Automated Evaluation of User-Generated Answer Quality in Social Q&A Communities: A Case Study of Zhihu Postprint

Authors: Guo Shunli, Zhang Xiangxian, Tao Xing, Zhang Liman

Date: 2023-07-26T00:00:00+00:00

Abstract

[Purpose/Significance] Aims to construct a quality evaluation index system for user-generated answers in social Q&A communities, achieve automated evaluation and screening of answer quality oriented towards user needs, and enhance the knowledge service quality of social Q&A communities. [Method/Process] Introduces social-emotional features and user characteristics, employs factor analysis and structural equation modeling to empirically construct a quality evaluation index system for user-generated answers. An automated answer quality evaluation method is designed based on the GA-BP neural network model. Finally, data from the Zhihu website is selected to conduct applied research on the user-generated answer quality evaluation index system and the automated evaluation method. [Results/Conclusion] Constructs an evaluation index system comprising five dimensions: answer text features, answerer characteristics, timeliness features, user features, and social-emotional features. Experimental analysis reveals that the GA-BP neural network-based automated answer quality evaluation method demonstrates higher accuracy and lower average error compared with other methods, possesses feasibility and effectiveness, and can be further applied and promoted in practice.

Full Text

Preamble

Vol. 63 No. 11, June 2019
ChinaXiv Cooperative Journal

Research on Automated Evaluation of User-Generated Answer Quality in Social Q&A Communities—Taking "Zhihu" as an Example

Guo Shunli[1], Zhang Xiangxian[2], Tao Xing[2], Zhang Liman[2]
[1] School of Media, Qufu Normal University, Rizhao 276826
[2] School of Management, Jilin University, Changchun 130022

Abstract

[**Purpose/Significance**] This study aims to construct an evaluation index system for user-generated answer quality in social Q&A communities, enabling automated evaluation and screening of answer quality oriented toward user needs, thereby improving knowledge service quality in social Q&A communities. [**Method/Process**] The research introduces social-emotional features and user characteristics, employing factor analysis and structural equation modeling to empirically construct a user-generated answer quality evaluation index system. An automated answer quality evaluation method is designed based on the GA-BP neural network model. Finally, data from the Zhihu website is selected to conduct an applied study of the user-generated answer quality evaluation index system and automated evaluation method. [**Result/Conclusion**] The study constructs an evaluation index system comprising five dimensions: answer text characteristics, respondent characteristics, timeliness characteristics, user characteristics, and social-emotional characteristics. Experimental analysis reveals that the GA-BP neural network-based automated answer quality evaluation method achieves higher accuracy and lower average error compared to other methods, demonstrating feasibility and effectiveness that warrant further application and promotion in practice.

Classification Number: G203
Keywords: Social Q&A Community; User-Generated Answers; Quality Evaluation; User Demand

In recent years, social Q&A communities have developed rapidly, with registered user numbers showing exponential growth. Taking Zhihu as an example, since opening public registration in 2013, by the end of November 2018, Zhihu officially announced that registered users had exceeded 220 million, with over 30 million questions and 130 million answers. Social Q&A communities have evolved into diversified, well-established large-scale knowledge-sharing platforms, becoming an important channel for people to obtain information and knowledge daily. However, social Q&A communities are characterized by their social and open nature, with questions and answers primarily generated through user participation. Any user can freely ask and answer questions, resulting in uneven quality of user-generated answers. Moreover, due to limitations in their own experience and cognition, question-asking users may not necessarily select the best answers; some may even be malicious advertisements or false information. This forces users to expend significant time and effort searching for, identifying, and acquiring knowledge in social Q&A communities, leading to a phenomenon of "knowledge overload and disorientation," reducing the efficiency of knowledge seeking and acquisition, and resulting in poor user experiences. Additionally, as the user base of social Q&A communities expands and the volume of user-

generated questions and answers increases, manual quality evaluation becomes difficult and inefficient. Relying solely on manual review or annotation cannot adequately address current quality issues in social Q&A communities. Therefore, automated evaluation of user-generated answer quality has become an urgent problem for social Q&A community operations.

