Quality Evaluation for Answers in Chinese QA Community Based on Random Forest –– Examples from Zhihu

Hongwei Wang [1, a], Yuqiang Ji [1, b], Wei Wang [2, c]

[1]School of Economics and Management, Tongji University, Shanghai, 200092, China

[2]College of Business Administration, Huaqiao University, Quanzhou, 362021, China

[a]email: hwwang@tongji.edu.cn, [b]email: 2014yqji@tongji.edu.cn, [c]email: wwang@hqu.edu.cn

Keywords: QA Community, Quality Evaluation, Random Forest

Abstract. As user-generated content develops continuously, online QA communities have become a major way to access high quality knowledge and information. Common concern of these communities is therefore to determine high quality answers accurately. This paper focuses on the largest Chinese QA community website Zhihu, and studies part of answers of the twenty popular topics we choose on it. We implement feature extraction to these answers in basic information, diversity and correlation with questions, and then apply random forest classifier to model and analyze the data. The classification model can predict the quality of answers with reasonable accuracy of more than 86%. Robustness test shows that the model has high stability.

Introduction

With the rapid development of Web 2.0, the dissemination mode of network information shifts rapidly from one-way mode to user-centered divergent mode. On this background, UGC (User-Generated Content) develops super-fast and online QA communities springs up with it. In QA communities, users can both raise and answer questions and give feedback by voting, commenting, selecting the best answer etc. [1]. Famous online QA communities so far include Yahoo! Answers, Uclue, Answers.com, Quora. Baidu Zhidao, Zhihu are more successful ones in China then. By September 2015, Zhihu has gained tens of thousands of topics that include 6.1 million questions and 19 million answers have been created in total. Despite article [2] indicates that user generated contents have high reliability and inevitably, and the quality of data varies greatly because of the huge number. The criterion for users to identify whether an answer is of high quality or not is based on seven kind of values which are content value, cognitive value, social sentiment value, information source value, external value, practicability and language popularity according to Kim's research on Yahoo! Answers [3]. Harper's research shows that the answer quality in communities that need to pay is higher than free ones [4]. Researches from Shah claims that the information contained in answers, the completeness and the novelty of answers are all have influences on answer quality [5]. To predict answer qualities, paper [6] constructs a model with non-textual features based on maximum entropy method and kernel density estimation, and get a better predicting performance than traditional models. To accurately evaluate answer quality, not only the answers themselves should be considered, but questions and answer writers may also have critical effects, as per paper [7].

The studies of Chinese QA communities are still having many deficiencies. Considering the huge differences between English and Chinese, the model fits in English communities may not suitable with Chinese communities. We use data from Zhihu, the largest Chinese QA community, propose a model to predict answer quality.

Dataset Description

Every question in Zhihu has its own topic, and consists of question title and description, each question has several answers. This structure of the website gives us a hint to crawling some valuable information. All topics in Zhihu construct a rooted directed graph trough parent-child

relationship, there are numerous topics and they have different segmentations, therefore, we select 20 active topics as the entrance of the data crawling, and they all have relatively moderate segmentation. Table 1 shows the summary of parts of data.

Tab.1. Summary of data

Topic	Question number	Answer number	Question Avg. comments	Question Avg. followers	Question Avg. answers	Answer Avg. comments	Answer Avg. votes
Animation	10000	82006	0.69	48.72	8.20	4.42	20.70
Arts	10000	56967	0.41	57.22	5.70	3.99	26.58
Automobile	10000	38700	0.24	26.08	3.87	2.98	10.11
Business	10000	36406	0.34	45.77	3.64	2.87	13.65
Culture	10000	41099	0.37	37.57	4.11	3.55	18.81
Economics	10000	47466	0.44	65.40	4.75	3.58	20.37
Entrepreneurship	10000	28327	0.26	38.90	2.83	2.92	25.74
…	…	…	…	…	…	…	…

Feature Extraction

We extract some features from answers themselves and the relationships between answers and questions. And more specifically, the characteristics of answer itself are divided into two parts: basic information and textual diversity.

Features from answer basic information are listed as follows.

- Answer length. In general, the text length of an answer is positively related to the information it contains.
- Average sentence length. The expression power is different between long sentences and a short one, so this feature could reflect answer quality at a certain degree.
- Punctuation ratio. Nowadays more and more Internet citizens like using punctuations to represent their emotions, but too much punctuations may have negative effects on answer expressions.
- Terms ratio. It is indicating an incorrect usage if too many single words occurred after word segmentation [8]. Therefore, this feature can reflect if the author's writing style has problems.
- Good format. Obviously, it is easier to read if an answer was writing in a good format.

Answer diversity features are as follows.

