofri, 2007). Special (2012) found that men were willing to disclose more personal information (Special and Li-Barber, 2012).

Researchers have made some predictive studies on self-disclosure behavior, Information disclosure is largely a predictor of user popularity (Christofides, Muise, and Desmarais, 2009).

2.2 User content contribution in the SQA community

Some researchers have studied the motivation of users to contribute content. Exposure of identity on social media, sharing advertising revenue with platforms, and desire for reputation are the main drivers of users' content contribution (Tang, Gu, and Whinston, 2012). An empirical study of Wikipedia by Xu et al. (2015) found that users have two kinds of behaviors: participation behavior and content contribution behavior. The research shows that users' participation behavior in the community is mainly influenced by internal driving factors such as altruism and a sense of belonging to the community, while users' content contribution is mainly influenced by external factors such as reciprocity and self-development needs (Xu and Li, 2015). Lampe et al. (2010) found five main motivations for users to contribute content: purposeful value, such as giving or receiving information. Self-discovery, such as acquiring social resources and self-knowledge. Maintaining interpersonal relationships, such as obtaining social support and friendship. Social enhancement is related to a value derived from users' state in the community, such as the best answerer, opinion leader, etc. And entertainment, which is the pleasure and relaxation of interacting with other users (Lampe et al., 2010).

2.3 Social capital theory

Some studies have shown that trust directly promotes user’s contribution behavior. From the perspective of social capital theory, the establishment of a trust relationship becomes an important condition for the generation of social capital to some extent. Trust between members is a private asset for individuals and a public asset for communities. Trust will significantly affect the knowledge-sharing behavior of community members (Chang and Chuang, 2011; Chiu, Hsu, and Wang, 2006). The type and amount of personal information disclosed indicates the degree of trust people have in online communities (He and Wei, 2009).

Offline privacy preference will affect users' online behavior (Spiekermann, Grossklags, and Berendt, 2001), and researchers have also explored the relationship between the two. For example, Dwyer et al. (2007) explored the influence of users' privacy concerns on users' information sharing behavior based

on a questionnaire survey and conducted a comparative study on Facebook and MySpace platforms (Dwyer et al., 2007). This study has some limitations. First, small sample size will have some influence on the measurement of correlation. Second, researchers did not go to social networking sites to verify the accuracy of information provided by users (Dwyer et al., 2007). Liu used the structural equation method to analyze the data collected through the questionnaire survey and found that users' emotional trust in the community and users' concern about privacy would significantly affect users' willingness to contribute knowledge and users' willingness to seek knowledge (LIU, 2013). Meng and Agarwal (2007) found that disclosure of personal information would enhance users' perceived authentication and promote their knowledge contribution behavior (Meng and Agarwal, 2007). The data used in the research of this kind of problem are mainly collected by questionnaire survey. The most common method is to judge whether the relationship between the two factors is significant or not through a structural equation.

However, the research on whether the disclosure preference of different types of personal information affects the way and the degree that users contribute to content has not been involved.

In addition, users have different attitudes toward various personal information disclosure behavior (Shin and Biocca, 2018). Different types of personal information may lead to users' different willingness to provide information and different ways of providing information (Bansal and Gefen, 2010; Shin and Biocca, 2018). However, the research on how the disclosure preference of different types of personal information affects the way and the degree that users contribute to content has not been involved.

3 Method

This paper mainly conducts the following three studies. First, we studied the changing relationship between individual privacy disclosure behavior and user's content contribution degree. Second, we verified whether user personal information disclosure behavior is related to user content contribution behavior or not through cross-analysis. Finally, we build a Multinomial Softmax Regression classifier to predict users' content contribution type and degree according to their privacy disclosure behavior.

We crawled 4 million users' information data and content contribution data on Zhihu, the largest SQA Chinese website, and conducted binary classification of the user's personal information according to whether the user disclosed it or not. As for the content contribution data, we calculated the mean of the number of questions and answers and divided the number of questions and answers into three degrees according to the mean value. On this basis, we conducted a pairwise combination to form nine content contribution types.

3.1 Data collection

This paper chooses the Zhihu website as our research object. Zhihu is a quora-like community in China, and it is the largest Chinese SQA community. In 2018, the number of registered users of the Zhihu community reached 180 million. The website USES artificial intelligence to quickly match interesting questions for potential respondents. We choose this website for two reasons. First, the large amount of user-generated data on Zhihu provides a sufficient and reliable basis for our research. Secondly, as the largest Chinese SQA community, Zhihu has a large user base, and they have sufficient needs to identify potential content contribution users in the user base to conduct more accurate question and answer matching.

In the Zhihu community, the personal information that users can disclose includes gender, location, education background, and employment experience, so these four indicators are used as indicators of users' personal information disclosure behavior. The content contribution index of users mainly includes questions and answers. This paper collected the above six data of 4 million users in the Zhihu community through a python web crawler, including users' personal information data and users' content contribution data.