Although scholars both domestically and internationally have conducted extensive research on answer quality evaluation in Q&A communities, a unified evaluation standard has yet to emerge.

1. Current Research Status at Home and Abroad

1.1 Feature Selection for Answer Quality Evaluation in Q&A Communities

Internationally, S. Kim et al. [1] studied the criteria for best answer selection in Yahoo! Answers from a user-oriented relevance perspective, finding that users consider social-emotional factors, content, and utility-related evaluation standards when selecting and adopting best answers, with evaluation standards varying across different topics. D. Ishikawa et al. [2] constructed a 12-dimensional Q&A community answer quality evaluation index system including respondent experience, evidence sources, politeness, detail level, opinions, relevance, specificity, and comprehensiveness. S. Oh et al. [3] selected 10 indicators—information accuracy, completeness, relevance, source reliability, respondent empathy, objectivity, readability, politeness, confidence, and respondent effort—as measures of answer quality, comparing evaluation differences across different professional groups. P. Fichman [4] evaluated Q&A community answer quality from three aspects—accuracy, completeness, and verifiability—finding that some non-mainstream Q&A websites also had high-quality answers, with answer quality being less related to the platform itself. A. Y. K. Chua et al. [5] examined the relationship between answer speed and answer quality, discovering significant differences between answer quality and speed across different question types, with the highest-quality answers demonstrating better overall quality than the fastest answers.

Domestically, scholars have primarily selected feature indicators from answer text and non-text perspectives to construct answer quality evaluation index systems. Sun Xiaoning et al. [6] empirically constructed a social search answer quality evaluation model from four dimensions: content quality, contextual quality, source quality, and emotional quality. Li Xiangyu et al. [7] combined expert scoring with triangular fuzzy weighted average G1 method to construct an SQA platform answer quality evaluation index system, confirming its scientific validity. Zhang Yuxuan [8] applied signaling theory from a user perspective, finding that seven types of external cues—including information utilization cues, information recognition cues, information reporting cues, information negation cues, information capability cues, information appearance cues, and system recommendation cues—influence users' perceived judgment of information quality in

social Q&A communities, proposing an external cue-based information quality perception model for social Q&A platforms. Jiang Wen et al. [9] introduced emotional features into online Q&A community information quality evaluation, assessing quality from four dimensions: text features, user features, timeliness features, and emotional features. Yuan Hong et al. [10] constructed a three-dimensional Q&A community answer quality evaluation index system from the definition of information quality: answer form, answer content, and answer utility. Kong Weize et al. [11] evaluated Q&A community answer quality from perspectives of text features, temporal features, link features, question granularity features, and Baidu Knows community user characteristics. Luo Yi et al. [12] introduced the new RIPA theory, arguing that three indicators—completeness, professionalism, and authority of user-generated content—are key factors affecting answer quality on social Q&A platforms.

1.2 Research on Answer Quality Evaluation Methods in Q&A Communities

Currently, scholars generally treat answer quality evaluation as a machine learning-based classification problem [13], applying machine learning methods to Q&A community answer quality evaluation, such as maximum entropy, support vector machines, decision trees, random forests, logistic regression, and neural networks. Some scholars employ traditional evaluation methods like hierarchical analysis and fuzzy comprehensive evaluation, while others conduct manual annotation-based evaluations using combined manual and automated approaches. Internationally, to improve best answer discovery and prediction accuracy, some scholars treat answer quality evaluation as a classification problem, enhancing precision and recall through improved classification algorithms. For example, J. Jeon et al. [14] proposed a Q&A community answer quality prediction method based on non-textual features, empirically demonstrating significant improvement over basic feature-based prediction. C. Shah et al. [15] used Yahoo! Answers as a case study, first employing manual annotation to evaluate answer quality for given questions, then extracting various features of questions, answers, and users to train classifiers for best answer selection.