- Sentiment. Many questions in online QA communities are regarding to trending news and topics, and in these cases, the sentiment orientation of the answer writer may have critical effects on the adoption of the answer.
- References. It can greatly increase the credibility of the information delivered by the answer if the answer explicitly indicates the source it referenced.
- Images. As the old saying goes "seeing is believing", an image can sometimes give us more powerful information than a paragraph of text.
- External links. To have a clear thought on professional cases, some pilot information may be required, so the answer writer often add some external links to guide answer viewers to the knowledge they needed.
- Text entropy. The entropy based on characters and terms can reflect the diversity of a text [9]. For a text T, with length of λ and n different terms, we can get its text entropy by (1), where p_i, i =1,2, ... , n is the frequency of term i in text T.

$$\mathrm{E}_T(p_1, p_2, ..., p_n) = \frac{1}{\lambda}\sum_{i=1}^{n} p_i(\log\lambda - \log p_i) \quad (1)$$

The last one is about the correlation between answers and questions.

- Topic similarity between questions and answers. An answer of high quality usually have the same topic with the question it belongs to. We use the TF-IDF algorithm to get topics of

answers and questions separately, and then calculate the topic similarity score through topic vectors.

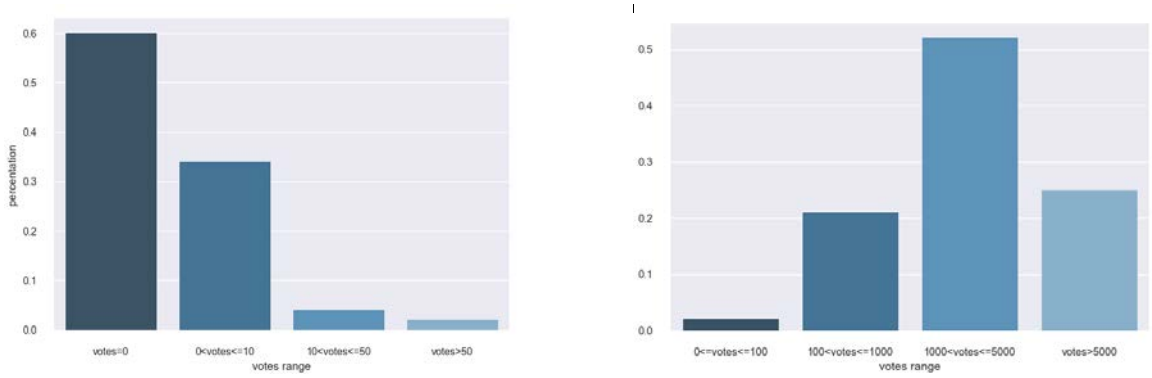
Model Training

In the default sorting mechanism on Zhihu website, a high-quality answer can easily get vote from viewers, and the more votes an answer get the more top it will be listed. But this is not meaning that a high-quality answer can get numerous votes and an answer with few votes is must of low quality. We use manual labeling method to annotate answers' quality. Label 1 means high quality while 0 means low quality. Table 2 shows the labeling results about a sample data which we selected 5000 data randomly from each category.

Tab.2. Summary of labeling results

Label	Avg. votes	Std. votes	Max. votes	Min. votes
0(low-quality)	8.99	131.44	6263	0
1(high-quality)	4532.66	6600.11	105156	11

To verify the manual labeling result, we compared it with the actual votes amount, and it shows that answers with label 1 have much more votes than the lable-0 answers, which means the manual labeling result is trustworthy. Figure 1 shows the comparison of the votes distribution in each category.



(a) votes distribution of class 0 (b) votes distribution for class 1

Fig.1. Votes distribution of each class

The classification problem presents a class imbalance. The number of high-quality answers is much less than answers of low quality. Experiments in paper [10] shows that the random forest algorithm outperforms other models such as C4.5, k-NN, SVM, LDA etc. in imbalanced classification problems, so we choose the random forest algorithm as the classification model to fit the problem proposed above. The model inputs are the features we extracted in the third section, and descriptive statistics of each variable is shown in Table 3.

Tab.3. Descriptive statistics

feature name	feature code	Avg.	Std.	Min.	Max.
Answer text length	ans_length	0.07	5.54	0.69	11.5
Average sentence length	avg_sentence_length	0.11	3.33	0.00	7.82
Punctuation ratio	punctuation_ratio	0.00	0.11	0.00	1.00
Terms ratio	terms_ratio	0.05	0.32	0.00	1.00
Good format	well_formed	0.13	0.69	0.00	1.00
Sentiment	sentiment	0.07	0.48	-1.00	1.00
References	has_reference	0.06	0.01	0.00	1.00
Images	img_num	0.22	1.63	0.00	6.88
External links	href_num	0.19	0.74	0.00	6.39
Text entropy	entropy	0.13	-4.37	-9.43	-1.2
Topic similarity	qa_similarity	0.00	-1.44	-4.96	0.00

Results

To evaluating the performance of a classifier, we should take percentage of correctly classifier (PCC), sensitivity, specificity, and receiver operating characteristic curve (ROC curve) etc. into consideration [11].