We use the social relationships of the users in the Zhihu community to recursively crawl user information. Start from an opinion leader with a large number of followees and followers, and then start

from each user in the list, climb the user's followees and followers list, and recursively climb until no new users are crawled. Under such a crawling strategy, we can get nearly all the users in the social network in the Zhihu community (that is, those users who have followed others or been followed by others), and those users who do not exist in the social network are not considered in this paper.

3.2 Data processing

First, we conducted the binary classification on the self-disclosing behavior dataset. There are four categories of personal information that users can disclose in the Zhihu community: gender, location, education experience, and employment experience. For certain personal privacy information, there are two behaviors of disclosure and non-disclosure. We set the disclosure of the personal information to be 1 and the non-disclosure of the personal information to be 0.

We then divided the degree of content contribution. In the SQA community, there are two main content contribution behaviors, questioning and answering (Wang and Zhang, 2016). The number of questions and answers varies among users. To rank-order the degree of user's contribution to the content, we calculated the mean of user's questions and answers as a threshold of contribution (Seo, Lim, and Choi, 2014), and the other threshold is 0. Therefore, we divide the contribution degree of the questions into three degrees according to the 0 value and the mean value of the questions. Similarly, the contribution degree of answers were divided into three degrees according to the 0 value and the mean value of answers. The pairwise combinatorial method is used to form the nine degrees of user's contribution.

We calculated the mean value of questions (mean=0.82) and the mean value of answers (mean=6.24) for more than 4 million users. Since both the question count and answer count are integers, decimals are rounded up and rounded down to 1 and 6 respectively. As a result, the degree of contribution to the answer can be composed of three degrees. When the answer count is equal to 0, it belongs to the first degree, the second degree is between 1 and 6, and the third degree represents the answer count is greater than 6. Similarly, when the question count is equal to 0, it is the first degree, when the question count is equal to 1, it is the second degree, and if the question count is greater than 1, it belongs to the third degree. Based on three types of questions and answers respectively, nine degrees of user content contribution are identified.

The specific classification results are shown in table 1.

<table>
<tr><td colspan="2" rowspan="2"></td><td colspan="3">Question count</td><td rowspan="2">Answer degree</td><td rowspan="2">User type</td></tr>
<tr><td>=0</td><td>=1</td><td>>1</td></tr>
<tr><td rowspan="3">Answer count</td><td>=0</td><td>1</td><td>2</td><td>3</td><td>=1</td><td>non-answer users</td></tr>
<tr><td>∈[1,6]</td><td>4</td><td>5</td><td>6</td><td>=2</td><td>Superficial contribution answerers</td></tr>
<tr><td>>6</td><td>7</td><td>8</td><td>9</td><td>=3</td><td>Depth-contribution answerers</td></tr>
<tr><td colspan="2">Question degree</td><td>=1</td><td>=2</td><td>=3</td><td colspan="2" rowspan="2"></td></tr>
<tr><td colspan="2">User type</td><td>non-question users</td><td>Superficial contribution questioners</td><td>Depth-contribution questioners</td></tr>
</table>

Table 1. Content contribution degrees and their corresponding question and answer features.

As shown in TABLE 1, if the number of questions and answers of the user is equal to 0, the user belongs to class 1. When the number of questions of the user is 0, and the number of answers is between 1 and 6 (including 1 and 6), then the user belongs to class 4, and the same is true for other categories.

3.3 Research model

We try to predict the content contribution degree of users based on their personal information disclosure behavior.

In the previous study, we divided the content contribution degree into nine degrees. However, in the actual prediction, the number of independent variables predicted is less than the number of predicted categories, which is not conducive to the effectiveness of the prediction. Therefore, through the integration of nine categories of users, we respectively predict the degree of users' contribution to questions and answers.

We merged nine degrees into three categories of contribution to the answer. Degree 1, Degree 2, and Degree 3 are users with no answer contribution, Degree 4, 5, and 6 are users with a superficial contribution to answering, while Degree 7, Degree 8, and Degree 9 are users with depth contribution to answers.

User contribution to the question is also divided into three degrees. Degree 1, degree 4, and degree 7 are question-free users, degrees 2, 5, 8 are shallow question contribution users, and degree 3, degree 6, and degree 9 are deep question contribution users. We use Tensorflow to program to construct the softmax function and predict the user's question contribution degree and user's answer contribution degree respectively.

We built the research model of this paper based on the above method, as shown in figure 1.

We seek to explore the relationship between different types of personal information disclosure behavior and user content contribution behavior. We explored the disclosure behavior of five kinds of personal information, including gender, location, education, and employment. In addition, we added the amount of information disclosed by users to quantify the degree of self-disclosure. We measure the content contribution degree of users from two dimensions, including answer contribution degree and question contribution degree. Each dimension contains three degrees: no contribution, shallow contribution degree, and deep contribution degree.