Domestically, Li Chen et al. [16] designed and implemented a Q&A quality classifier based on feature sets by extracting textual and non-textual features according to given Q&A quality criteria. Wang Wei et al. [17] introduced structured features, text features, and user social attributes into a Chinese Q&A community answer quality evaluation feature system, then selected three evaluation methods—logistic regression, support vector machines, and random forest—combined with three newly designed features and classical text/link features to classify high-quality and non-high-quality answers. Cui Minjun et al. [18] extracted four types of features—text, non-text, language translatability, and link count in answers—based on question types, using logistic regression algorithms to evaluate answer quality for various question types. Hu Haifeng [19] studied feature representation and fusion of textual and non-textual information for

user-generated answer quality evaluation.

1.3 Literature Review Summary

Reviewing existing research reveals that current studies primarily employ single or multiple feature indicator combinations to construct user-generated answer quality evaluation index systems. However, these systems suffer from incompleteness, lack of unified standards, subjectivity and ambiguity in some indicators, and difficulty in quantification and judgment. Few studies consider the impact of user social-emotional features on answer quality evaluation, nor do they account for individual differences such as user needs, interests, and cognitive levels, lacking personalized evaluation index systems oriented toward user needs. Scholars have treated answer quality evaluation as a machine learning classification problem, employing methods such as SVM, stochastic gradient boosting, decision trees, maximum entropy, logistic regression, Bayesian, and J48, all achieving good experimental results. Although numerous studies address automated answer quality evaluation, few scholars have used neural network methods for evaluation or compared their effectiveness and accuracy differences with other methods.

Given these gaps, this study combines previous research findings to construct a user-generated answer quality automated evaluation index system from a user needs perspective, attempting to solve problems of indicator ambiguity, incompleteness, and lack of personalization. Treating answer quality automated evaluation as a machine learning problem, we select the typical machine learning method—genetic algorithm optimized BP neural network model—to conduct empirical applied research based on the constructed user-generated answer quality automated evaluation index system, proposing an automated evaluation method for user-generated answers in social Q&A communities.

2. Construction of User-Generated Answer Quality Evaluation Index System

2.1 Initial Selection of Evaluation Indicators

Referring to literature [13] on answer quality evaluation indicators, this study posits that users are influenced by multiple factors when evaluating answer quality, generally considering answer text content quality, respondent quality, and timeliness—three categories of features whose impact on answer quality has been confirmed by most research. However, as an open social website, social Q&A community users also consider other users' evaluations of answer quality (such as likes, forwards, comments, etc.) when screening and evaluating answers. They are easily influenced by interpersonal relationships, community opinion leaders, and interactive communication. Opinion leaders in Q&A communities can influence other users' cognition, and their answers can gain more support and agreement from followers [17]. Moreover, respondents' emotional attitudes and positivity levels also affect users' answer adoption. Therefore, this study

introduces users' social-emotional attitude features toward answers into answer quality evaluation. Additionally, different users in social Q&A communities are affected by individual differences in cognition, needs, and interests, holding different standards and requirements for answer quality evaluation. Thus, user characteristics must also be considered in the answer quality evaluation process to make screened answers more aligned with personalized user needs.

Consequently, this study introduces user social-emotional and user characteristics into answer quality evaluation, dividing user-generated answer quality evaluation indicators into five dimensions: answer text characteristics, respondent characteristics, timeliness characteristics, user characteristics, and social-emotional characteristics. Through extensive reading and review of information quality evaluation literature, and based on theoretical frameworks such as the Information Systems Success Model, Uses and Gratifications Theory, and Data Quality Framework [20], 24 evaluation indicators were initially selected, as shown in Table 1 .