In binary classification problems, we can get a confusion matrix as shown in Table 4.

Tab.4. Confusion matrix

	Positive(Actual)	Negative(Actual)
Positive(Predicted)	True Positive(TP)	False Positive(FP)
Negative(Predicted)	False Negative(FN)	True Negative(TN)

We can easily get the performance indicators by using the contents above.

- PCC is a description of systematic errors, we can get it by (2).

$$PCC = (TP + TN) / (TP + FN + FP + TN) \quad (2)$$

- Sensitivity, aka the true positive rate, measures the proportion of positives that are correctly classified. we can get this measurement by (3).

$$sensitivity = TP / (TP + FN) \quad (3)$$

- Specificity, aka the true negative rate, measures the proportion of negatives that are correctly classified. The calculation is defined by (4).

$$specificity = TN / (TN + FP) \quad (4)$$

- ROC curve, is a plot that illustrate the performance of a binary classifier. It is created by plotting the true positive rate against the false positive rate at various threshold settings.

As is shown in Table 5 and Figure 2, the classification model we propose can perform an accurate classification with satisfactory results.

Tab.5. Performance indicators of the model

accuracy	sensitivity	specificity	precision	recall	F1 score
0.86	0.87	0.89	0.86	0.87	0.86

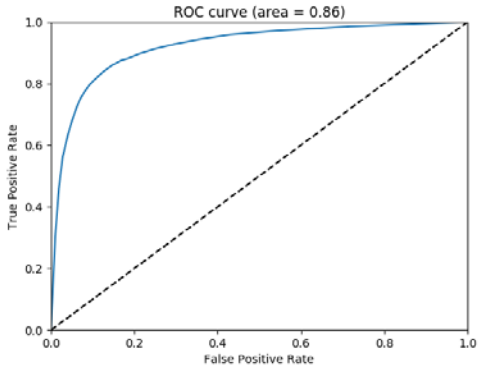


Fig.2. ROC curve of the model

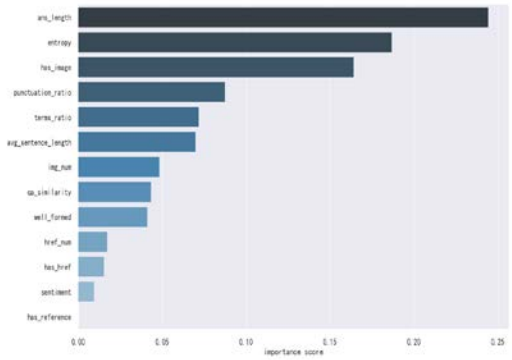
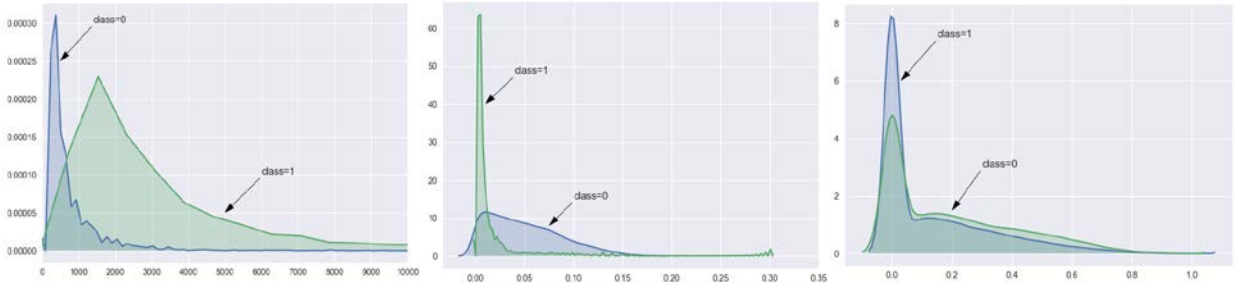


Fig.3. Feature importance of the model

The random forest classifier uses an implicit procedure to select features. It has a superior performance on high-dimensional datasets [12]. The process of implicit feature selection is represented by Gini importance. The feature importance is calculated by the Gini impurity of each node in the binary tree, which reflects the separation degree of two classes at each node [13].

The feature importance score of the model we propose is shown in Figure 3. The score of the answer length is much higher than others, the second one is the text entropy, and the followings are at the same level while the features of sentiment and references are in a lower degree, which means these features can hardly distinguish the two classes. To see more details of the distributions of the features, we get the Figure 4.