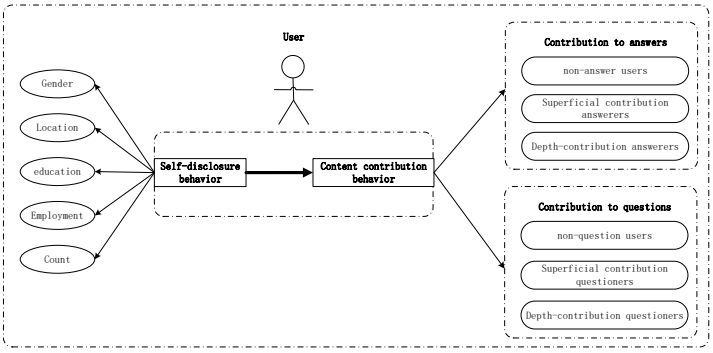

Figure 1. Research Model.

3.4 Predicting method

We need to use the loss function to solve the optimal W value and B value.

In machine learning, loss function (Loss) can be regarded as an indicator of the quality of a model. The function is usually minimized to find the optimal solutions for W and B. We use a common loss function, cross-entropy. It measures the inefficiency of our predictions in describing the truth. It is defined as follows:

$$H(y_i) = -\sum_i y_{label} \ln y_i$$

Where yi is the predicted probability that the sample belongs to category i, and ylabel is the actual probability that the sample belongs to category i.

Since we output more than two categories, the prediction in this paper belongs to multi-category prediction. For multiple classification problems, there are two solutions. The first is to construct a softmax

regression classifier, and the second is to construct multiple independent logistic classifiers. When the predicted categories are mutually exclusive, we choose the former; when the predicted categories are not completely mutually exclusive, we choose the latter. Since each user can only have one con-tent contribution degree, the relationship among categories is completely mutually exclusive. Softmax regression classifier was selected to predict the content contribution degree of users.

Softmax regression is an extension of binary logistic regression, specifically, the use of line-ar predictors and additional normalized factors (a logarithmic form of a partition function) to model the logarithm of the probability of a result.

We divided 4,290,484 pieces of user data according to the ratio of 9:1 (explanation for it), 90% of which were used as the training set, including 3861,435 pieces of data. The other 10% was our test set, containing 429,048 pieces of user data.

4 Result

4.1 Descriptive Statistics

First, our study carries out a descriptive analysis, which is shown in table 2.

Variable	Mean	Std. Deviation	Min	Max
gender	0.7302374	0.443837	0	1
location	0.1398221	0.3468028	0	1
education	0.2151186	0.4109051	0	1
employments	0.0638992	0.2445735	0	1
count	1.149077	0.9790534	0	4

Table 2. Descriptive statistics result.

4.2 The significant effect of user's self-disclosure behavior on content contribution degree

Based on the statistical description, we try to further affirm the significant effect of users' self-disclosure behavior on content contribution degree.

We use cross-test to examine the relationship between the two mentioned above. The former is the independent variable, namely gender disclosure behavior, location disclosure behavior, education background disclosure, professional experience disclosure, and the total number of information types disclosed by users. And the latter serves as the dependent variable, including nine degrees of user's content contributions. For meaningless variables, we will delete them in the later prediction to improve the accuracy of the prediction. The results are shown in Table 3, including the correlation between gender disclosure and content contribution degree, the relationship between the content degree of contribution and home to disclosure, the relationship between education disclosure and content contribution degree, as well as the correlation between employment disclose and content contribution degree.

	Pearson Chi-Square		Phi	Cramer's V
	Values	Asymp.Sig		
Class * gender	318502.620[a]	**.000**	**.272**	**.272**
Class * location	260860.968[a]	**.000**	**.247**	**.247**
Class * education	192297.317[a]	**.000**	**.212**	**.212**
Class * employments	109410.093[a]	**.000**	**.160**	**.160**
Class * Total-count	506394.273[a]	**.000**	**.344**	**.172**

Table 3. Chi-Square Tests.

The correlation is mainly determined by the Pearson chi-square test. It is considered that whether the user has disclosure behavior has a significant impact on the degree of content contribution if the sig value is less than 0.05. In addition, the phi value and V value are used to test the tightness of the relationship between the two variables.

In the cross-examination of gender disclosure behavior and content contribution degree, the sig value of the Pearson chi-square test is equal to 0.000, showing that whether the gender information is dis-closed has a significant impact on the content contribution degree. Phi value and V value are 0.272, greater than 0.1, indicating a close relationship, that is, the degree of contribution content has an obvious relationship with whether gender information is disclosed. This conclusion is consistent with the chi-square test results above.

The disclosure of location information has a significant impact on the degree of content contribution (sig=.000). Phi value and V value are 0.247, greater than 0.1, indicating whether the disclosure of local information has an obvious relationship with content contribution degree.

The disclosure of education information has a significant impact on the content contribution degree of users (sig=.000). Similarly, the phi value and V value are 0.212, greater than 0.1, indicating whether disclosure of education information or not is significantly related to the degree of content contribution.

