Summer 6-30-2018

# Quality Prediction of Answers in Community of Questions and Answers of Zhihu

Ming Li
*School of business administration, China University of Petroleum-Beijing, China*

Yi Zhang
*School of business administration, China University of Petroleum-Beijing, China*

Kaijuan Xing
*School of information, University of Texas at Austin, USA*

Xiaoyu Qi
*School of business administration, China University of Petroleum-Beijing, China*

# Quality Prediction of Answers in Community of Questions and Answers of Zhihu

*Ming Li,* [1]*Yi Zhang*[1,*]*, Kaijuan Xing*[2]*, Xiaoyu Qi*[1]
[1]School of business administration, China University of Petroleum-Beijing, China
[2]School of information, University of Texas at Austin, USA

**Abstract:** The participation of Web2.0 in collaboration makes the Q&A social platform develop rapidly, but the problem that the answering quality is difficult to distinguish has gradually become apparent, even it hindered the healthy development of social platforms. In order to predict the Answers in Zhihu, which is the famous Community of Questions and Answers website, the quality prediction method based on Logistic Regression is proposed. Firstly, we collect dataset from Zhihu. Based on the factors, the training set is constructed. After that the two answer quality prediction models based on Logistic Regression is constructed. With the comparison of the two models, the final model is determined. It can be used to predict the answer quality of Zhihu.

Keywords**:** Community of questions and answers; Answer quality prediction; Zhihu; Logistic Regression

## 1. INTRODUCTION

With the coming of Web 2.0, the way people obtain information has changed from the original one-way transmission to user-centered, which is a network communication mode emphasizing collective sharing [1,2]. Various Community of questions and answers (CQA) platforms emerged. Their emergence greatly improved the efficiency of the Internet use, and to a certain extent, changed the inherent mode of the Internet. Q&A social networking sites use questions and answers as the primary mode of communication, where users can obtain the similar question and answer with that users posted previously, or invite users on the site to answer the question, and then choose an answer with high quality [3,4]. Zhihu is such a Q&A community platform. In recent years, the development of Zhihu is very fast, and registered users have also increased significantly [5,6]. The information on Zhihu is mainly based on questions and answers of various topics, and the questioner and the responder can communicate through the platform. The main service of Zhuhu is to provide users with high quality answers, but with the amount of users and information increasing, the quality of the answer becomes more and more difficult to distinguish, so the quality of the information provided to users is difficult to guarantee, which becomes a challenge for the development of community websites like Zhihu [7-9]. Therefore, it is essential to build a scientific method of evaluating and predicting the quality of the answers.

The evaluation of answer quality is one of the hot topics. Shah C took Yahoo Answers as the research object, let five manual scorer from Amazon score from thirteen dimensions including informative, polite, complete readable, relevant, brief, convincing, detailed, original, objective, novel, helpful and expert, and then analyzed the data and set up the model based on Logistic Regression. Then he further extracted questions, answers and users characteristics and used the same method for analysis and modeling [10,11]. Agichtein E [12] is also a study of Yahoo Answers, based on the analysis to variable factors associated with high quality contents, ultimately created an icon based quality classification framework with the algorithm model of the relationship between questioners and respondents, text features of the content and features of the usage. Toba et al [13] used Support Vector Machines, Logistic Regression and other basic classification methods to analyze the extracted

text and non-text features, and then by comparison, a classification method with the best effect was obtained. They also analyzed the relationship between the types of questions and quality of answers. Hoang, Lee et al extracted the characteristics of four aspects: the answer, the rigor, the readability and the subjectivity, then used the classification method of maximum entropy to analyze the data, and then established the quality classification model [14]. Jia Jia et al. assessed the model by the answer quality with thirteen dimensions, and evaluated the answering quality of the two domestic social Q&A platforms—Zhihu and Baidu knows through the questionnaire [15].

These researches mostly focused on the famous CQA website Yahoo! Answers, but less research is targeted at the Chinese CQA website. In this paper, we take the Zhihu as the main research object and establish the experimental training set. Different classification methods are used to analyze the data. An optimal algorithm model is determined through comprehensive analysis and comparison. The second part of the paper is a detailed introduction to the adopted Logistic Regression. The third part introduces the related contents of the experiment, including the introduction to the specific process of the experiment, the analysis of the experimental results and the quality prediction model established. The fourth part is the summary of the research, the significance of the study, the analysis of the existing problems and the prospect of the future research work.