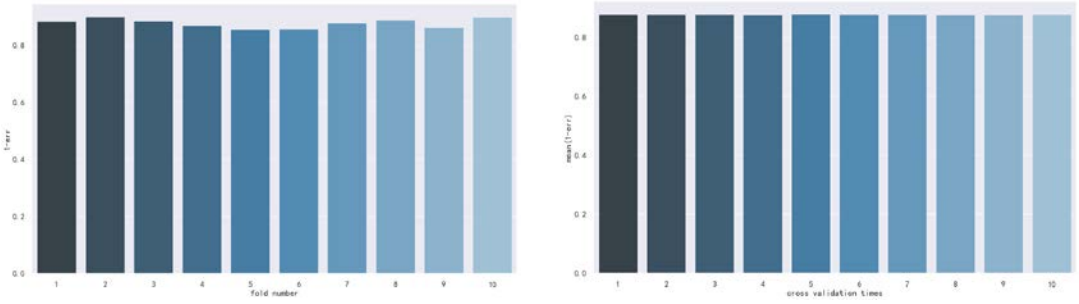


(a) answer length distribution (b) entropy distribution (c) topic similarity distribution

Fig.4. Distributions of some features

Robustness Test

Cross validation, aka rotation estimation, is a method to evaluate the generalization ability of a model raised by Seymour Geisser [14]. The goal of cross validation is extracting part of data as the testing set, to avoid problems like overfitting. There are several methods to perform cross validation, leave-p-out cross-validation, leave-one-out cross-validation, k-fold cross-validation etc. In binary classification cases, the misclassification error rate is used to evaluate the fitting. We use 10-fold cross validation method to give our model a robustness test. The results are shown in Figure 5(a).



(a) Single 10-fold cross validation (b) Ten times 10-fold cross validation

Fig.5. Cross validation results

To evaluate the model more accurately, we perform 10 times of 10-fold cross validation. The mean score of each validation is shown in Figure 5(b).

The results show that the quality evaluation model has a stable performance in both single and multiple cross validation tests, thus indicating that the model has good robustness.

Conclusion

We take Zhihu as the research object, after manually label 33907 answers on their quality, we find that an answer of high quality can get more votes. We extract some features from three aspects, answer basic information, diversity and the correlation with questions. By using these features and the labeled dataset, we construct a classification model based on the random forest algorithm, which has a superior performance on answer quality classification with the accuracy of more than 86%. The robustness testing shows that this model has stable performance, so it is worthy to popularize.

This paper is not free of deficiencies. We define the answer quality as a binary classification problem, but we cannot distinguish the differences among the answers with positive labels. The social attributes of community are not taken into consideration, which may be a critical factor of the answer quality problem.

Acknowledgement

This work is partially supported by the Natural Science Foundation of China [71371144, 71601119, 71601082].

References

[1] Yao, Y., Tong, H., Xie, T., & Akoglu, L. (2015). Detecting high-quality posts in community question answering sites. Information Sciences, 302, 70-82.

[2] Burgess, S., Sellitto, C., Cox, C., & Buultjens, J. (2009, June). User-generated content in tourism: Benefits and concerns of online consumers. In ECIS (pp. 417-429).

[3] Kim, S., Oh, J. S., & Oh, S. (2007). Best‐answer selection criteria in a social Q&A site from the user‐oriented relevance perspective. Proceedings of the American Society for Information Science and Technology, 44(1), 1-15.

[4] Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008, April). Predictors of answer quality in online Q&A sites. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 865-874). ACM.

[5] Shah, C., & Pomerantz, J. (2010, July). Evaluating and predicting answer quality in community QA. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (pp. 411-418). ACM.

[6] Jeon, J., Croft, W. B., Lee, J. H., & Park, S. (2006, August). A framework to predict the quality of answers with non-textual features. In Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 228-235). ACM.

[7] Agichtein, E., Castillo, C., Donato, D., Gionis, A., & Mishne, G. (2008, February). Finding high-quality content in social media. In Proceedings of the 2008 international conference on web search and data mining (pp. 183-194). ACM.

[8] Yang-sen, Z., Da-yuan, C., & Shi-wen, Y. (2006). A Hybrid Model of Combining Rule-based and Statistics-based Approaches for Automatic Detecting Errors in Chinese Text. Journal of Chinese Information Processing, 20, 1-7.

[9] Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446-3453.

[10] Martens, D., Vanthienen, J., Verbeke, W., & Baesens, B. (2011). Performance of classification models from a user perspective. Decision Support Systems, 51(4), 782-793.

[11] Breiman, L. (2004). Consistency for a simple model of random forests.

[12] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[13] Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., & Bachert, P. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC bioinformatics, 10(1), 1.

[14] Geisser, S. (1993). Predictive inference (Vol. 55). CRC press.