The disclosure of employment information has also been tested to have a significant impact on the degree of content contribution (sig=.000), and the two are closely related (Phi value = V value =.160).

Finally, we examined the correlation between the total number of personal information types that users disclose and the degree of content contribution of users. The results are still significant (sig=.000).

4.3 Prediction of content contribution degrees through self-disclosure behavior

4.3.1 Prediction of user's answer contribution

First, we predict the degree of the user's contribution to the answer. There are three degrees, namely non-answer contribution, superficial contribution answerers, and Depth-contribution answerers.

Softmax regression prediction was implemented using Tensorflow, a deep learning framework. After 10 iterations of calculation, the loss function reaches a plateau and does not decline.

We drew the curve graph of the loss function changing with the increase of iteration times. We found that after the first iteration, the value of the loss function showed the most obvious decline. With the increase in iteration times, the curve tended to be flat and loss reached the minimum value

We drew the change diagram of the loss function with iteration times, as shown in figure 2.

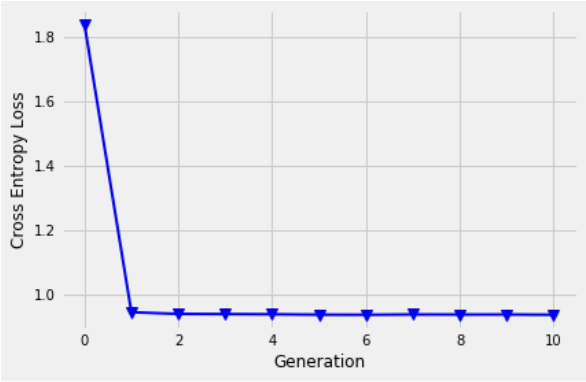

Figure 2. Loss function of answering contribution.

Thus, optimal weight value and bia value are obtained based on the minimum loss value:

$$\begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = softmax\begin{pmatrix} -0.679X_1 + 1.122X_2 + 0.874X_3 + 0.784X_4 - 0.290X_5 + 0.296 \\ -0.476X_1 + 1.123X_2 + 0.840X_3 + 0.649X_4 + 0.152X_5 - 0.740 \\ 0.428X_1 + 1.514X_2 + 1.401X_3 + 0.954X_4 + 0.016X_5 - 2.190 \end{pmatrix}$$

According to the optimal weight, we analyze the relationship between the disclosure behavior of different types of personal information and the degree of user content contribution.

If a newly registered user discloses his/her gender, the probability that he/she makes a deep answer contribution will increase by 0.428, and the probability that he/she does not make an answer contribution will decrease by 0.679. If users disclose their location information, they are most likely to make in-depth contributions, and the possibility of their contribution will increase by 1.514. If the education background is disclosed by the newly registered user, the possibility of the user making a deep answer contribution will increase by 1.401, which is much more than the possibility of not making a contribution or making a superficial contribution to answers. If users disclose their employment experience, the probability of making a depth contribution to an answer will increase by 0.954.

For every additional information disclosed by users, the probability that they do not contribute to answers will decrease by 0.29, the probability that they make a superficial contribution to answers will increase by 0.152, and the probability that they make depth contribution to answers will increase by 0.016.

In order to verify this model, we drew a graph of the prediction accuracy changing with the number of iterations (Figure 3).

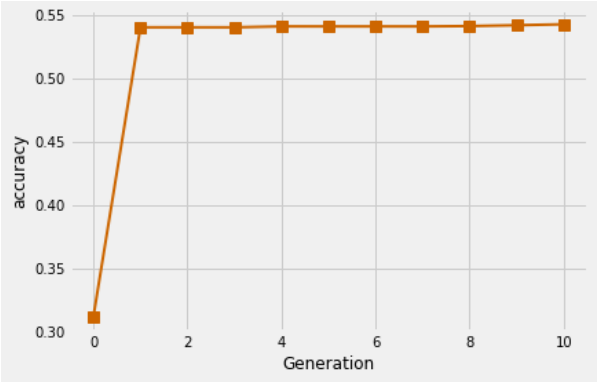

Figure 3. Accuracy of answering contribution.

We found that after the first iteration, the accuracy improved the fastest, and with the increase in iteration times, the accuracy degreed off. Finally, it reached 54.28%.

4.3.2 Prediction of user's question contribution

Then, we use the same method to predict the degree of the user's contribution to the problem, there are three categories, namely non-question contribution, superficial contribution questioners, and depth-contribution questioners. After 10 iterations, loss reaches the minimum value of 0.782. The change curve of loss as the number of iterations increases is shown in figure 4.