After initially selecting social Q&A community user-generated answer quality evaluation indicators, the authors employed expert interviews to revise relevant expressions, focusing on listening to experts' opinions regarding indicator rationality and completeness. Discussions explored whether the dimensional division and indicator selection were reasonable, whether indicator names were appropriate, and whether ambiguity or difficulty in measurement existed, eliminating indicator ambiguity and vagueness to achieve preliminary standardized screening. Based on expert suggestions and feedback, the user-generated answer quality automated evaluation indicators were formed, as shown in Figure 1 [Figure 1: see original paper]. Specific revisions were as follows:

(1) Deletion of evaluation indicators. The indicator "Number of questions asked by respondent" under respondent characteristics was deleted because it reflects respondent needs rather than their ability and experience in generating answers, having insufficient impact on answer quality. Three indicators under answer text characteristics were deleted: "Stopword count," "Paragraph count," and "Question-answer coupling degree." The stopword count in answer text should be answer text length minus keyword count, creating indicator redundancy. "Paragraph count" in answer text characteristics cannot reflect answer quality and has minimal impact. Additionally, "Question-answer coupling degree" duplicates "User preference-answer coupling degree," as user questions directly reflect user needs and preferences. Therefore, "Stopword count," "Paragraph count," and "Question-answer coupling degree" were deleted from answer text characteristics. The indicator "User education level" under user characteristics was deleted because it has minimal impact on users' judgment of answer text quality.

(2) Supplement of evaluation indicator factors. The indicator "Number of images or animations" was added to answer text characteristics. In the mobile Internet environment, many users prefer understanding knowledge

through images or animations, which contain large amounts of information and facilitate comprehension. Therefore, the number of images or animations in user-generated answers affects answer quality. The indicator "Match degree between professional field and question" was added to respondent characteristics to reflect respondents' professionalism and familiarity with the question domain.

2.2 Empirical Analysis of User-Generated Answer Quality Evaluation Index System

2.2.1 Questionnaire Design and Distribution The questionnaire primarily measured user-generated answer quality in social Q&A communities, using declarative sentences to express the feasibility and rationality of each evaluation indicator for answer quality measurement. A total of 610 questionnaires were collected through online and field distribution, with 580 valid questionnaires obtained. The collected questionnaires were randomly divided into two parts (290 samples each) for Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA).

Reliability testing of the survey sample data revealed that Cronbach's alpha values for all five dimensions exceeded 0.8, with overall sample reliability at 0.846, indicating good data reliability. However, after deleting indicators "Question-answer length ratio C5" and "Follow relationship C20," the reliability of answer text characteristics dimension B1 and social-emotional dimension B5 improved significantly, as did overall questionnaire reliability. This suggested these indicators failed reliability testing and should be deleted. KMO and Bartlett's sphericity tests were then conducted, revealing significant approximate chi-square values in Bartlett's test, indicating common factors in the correlation matrix and good overall sample validity suitable for further factor analysis.

2.2.2 Exploratory Factor Analysis This study employed principal component analysis for EFA. When five common factors were extracted, the cumulative variance contribution rate reached 55.051%. Maximum variance rotation was applied, converging after 10 iterations, yielding the rotated factor matrix shown in Table 2 .

Table 2 shows that evaluation indicator variable "Emotional feature word count C15" had similar loading factors on common factors 2, 4, and 5, with insignificant differences and poor validity, warranting deletion. Common factor 1 explained five indicator variables (C1, C2, C3, C4, C6), corresponding to all indicators of answer text characteristics dimension. Common factor 2 explained two indicator variables (C13, C14), corresponding to all indicators of user characteristics dimension. Common factor 3 explained four indicator variables (C7, C8, C9, C10), corresponding to all indicators of respondent characteristics dimension. Common factor 4 explained four indicators (C16, C17, C18, C19), corresponding to four indicators of social-emotional dimension excluding C15. Common factor 5 contained only two indicators (C11, C12), corresponding to

timeliness dimension indicators. This aligned with our preliminary dimensional assumptions, confirming the rationality of dividing user-generated answer quality evaluation indicators into five dimensions, with further refinement based on CFA results.

2.2.3 Confirmatory Factor Analysis Structural Equation Modeling software AMOS 17.0 was used for Confirmatory Factor Analysis (CFA). Another portion of sample data (290 copies) was used to further test indicator validity, with 17 observed variables, 5 latent variables, and 17 residual variables configured. The maximum likelihood estimation method was employed to estimate loading coefficients between observed variables and their corresponding latent variables, as shown in Table 3 .