## 2. LOGISTIC REGRESSION

As one of the most important classification models in pattern recognition and machine learning, Logistic Regression is an interpretable model and has a good generalization ability [16]. Logistic Regression is a regression model in which variables are classified and are applied to various fields including machine learning, medical field and social science [17]. For the two-element Logistic Regression function (that is, there are only two types of the dependent variable) it can take any actual input T and the output always values between 0 and 1. Logical Regression is a typical probability statistical classification model. It fits logarithm by the linear function and can be expressed as:

$$\ln\left(\frac{p(y=1|x)}{p(y=2|x)}\right) = w^{\mathrm{T}}x \tag{1}$$

$w$ is a fitting parameter, $x$ is an instance, y represents the label of the class and the $p$ represents the conditional probability.

Because the sum of the probability of each class is 1:

$$\sum_{j=1}^{2} p(y=j|x) = 1 \tag{2}$$

Combine formula (1) and formula (2), available:

$$\mathrm{p} = (\mathrm{y}=1|\mathrm{x}) = \frac{\exp\,(w^{T}x)}{1+\exp\,(w^{T}x)'} \tag{3}$$

$$\mathrm{p} = (\mathrm{y}=2|\mathrm{x}) = \frac{1}{1+\exp\,(w^{T}x)'} \tag{4}$$

The value of $w$ is often estimated by the maximum likelihood method, and its logarithmic likelihood function is:

$$\mathrm{L}(w) = \Sigma_{j=1}^{2}\Sigma_{i=1}^{N_j} lnp(x_i{}^{(j)}|y=j;w) \tag{5}$$

$N_j$ represents the number of class j instances[16].

In conclusion, it is a generalized linear model with sigmoid function and can also be expressed as:

$$P_i = Logit^{-1}\,(\beta\cdot \mathrm{X}_i) = \frac{1}{1+e^{-\beta\cdot \mathrm{X}_i}} \tag{6}$$

$\beta \cdot \mathrm{X}_i = \beta_0 + \beta_1 \cdot x_{1,i} + \cdots + \beta_m \cdot x_{m,i}$ , $\beta$ is the Regression coefficient , $x_i$is the explanatory variable[17].

## 3. EXPERIMENT

### 3.1Logistic Regression [1]

We collect 500 answers from Zhihu. Then the quality of the answers is evaluated manually. The standard of assessment is whether the content of the answer is related to the question, whether the information provided in the answer is true and reliable, whether it can solve the questioner's question and so on. We refer to the thirteen dimensions scoring model of Community answer quality proposed by Shah C et al [11]., let each scorer scores for answers respectively from the angels of Informative, Polite, Complete, Readable, Relevant, Brief, Convincing, Detailed, Original, Objective, Novel, Helpful, and Expert in the range of 0-5, and according to those scoring points and the assessment data of answer quality, we set up a training set. The prediction of answer quality is actually a classification problem. We will divide the answer quality into two categories, including good quality and poor quality, which are expressed by 1 and 0 respectively. 1 corresponds to good quality, while 0 represents poor quality.

The essence of Logistic Regression is a classification model, the input value is random, but there are only two types of output result, and Shah C et al [11] also used Logistic Regression to model. So we analyze the data of the training set with the method of Logistic Regression using SPSS statistical analysis software. The results are as follows:

**Table 1 Variables in the Equation**

| Step1 | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| Informative | -.001 | .087 | .000 | 1 | .993 | .999 |
| Polite | .008 | .087 | .009 | 1 | .923 | 1.008 |
| Complete | .107 | .089 | 1.426 | 1 | .232 | 1.113 |
| Readable | .065 | .098 | .443 | 1 | .506 | 1.067 |
| Relevant | .711 | .106 | 44.659 | 1 | .000 | 2.036 |
| Brief | .086 | .098 | .772 | 1 | .379 | 1.090 |
| Convincing | -.017 | .103 | .026 | 1 | .871 | .983 |
| Detailed | .917 | .110 | 68.962 | 1 | .000 | 2.501 |
| Original | .007 | .092 | .006 | 1 | .940 | 1.007 |
| Objective | -.057 | .092 | .381 | 1 | .537 | .945 |
| Novel | -.106 | .097 | 1.206 | 1 | .272 | .899 |
| Helpful | -.020 | .089 | .052 | 1 | .819 | .980 |
| Expert | .040 | .092 | .188 | 1 | .664 | 1.041 |
| Constant | -5.052 | .775 | 42.454 | 1 | .000 | .006 |