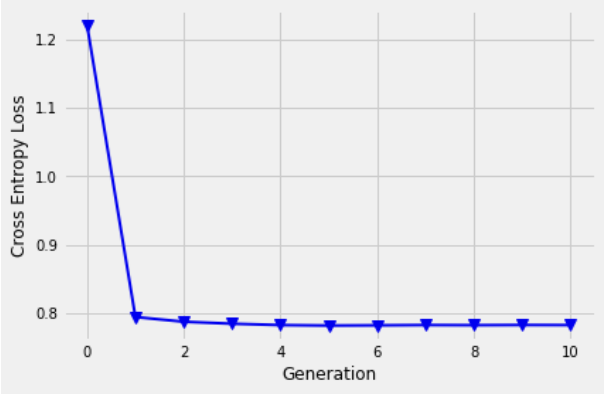

Figure 4. Loss function of questioning contribution.

On this basis, we output the optimal weight and the optimal bias value and construct an equation to predict the degree of the user's contribution to questions.

$$\begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = softmax\begin{pmatrix} -0.128X_1 - 0.202X_2 - 0.545X_3 + 0.080X_4 - 0.607X_5 - 0.035 \\ 0.551X_1 + 0.406X_2 - 0.054X_3 + 0.458X_4 - 0.599X_5 - 0.342 \\ 1.110X_1 + 0.649X_2 + 0.381X_3 + 0.759X_4 + 0.391X_5 - 0.469 \end{pmatrix}$$

Based on the prediction equation, we discuss the possible relationship between user disclosure behavior and the degree of contribution to the question. If users disclose their gender, they are 0.128 less likely not to ask questions, 0.551 more likely to make a superficial contribution to the question, and 1.110 more likely to make a depth contribution to the question. If the user discloses the location information, the possibility of the user not asking questions decreases by 0.202, the possibility of making a superficial contribution to the question increases by 0.406, and the possibility of making a depth contribution to the question increases by 0.649. If the user discloses his educational background, the probability of the user making depth contributions to the question increases by 0.381, while the probability of the user not contributing to the questions increases by 0.545. If the user discloses his employment experience, the possibility of making depth contribution to questions will increase by 0.759, which is the highest probability compared with the other two degree contribution. For every additional information disclosed by users, their probability of becoming depth-contribution questioners increases by 0.391. We drew a graph of the accuracy changing with the number of iterations (Figure 5).

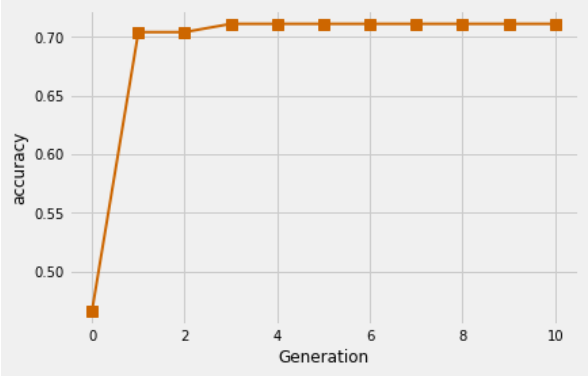

Figure 5. Accuracy of questioning contribution.

Finally, the prediction accuracy of the user's contribution to questions reached 71.126%.

4.3.3 The relationship between user's self-disclosure behavior and the way of contribute content

We further explore whether the disclosure of different types of personal information will lead to differences in the way users contribute to the content. On this basis, we attempt to study whether these differences are obvious. Because user disclosure behavior has the greatest impact on users' deep contribution behavior, we only explore the differences of deep contribution behavior caused by user self-disclosure behavior. By comparing the coefficients of the above two equations, we measure the difference in users' preference for content contribution. Since the measurement units of variables are consistent and the coefficient has not been standardized, the coefficient represents the degree of influence. As a result, the coefficients of the two equations are of comparative significance. The results are shown in Table 4.

	gender	locations	educations	employments	counts
Depth-contribution answerer	0.428	1.514	1.401	0.954	0.016
Depth-contribution questioner	1.110	0.649	0.381	0.759	0.391
The difference between the two	0.682	0.865	1.02	0.195	0.375

Table 4. Preference comparison of user content contribution behavior.

By comparing the coefficients of the two equations, we found that if a user discloses his/her gender, the probability that he/she will become a depth-contribution questioner will increase by 1.10, and the probability that he/she will become a depth-contribution answerer will only increase by 0.428. Therefore, users with gender disclosure behaviors are more likely to be inclined to contribute to questions. If users disclose their location information, the probability that they will become deep-contribution answerers will increase by 1.514, while the probability that they become depth-contribution question-ers will increase by 0.649. In contrast, users who reveal their location information are more likely to have depth-contribution answer behavior. Similarly, users who disclose educational backgrounds and professional experience are more likely to contribute in-depth to the answer. In addition, the amount of information disclosed by users has a greater impact on their questioning behaviors than on their answering behaviors.

In addition, we calculated the difference of regression coefficients between the two kinds of contribution behaviors to study the preference difference of user content contribution behaviors. We found that the preference for content contribution mode caused by education background disclosure is the most

obvious. In other words, users who disclose education are most likely to contribute to the answer, and this probability is much higher than the probability of contributing to questions. In addition, location information disclosure behavior can also lead to the probability that users make in-depth contributions to the answer being much higher than the probability that users make in-depth contributions to the question. Gender disclosure behavior will lead to the probability that users contribute questions far more than the probability of contributing answers. Employment disclosure has the least impact on the content contribution method. The more personal information is disclosed, the higher the probability of users asking questions, and the more obvious their preference for content contribution.