According to general empirical rules, if the absolute C.R. value exceeds 2.58, the model's parameter estimates reach the 0.01 significance level, with path coefficients receiving data support; when P-values are less than 0.001, "***" is displayed, indicating the model reaches the 0.001 significance level [34]. Table 3 shows that in the evaluation index system significance test, the standardized loading factor estimate for "Disagree count C18" was less than 0.5, indicating the indicator failed validity testing and should be deleted. Model fit indices provided by AMOS were then used to evaluate the rationality of the constructed evaluation index system. According to various indicator testing standards, all test results fell within acceptable ranges, indicating the overall evaluation index system basically met testing requirements. After deleting observed variable C18, the model's absolute fit index ^{2} value decreased from 186.125 to 150.568, and the CMIN/DF value decreased from 1.708 to 1.602, indicating improved absolute fit performance. This further confirmed that deleting indicator C18 enhanced the rationality of the constructed evaluation index system.

2.3 Revision and Establishment of Evaluation Indicators

Following empirical analysis using both EFA and CFA, and considering test results comprehensively, the indicator "Question-answer text length ratio C5" in answer text dimension was deleted due to failing reliability testing and having redundancy with "Text length C1." The indicator "Follow relationship C20" also failed reliability testing and was deleted. For the social-emotional dimension, "Emotional feature word count C15" showed poor validity during principal component analysis with similar loadings across multiple common factors, correlating with "Respondent emotional attitude tendency," warranting deletion. Additionally, indicator "Disagree count C18" had loading coefficients less than 0.5, failing significance testing, and its deletion significantly improved model fit and loading coefficients for other indicators in the same dimension. In summary, the final selected social Q&A community user-generated answer quality automated evaluation indicators include 5 dimensions and 16 indicators, as shown in Figure 2 [Figure 2: see original paper].

3. Automated Evaluation of User-Generated Answer Quality Based on GA-BP Neural Network

3.1 Acquisition and Quantification of Evaluation Indicators

Automated evaluation of user-generated answer quality in social Q&A communities requires algorithmic implementation and automatic quantification of evaluation indicators. The GoSeeker web scraping software was used to directly collect data and text information. Tools and methods such as jieba segmentation, HowNet emotional dictionary, and text processing techniques were employed, with Python and Matlab programs written to implement indicator statistics and quantification.

(1) **Text characteristic dimension indicators acquisition and quantification.** This includes: Text length: directly quantified using character count of answer text. Within a certain threshold, longer answer text generally indicates richer, more complete information and better satisfaction of user knowledge needs, implying higher answer quality. Keyword count: quantified using word frequency statistics of answer text excluding stopwords. Total word frequency minus stopword frequency yields keyword count. Sentence count: quantified by counting occurrences of sentence-ending punctuation marks (periods, question marks, etc.). External link count: refers to reference sources and answer expansion links appearing in answer text, quantified by counting hyperlinks. Image/animation count: directly obtained by counting images or animations in answer text.

(2) **Respondent characteristic dimension indicators acquisition and quantification.** This includes: Best answer count: quantified by the number of answers adopted as best answers among all respondent's answers, or by best answer adoption rate. In Zhihu, since no explicit best answer designation exists, the number of answers featured in Zhihu Daily or Zhihu Roundtable is used for quantification. Total answers count: quantified by the total number of questions answered by the user, indicating experience and participation enthusiasm. User authority: quantified directly by respondent's user level or points, where higher levels/points indicate higher community recognition and influence. Professional field-question match degree: coded as 1 if professional field matches question domain, 0 otherwise.

(3) **Timeliness dimension indicators acquisition and quantification.** This includes: Answer relative response order refers to the sequential position of the current answer among all answers to the same question, sorted by response time. Quantified as:

$$\text{Answer relative response order} = \frac{\text{Answer response time order}}{\text{Total number of answers}}$$

Answer-question generation interval can be quantified by the day difference

between answer date and question date. To avoid bias from excessively large values, grouping methods are applied. After questionnaire surveys and interviews, the day difference ranges and corresponding 10-point scale values are shown in Table 4 .