In table 1, B represents the coefficient of variables. Sig. reacts to the significance of variables. If the Sig. <0.05 of a equation’s variable, it means that this variable has statistical significance, and vice versa. According to the data in the table, it can be found that only "Relevant" and "Detailed" variables have statistical significance. Therefore, the poorly significant variables are removed and the data are analyzed again, and the results are as follows:

**Table 2Variables in the Equation**

| Step 1 | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| Relevant | . 734 | . 101 | 52. 440 | 1 | . 000 | 2. 084 |
| Detailed | . 875 | . 106 | 68. 185 | 1 | . 000 | 2. 398 |
| Constant | −4. 803 | . 415 | 133. 710 | 1 | . 000 | . 008 |

The model of Logistic regression that can be obtained from the above table is as follows:

$$P = \frac{1}{1 + e^{-(-4.803 + 0.734 * Relevant + 0.875 * Detail)}}$$

When the probability of prediction is P>0.5, the quality of the answer is good; when P<0.5, the quality of the answer is poor.

**Table 3. Classification Table**

| Observed | | | Predicted | | |
|---|---|---|---|---|---|
| | | | The quality of the answer | | Percentage |
| | | | NO | YES | Correct |
| Step 1 | The quality of the answer | NO | 313 | 38 | 89.2 |
| | | YES | 65 | 84 | 56.4 |
| | Overall Percentage | | | | 79.4 |

This table responds to the accuracy of the model. Known from the table, the accuracy rate of the answer predicted with pool quality is 89.2%, and the accuracy of the answer predicted with good quality is 56.4%, and the total accuracy is 79.4%.

**Table 4. Model Summary**

| Step | -2 log likelihood | Cox&Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 432.135[a] | .296 | .420 |

Table 4 reflects the fitting degree of the model. The closer the value of Cox&Snell R Square and Nagelkerke R Square is to 1, the better the model fits. According to the table, the two values are small, so the fitting degree of the model is not very ideal. The requirements of the model algorithm established by the thirteen dimensional method are relatively high for the user itself, because users' scoring on different aspects directly affects the prediction of answer quality, so the final predictive results are highly uncertain and the precision and the fitting degree of the model is not high.

### 4.2 Logistic Regression [2]

In order to improve the practicability of the model, we decide to establish a new model, Toba [7], Agichtein E[12] , Shah C et al [10,11],analyzed directly to the question text and non-text features, so we decide to extract relevant features from questions and answers on Zhihu. Because the quality of the answer is not only related to the answer itself, also may be associated with the problem and answer users, we decide to extract the

characteristics of questions, answers and respondents. According to the characteristics of Zhihu operating mechanism, we finally decide to extract the following features: the length of the question, the number of labels of the question, the number of comments on the question, the number of answers to the question; the order of the answer, the length of the answer, the number of sentences in answers, the number of comments of the answer, the number of approval of the answer; the number of answers of the respondent, the number of share of the respondent, the number of questions of the respondent, The number of respondents' attention, the number of concerns of the respondent, the number of endorsed of the respondent, the number of thanks to the respondent and the number of collection to the respondent. The above characteristics almost involve all aspects of questions and answers. Therefore, the models built on this are more practical and objective.

We have collect another 500 answers from Zhihu. The criteria for judging the quality of answers remain unchanged, and then a new training set is set up. We still use Logistic Regression to analyze the relevant data of the training set, and the final results are as follows:

**Table 5. Variables in the Equation**

| Step1 | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| The length of the question | .034 | .013 | 6.460 | 1 | .011 | 1.034 |
| The number of labels of the question | .036 | .015 | 5.684 | 1 | .017 | 1.036 |
| The number of comments on the question | -.010 | .005 | 3.307 | 1 | .069 | .990 |
| The number of answers to the question | .000 | .000 | 3.988 | 1 | .046 | 1.000 |
| The order of the answer | .048 | .053 | .836 | 1 | .361 | 1.050 |
| The length of the answer | .001 | .000 | 26.993 | 1 | .000 | 1.001 |
| The number of sentences of the answer | -.008 | .008 | 1.017 | 1 | .313 | .992 |
| The number of comments of the answer | -.001 | .001 | 1.117 | 1 | .291 | .999 |
| The number of approval of the answer | .000 | .000 | 6.643 | 1 | .010 | 1.000 |
| the number of answers of the respondent | .000 | .000 | .061 | 1 | .805 | 1.000 |
| the number of share of the respondent | -.003 | .003 | .684 | 1 | .408 | .997 |
| the number of questions of the respondent | .001 | .001 | .461 | 1 | .497 | 1.001 |
| The number of respondents' attention | .000 | .000 | .002 | 1 | .962 | 1.000 |
| The number of concerns of the respondent | .000 | .000 | 6.362 | 1 | .012 | 1.000 |
| The number of endorsed of the respondent | .000 | .000 | .443 | 1 | .506 | 1.000 |
| The number of thanks to the respondent | .000 | .000 | .165 | 1 | .685 | 1.000 |
| The number of collection to the respondent | .000 | .000 | .031 | 1 | .861 | 1.000 |
| Constant | -3.860 | .605 | 40.708 | 1 | .000 | .021 |

From the Table 5, we can see that variables of Sig.<0.05 include: the length of the question, the number of the question, the number of answers to the question, the answer length, the answer agreement and the respondents' attention. Therefore, removing the variables with poor significance, the data of the training set are analyzed again and the results are as follows:

**Table 6. Variables in the Equation**

| Step1 | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| The length of the question | .024 | .012 | 4.035 | 1 | .045 | 1.024 |
| The number of labels of the question | .034 | .014 | 5.820 | 1 | .016 | 1.035 |
| The number of answers to the question | .000 | .000 | .389 | 1 | .533 | 1.000 |
| The length of the answer | .001 | .000 | 60.036 | 1 | .000 | 1.001 |
| The number of approval of the answer | .000 | .000 | 10.422 | 1 | .001 | 1.000 |
| The number of concerns of the respondent | .000 | .000 | 17.658 | 1 | .000 | 1.000 |
| Constant | -3.389 | .484 | 48.961 | 1 | .000 | .034 |

The table shows Sig=0.533>0.05 of "the number of answers to the question", so it is necessary to remove " the number of answers to the question " variable and reanalyze the data.

The result is the following table:

**Table 7. Variables in the Equation**

| Step1 | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|
| The length of the question | .024 | .012 | 3.893 | 1 | .048 | 1.024 |
| The number of labels of the question | .033 | .014 | 5.585 | 1 | .018 | 1.034 |
| The length of the answer | .001 | .000 | 60.077 | 1 | .000 | 1.001 |
| The number of approval of the answer | .000 | .000 | 12.013 | 1 | .001 | 1.000 |
| The number of concerns of the respondent | .000 | .000 | 17.969 | 1 | .000 | 1.000 |
| Constant | -3.348 | .480 | 48.741 | 1 | .000 | .035 |

A new Logistic Regression model can be obtained from the table as follows:

$$\mathbf{P}=\frac{1}{1+e^{-(-3.348+0.024a+0.033b+0.001c+0.000146d+0.000020e)}}$$

Among them, a, b, c, d and e respectively represent the length of the problem, the number of labels of the question, the length of the answer, the number of answers to the question, and the number of concerns of the respondent.

**Table 8. Classification Table**

| Observed | | | Predicted | | Percentage Correct |
|---|---|---|---|---|---|
| | | | The quality of the answer | | |
| | | | No | Yes | |
| Step1 | The quality of the answer | No | 321 | 21 | 93.9 |
| | | Yes | 68 | 90 | 57.0 |
| | Overall Percentage | | | | 82.2 |

From Table 8, the total probability of the prediction of the new Logistic Regression model is 82.2%, which is better than the previous one.

**Table 9. Model Summary**

| Step | -2 Log likelihood | Cox&Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 425.126[a] | .328 | .460 |

From the table, the values of Cox&Snell R Square and Nagelkerke R Square are all increased, and the fitting degree of the model is also improved.

Therefore, the effect of the established algorithm model is quite good, no matter the accuracy rate or the degree of fitting is ideal, and the variables are directly acquired characteristics eliminating the influence of human factors.