5 Discussion

This study has several major contributions. Firstly, we propose a new method to predict the content contribution intention and mode of newly registered users. Secondly, we take personal information disclosure behavior as a predictive factor to predict the content contribution behavior of users for the first time. Then, we have deeply studied the differences in user content contribution behaviors caused by different types of personal information disclosure. In addition, we use a large amount of data on users' online behavior to add credibility to our findings.

In previous studies, the researchers found that the disclosure of personal information would promote the content contribution behavior of users (Meng and Agarwal, 2007), but what is the promotion mechanism? Whether different types of personal information disclosure will lead to differences in content contribution behavior has not been further studied. In addition, the data of these studies are mostly obtained in the form of questionnaires, and their authenticity is difficult to be verified. We use a large number of online behavior data of users to prove the previous research, and find that the user's personal information disclosure behavior has a positive effect on the content contribution of users. In addition, we further studied the impact of different types of personal information disclosure on the degree and mode of the content contribution of users.

There are few studies on early users, and researchers often extract the characteristics of user behavior after users have a certain degree of community participation, to predict the later behavior of users. However, users are easily lost in the early stage. Without early intervention, the community can easily lose potential users. We propose a new idea to predict user's content contribution behavior, that is, to obtain users' intention of content contribution according to the personal information disclosed by users in the early stage, to predict the way and degree of users' content contribution, which is of great significance for improving the accuracy of question and answer matching on the platform and mining potential content contribution users.

The results of previous research show that the disclosure of personal information will promote users' content contribution behavior (Meng and Agarwal, 2007), and this paper verifies the result. We found that the more personal information users disclose, the higher the degree of user content contributions. And through the analysis of the characteristics of disclosure behavior, it is found that users who disclose personal information are often accompanied by a higher degree of content contribution behavior than those who do not disclose personal information. We analyzed the possible reasons. The amount of information disclosed by users represents the degree of trust in the community (He and Wei, 2009). The more information disclosed the more trust they have in the community, thus increasing their willingness to contribute to the content (Liu, 2013).

6 Conclusion

This paper puts forward a new way of predicting users' content contribution behavior, which makes it possible to predict the content contribution intention, mode, and contribution degree of newly registered users. We study the characteristics of information disclosure behavior of users with different content contribution degrees and predict their content contribution willingness and content contribution degree according to their self-disclosure behavior.

Statistical analysis of 4 million user data in the Zhihu community shows that users are more willing to disclose information that is more beneficial to their community images, such as highly educated, economically developed locations, popular majors, and occupations with higher social impact and social status. In addition, the deeper the contribution of user content, the higher the proportion of users who disclose personal information.

In addition, users who neither ask nor answer rarely disclose personal information, those who answered a lot but didn't ask questions disclosed the most personal information. Users who answered the most and questioned the most also disclose much personal information.

Furthermore, users who disclose gender have the lowest degree of content contribution, followed by users who disclose their location. The content contribution degree of users who disclose their education information reaches the medium degree, while the content contribution degree of users who disclose their employment is the highest

We construct the soft Max function to make a multi-classification prediction. The results of machine learning show that if users disclose their gender, the probability of depth contribution to the answer will increase by 0.428, and the probability of depth contribution to the question will increase by 1.110. If the user discloses the location information, the probability of making a depth contribution to the answer will increase by 1.514, and the probability of making a depth contribution to the question will increase by 0.649. If users disclose their educational background, the probability of their in-depth contribution to answers increases by 1.401, and the probability of their in-depth contribution to questions increases by 0.3811. If users disclose their professional experience, the possibility of contributing in-depth to the answers will increase by 0.954, and the possibility of contributing in-depth to the questions will increase by 0.759. For every additional personal information disclosed by a user, the probability of superficial contribution to the answer increases by 0.152, the probability of depth contribution to the answer increases by 0.016, and the probability of depth contribution to the question increases by 0.39.

In addition, the disclosure of different types of information will lead to different ways of user content contribution. Our study found that users who disclose gender are more likely to respond, while users who disclose their location, educational experience, and professional background are more likely to ask questions. Moreover, the disclosures of education experience lead to the most obvious differences in users' preferences for content contribution modes.

This paper proposes a new method to predict the willingness and degree of contribution of newly registered users, and deeply studies the specific relationship between users' disclosure behavior of different types of personal information and users' content contribution behavior. We provide a new method for the SQA community platform to identify potential content contribution users. This paper also overcomes the limitations of previous studies through a large amount of user data.

7 Limitation and future study

The contribution form of users in the question-and-answer community is diverse. This paper only studies the user's answer behavior and question behavior, but does not mention other contribution behaviors. For example, clicking the "thumb up" button to express recognition of others' content is also a kind of contribution behavior. In future studies, we can add other forms of contribution.