(4) **User dimension indicators acquisition and quantification.** This includes: Asker question count: generally available in user profiles on social Q&A communities, directly quantified through web scraping. For example, Zhihu user profiles include "question count" information. User preference-answer coupling degree: quantified using similarity between answer text vector and user preference vector. Since questions most directly reflect user knowledge needs, similarity between question and answer text can be used for measurement.

(5) **Social-emotional dimension indicators acquisition and quantification.** This includes: Agree count and comment interaction can be directly quantified through web scraping. Respondent emotional attitude includes three polarities: positive, negative, and neutral, quantified by counting emotional words in answer text based on an emotional baseline dictionary. Comment interaction quantity is quantified by the number of comments under answers.

3.2 User-Generated Answer Quality Evaluation Based on GA-BP Neural Network

The BP neural network is a neural network trained using error backpropagation, the most widely applied artificial neural network algorithm, comprising three layers: input, hidden, and output. Standard BP neural networks learn through supervised learning, employing gradient descent on error functions to minimize mean square error between actual and expected outputs [35]. Although widely applied, BP neural networks suffer from easily falling into local minima, inability to guarantee convergence to global minima, slow convergence speed, and long training times. However, Genetic Algorithms (GA) employ probabilistic optimization methods, automatically obtaining and guiding optimization search spaces, adaptively adjusting search directions without requiring deterministic rules, possessing strong global search capabilities and optimization performance [36]. Genetic algorithms offer good global search ability, easily obtaining global optimal solutions, effectively overcoming BP algorithm's local optima defects while optimizing BP neural network initial weights and thresholds. Therefore, selecting genetic algorithms to optimize BP neural networks (hereafter "GA-BP neural network") can accelerate BP neural network convergence, improving prediction accuracy and stability.

User-generated answer quality in social Q&A communities is influenced by 16 feature factors across five dimensions, making its automated evaluation result difficult to express with mathematical formulas—a typical nonlinear problem. However, as a multi-layer feedforward network, BP networks possess powerful

nonlinear mapping capabilities, capable of simulating and analyzing nonlinear relationships among the five dimensions and 16 evaluation indicators, enabling nonlinear classification and prediction. After repeated learning and training, they can sufficiently approximate any complex nonlinear relationship. Moreover, GA-BP neural network algorithms have been widely applied in other fields with fruitful results, providing sound theoretical and practical foundations that make social Q&A community user-generated answer quality evaluation more objective and rational. Therefore, this study employs genetic algorithm-improved BP neural networks for automated evaluation of user-generated answer quality in social Q&A communities. Before training the BP neural network, genetic algorithms are used to optimize initial weights and thresholds, narrowing the search range before applying BP neural network algorithms for automated evaluation.

The GA-BP neural network-based social Q&A community user-generated answer quality evaluation process is shown in Figure 3 [Figure 3: see original paper].

3.2.1 Indicator Feature Extraction, Quantification, and Normalization First, the GoSeeker web scraping software was used to automatically crawl data, extracting each evaluation indicator feature using the quantification methods described in Section 3.1. Since extracted sample data had indicators with different magnitudes and significant gaps between them, direct application in GA-BP neural network computation could lead to similarly scaled network weights, making the constructed network overly "sensitive." To ensure BP neural network training speed and accuracy, and avoid errors from excessively large or small data, collected data must be normalized. This study uses S-type functions as activation functions, with value domains limited to [-1, 1], requiring sample data normalization to the interval [-1, 1]. The premnmx function was adopted for normalizing extracted sample data, as shown in formulas (1) and (2):

$$p_n = 2\frac{t - \min t}{\max t - \min t} - 1 \quad (1)$$

$$t_n = 2\frac{p - \min p}{\max p - \min p} - 1 \quad (2)$$

In formulas (1) and (2), p and t represent input and output data values, min p and max p represent input data minimum and maximum values, and min t and max t represent output data minimum and maximum values.