According to the results of the above experiments, the accuracy of each algorithm model is as follows:

**Table 10. Accuracy**

| Algorithm model | Logression[1] | Logression[2] |
|---|---|---|
| accuracy rate | 79.4 | 82.2 |

According to the above table, the best model in the modeling is the Logression[2] model.

## 5. Conclusions

In this paper, the method for quality prediction of answers in Community of Questions and Answers is proposed. Based on the collected dataset from the famous Community of Questions and Answers website Zhihu and the factors that have influence on the quality of answers, the quality prediction method based on Logistic Regression is proposed. After that the two answer quality prediction models based on Logistic Regression is constructed. With the comparison of the two models, the final model is determined. It can be used to predict the answer quality. Our experiments also have some shortcomings and improvements, such as the training dataset is too little. We can further explore the optimal model; the quality and the type of problems may affect the quality of the answer; respondents' mood may also be related to the quality of the answer and all of these can further be researched.

## ACKNOWLEDGEMENT

This research was supported by the National Natural Science Foundation of China under Grant 71571191, the Humanity and Social Science Youth Foundation of Ministry of Education in China (Project No. 15YJCZH081), National Natural Science Foundation of China under Grant 91646122, the Science Foundation of China University of Petroleum, Beijing (No. 2462015YQ0722).

## REFERENCES

[1] Kamel B, M. N, & Wheeler S.(2007).The emerging Web 2.0 social software: an enabling suite of sociable technologies in health and health care education. *Health Information & Libraries Journal*, *24*(1): 2-23.

[2] Deng Z, Luo L. (2007). An exploratory discuss of new ways for competitive intelligence on Web2. 0. *Integration and Innovation Orient to E-Society Volume 2*, 597-604.

[3] Liu J, Shen H, Yu L. (2017). Question quality analysis and prediction in community question answering services with coupled mutual reinforcement. *IEEE Transactions on Services Computing*, *10*(2): 286-301.

[4] Figueroa A.(2017).Automatically generating effective search queries directly from community question-answering questions for finding related questions. *Expert Systems with Applications*, *77*: 11-19.

[5] Zhang J., Ackerman M S, Adamic L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web* ,221-230

[6] Dror G, Koren Y, Maare k , Szpektor I.(2011).I want to answer; who has a question?: Yahoo! answers recommender system. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining* , 1109-1117.

[7] Toba H, Ming Z Y, Adriani M, Chua T S.(2014).Discovering high quality answers in community question answering archives using a hierarchy of classifiers. *Information Sciences, 261*: 101-115.

[8] Liu Y, Agichtein E. (2008).On the evolution of the Yahoo! Answers QA community. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*: 737-738.

[9] Shtok A, Dror G, Maarek Y, Szpektor I.(2012).Learning from the past: answering new questions with past answers. In *Proceedings of the 21st international conference on World Wide Web*: 759-768.

[10] Shah C,Kitzie V,Choi E.(2014).Questioning the Question-Addressing the Answerability of Questions in Community Question-Answering.IEEE Xplore,1530-1605.

[11] Shah C, Pomerantz J.(2010).Evaluating and Predicting Answer Quality in Community QA.ACM SIGIR conference ,411-418.

[12] Agichtein E,Castillo C,Donato D,Gionis A,Mishne G.(2008). Finding High-Quality Content in Social Media. International Conference on Web Search and Data Mining,183-194.

[13] Toba H, Zhao Y M, Adriania M,Chua T-S.(2014).Discovering high quality answers in community question answering archives using a hierarchy of classifers. Infornmation Sciences 261,101-115.

[14] Hoang L,Lee J-T,Song Y-I,Rim H-C. (2008). A Model for Evalating the Quality of User-Created Documents. Information Retrieval Technology496-501.

[15] Jia Jia,Song Enmei,Su Huan.(2013). The quality assessment of the social question and answer platform - "Zhihu", "Baidu Zhidao" as an example [J]. Journal of information resources management, 02: 2095-2171. (in Chinese)

[16] Guo Huaping.,Dong Yadong.,Wu Changan&Fan Ming.(2015). Logistic Regression Method for Class Imbalance Problem.Pattern Recognition and Artificial Intelligence,8.

[17] Le L T, Shah C, Choi E.(2016).Evaluating the Quality of Educational Answers in Community Question-Answering[J],129-138.