We assume that all personal information disclosure behaviors are generated when users register their accounts, and we do not consider that the disclosed information may be modified by users later. We believe that users' behavior of modifying or supplementing their personal information at a later stage is also worth studying.

We have only studied the user data of a Chinese SQA platform, but this method is universal, and we think this idea and method are still applicable to other online communities with the same platform nature.

References

Acquisti, A., and Gross, R. (2006). Imagined communities: Awareness, information sharing, and privacy on the Facebook. In *International workshop on privacy enhancing technologies*, 36–58. Springer.

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In: *Proceedings of the international conference on Web search and web data mining - WSDM '08* (p. 183). https://doi.org/10.1145/1341531.1341557

Bansal, G., and Gefen, D. (2010). The impact of personal dispositions on information sensitivity, privacy concern and trust in disclosing health information online. *Decision Support Systems*, *49*(2), 138–150.

Bansal, G., Zahedi, F. M., and Gefen, D. (2016). Do context and personality matter? Trust and privacy concerns in disclosing private information online. *Information and Management*, *53*(1), 1–21. https://doi.org/10.1016/j.im.2015.08.001

Barak, A., and Gluck-Ofri, O. (2007). Degree and Reciprocity of Self-Disclosure in Online Forums. *CyberPsychology & Behavior*, *10*(3), 407–417. https://doi.org/10.1089/cpb.2006.9938

Batenburg, A., and Bartels, J. (2017). Keeping up online appearances: How self-disclosure on Facebook affects perceived respect and likability in the professional context. *Computers in Human Behavior*, *74*, 265–276. https://doi.org/10.1016/j.chb.2017.04.033

Benndorf, V., Kübler, D., and Normann, H.-T. (2015). Privacy concerns, voluntary disclosure of information, and unraveling: An experiment. *European Economic Review*, *75*, 43–59.

Benson, V., Saridakis, G., and Tennakoon, H. (2015). Information disclosure of social media users: Does control over personal information, user awareness and security notices matter? *Information Technology and People*, *28*(3), 426–441. https://doi.org/10.1108/ITP-10-2014-0232

Bryce, J., and Fraser, J. (2014). The role of disclosure of personal information in the evaluation of risk and trust in young peoples' online interactions. *Computers in Human Behavior*, *30*, 299–306.

Chang, H. H., and Chuang, S.-S. (2011). Social capital and individual motivations on knowledge sharing: Participant involvement as a moderator. *Information & Management*, *48*(1), 9–18.

Chiu, C.-M., Hsu, M.-H., and Wang, E. T. G. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision Support Systems*, *42*(3), 1872–1888. https://doi.org/10.1016/j.dss.2006.04.001

Christofides, E., Muise, A., and Desmarais, S. (2009). Information Disclosure and Control on Facebook: Are They Two Sides of the Same Coin or Two Different Processes? *CyberPsychology & Behavior*, *12*(3), 341–345. https://doi.org/10.1089/cpb.2008.0226

Dwyer, C., Hiltz, S. R., and Passerini, K. (2007). Trust and privacy concern within social networking sites: a comparison of Facebook and MySpace. In: *Thirteenth Americas conference on information systems* (Vol. 123, pp. 339–350). https://doi.org/10.1.1.148.9388

Gross, R., Acquisti, A., and Heinz, H. J. (2005). Information revelation and privacy in online social networks. In: *Proceedings of the 2005 ACM workshop on Privacy in the electronic society - WPES '05* (p. 71). https://doi.org/10.1145/1102199.1102214

He, W., and Wei, K. K. (2009). What drives continued knowledge sharing? An investigation of knowledge-contribution and -seeking beliefs. *Decision Support Systems*, *46*(4), 826–838. https://doi.org/10.1016/j.dss.2008.11.007

Johnson, B. K., and Ranzini, G. (2018). Click here to look clever: Self-presentation via selective sharing of music and film on social media. *Computers in Human Behavior*, *82*, 148–158. https://doi.org/10.1016/j.chb.2018.01.008

Jourard, S. M., and Lasakow, P. (1958). Some factors in self-disclosure. *Journal of Abnormal and Social Psychology*, *56*(1), 91–98. https://doi.org/10.1037/h0043357

Kear, K., Chetwynd, F., and Jefferis, H. (2014). Social presence in online learning communities: The role of personal profiles. *Research in Learning Technology*, *22*. https://doi.org/10.3402/rlt.v22.19710

Kisekka, V., Bagchi-Sen, S., and Raghav Rao, H. (2013). Extent of private information disclosure on online social networks: An exploration of Facebook mobile phone users. *Computers in Human Behavior*, *29*(6), 2722–2729. https://doi.org/10.1016/j.chb.2013.07.023

Krasnova, H., and Veltri, N. (2011). Behind the curtains of privacy calculus on social networking sites: The study of Germany and the USA. In *10th International Conference on Wirtschaftsinformatik* (pp. 891–900).

Lampe, C., Wash, R., Velasquez, A., and Ozkaya, E. (2010). Motivations to participate in online communities. In *Proceedings of the 28th international conference on Human factors in computing systems - CHI '10* (p. 1927). https://doi.org/10.1145/1753326.1753616

Le, L. T., and Chirag, S. (2018). Retrieving people: Identifying potential answerers in Community Question-Answering. *Journal of the Association for Information Science and Technology*, *69*(10), 1246–1258.

Liu, W. (2013). Knowledge Sharing Mechanisms: Characteristics and Roles in Knowledge Sharing. PhD Thesis.

Meng, M., and Agarwal, R. (2007). Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Information Systems Research*, *18*(1), 42–67. https://doi.org/10.1287/isre.1070.0113

Milošević, M., Živić, N., and Andjelković, I. (2017). Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications*, *83*, 326–332.

Nosko, A., Wood, E., and Molema, S. (2010). All about me: Disclosure in online social networking profiles: The case of FACEBOOK. *Computers in Human Behavior*, *26*(3), 406–418. https://doi.org/10.1016/j.chb.2009.11.012

Seo, J. Y., Lim, H. S., and Choi, J. W. (2014). Threshold value of Benthic Pollution Index (BPI) for a muddy healthy benthic faunal community and its application to Jinhae Bay in the southern coast of Korea. *Ocean Science Journal*, *49*(3), 313–328.

Shin, D., and Biocca, F. (2018). Exploring immersive experience in journalism. *New Media & Society*, *20*(8), 2800–2823.

Special, W. P., and Li-Barber, K. T. (2012). Self-disclosure and student satisfaction with Facebook. *Computers in Human Behavior*, *28*(2), 624–630. https://doi.org/10.1016/j.chb.2011.11.008

Spiekermann, S., Grossklags, J., and Berendt, B. (2001). E-privacy in 2nd Generation E-Commerce : Privacy Preferences versus actual Behavior. In *ACM Conference on Electronic Commerce* (pp. 1–10). https://doi.org/10.1145/501158.501163

Tang, Q., Gu, B., and Whinston, A. (2012). Content Contribution for Revenue Sharing and Reputation in Social Media: A Dynamic Structural Model. *Journal of Management Information Systems*, *29*(2), 41–76.

Techweb. (2018). The number of registered users on zhihu exceeded 200 million, and the number of new users exceeded 80 million in 2018. URL: http://www.techweb.com.cn/internet/2018-11-07/2711199.shtml (retrieved 11/07/2018)

Tsay-Vogel, M., Shanahan, J., and Signorielli, N. (2018). Social media cultivating perceptions of privacy: A 5-year analysis of privacy attitudes and self-disclosure behaviors among Facebook users. *New Media & Society*, *20*(1), 141–161.

Van Gool, E., Van Ouytsel, J., Ponnet, K., and Walrave, M. (2015). To share or not to share? Adolescents' self-disclosure about peer relationships on Facebook: An application of the Prototype Willingness Model. *Computers in Human Behavior*, *44*, 230–239. https://doi.org/10.1016/j.chb.2014.11.036

Wang, C., Blei, D., and Fei-Fei, L. (2009). Simultaneous image classification and annotation. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009* (Vol. 2009 IEEE, pp. 1903–1910). https://doi.org/10.1109/CVPRW.2009.5206800

Wang, Y., Norcie, G., Komanduri, S., Acquisti, A., Leon, P. G., and Cranor, L. F. (2011). "I regretted the minute I pressed share": a qualitative study of regrets on Facebook. In *SOUPS* (p. 1). https://doi.org/10.1145/2078827.2078841

Wang, Z., and Zhang, P. (2016). Examining user roles in social Q&A: The case of health topics in Zhihu.com. *Proceedings of the Association for Information Science and Technology*, *53*(1), 1–6. https://doi.org/10.1002/pra2.2016.14505301103

WENTING, L. I. U. (2013). Knowledge Sharing Mechanisms: Characteristics and Roles in Knowledge Sharing.

Xie, W., and Kang, C. (2015). See you, see me: Teenagers' self-disclosure and regret of posting on social network site. *Computers in Human Behavior*, *52*, 398–407. https://doi.org/10.1016/j.chb.2015.05.059

Xu, B., and Li, D. (2015). An empirical study of the motivations for content contribution and community participation in Wikipedia. *Information and Management*, *52*(3), 275–286. https://doi.org/10.1016/j.im.2014.12.003

Zhang, X., Liu, S., Chen, X., Wang, L., Gao, B., and Zhu, Q. (2018). Health information privacy concerns, antecedents, and information disclosure intention in online health communities. *Information & Management*, *55*(4), 482–493.

Zhou, J. (2018). Factors Influencing People's Personal Information Disclosure Behaviors in Online Health Communities: A Pilot Study. *Asia-Pacific Journal of Public Health*, *30*(3), 286–295. https://doi.org/10.1177/1010539518754390