3.2.2 BP Neural Network Initialization

(1) **Input and output layer determination.** Kolmogorov's theorem [37] proves that a BP neural network with one hidden layer can approximate

any mapping relationship with arbitrary precision. Therefore, to simplify model complexity and improve GA-BP neural network learning speed and efficiency, this study configured the social Q&A community user-generated answer quality evaluation model as a 3-layer network structure with only one hidden layer. The five dimensions and 16 indicators of social Q&A community user-generated answer quality evaluation serve as GA-BP neural network input layer, resulting in 16 input layer neurons. The output layer result reflects user-generated answer quality level, so the output layer has 1 neuron.

Some studies manually annotate answer quality levels as output variables, dividing quality into five levels: very low, low, medium, high, and very high. However, manual annotation may differ from actual user needs. To reflect differences in user information needs, this study defines user-selected best answers as the highest level, then calculates similarity between other answer texts and best answers to classify answer quality levels, as shown in Table 5 . If no best answer exists, the answer with the most agree/support votes is selected as the best answer.

(2) **Hidden layer node determination.** This study uses the trial-and-error method to determine BP neural network hidden layer nodes. With other network parameters held constant, the same sample set is trained repeatedly by adjusting hidden layer node numbers, selecting the node count that yields the minimum MSE as optimal. Formula (3) provides an initial estimate, with trial-and-error determining the final optimal hidden layer node count.

$$n_l = \sqrt{n+m} + a \quad (3)$$

In formula (3), n_l represents hidden layer node count, n represents input layer node count, m represents output layer node count, and a is a constant between 1-10.

(3) **Initialization function setting.** BP neural network functions include transfer functions, learning functions, and performance functions. Transfer functions typically use S-type logarithmic or tangent functions. Since this study normalizes input/output data to [-1, 1], suitable for Sigmoid tangent function value ranges, the hidden layer transfer function tansig and output layer transfer function logsig were selected. Momentum gradient descent was adopted as the network training method, with learning function learngdm and training function trainingdm.

(4) **Initialization weights and thresholds.** Genetic algorithm optimization of BP neural network initial weights and thresholds proceeds as follows:

Individual encoding and initial population generation. Real-number encoding was used for individuals. The encoding string comprises hidden-input

layer connection weights, output-hidden layer connection weights, hidden layer thresholds, and output layer thresholds. Network weights and thresholds are concatenated in sequence to form a real-number array serving as a genetic algorithm chromosome. Encoding length is shown in formula (4). Within weight and threshold ranges, M chromosomes form the initial population. Since population size significantly impacts genetic algorithm global search performance, appropriate population sizes must be selected based on specific problems.

$$S = n \times n_l + n_l \times m + n_l + m \quad (4)$$

In formula (4), S represents population size, n_l represents hidden layer node count, n represents input layer node count, and m represents output layer node count.

Fitness function setting. Genetic algorithm evolutionary search relies on fitness functions, using fitness values of each chromosome in the population for searching. Individuals with higher fitness values have greater probability of being inherited by the next generation. The fitness function is set as the reciprocal of BP neural network error. When this fitness function reaches its maximum, BP neural network weights and thresholds are optimized, as shown in formula (5):

$$f(i) = \frac{1}{\text{MSE}} \quad (5)$$

In formula (5), f(i) represents the fitness value of chromosome i; MSE represents the sum of squared errors between BP neural network predicted output and expected output.

Individual selection. Selection operation uses ranking method, sorting individuals by fitness value from smallest to largest, with the smallest fitness individual assigned rank 1 and the largest assigned rank M. Individual selection probability is then calculated using fitness proportionate selection method, as shown in formula (6):

$$p_i = \frac{f_i}{\sum_{i=1}^{m} f_i} \quad (6)$$

In formula (6), f_i represents individual i's fitness value; m represents population individual count.

Crossover and mutation operations. Crossover uses single-point crossover, with the optimal individual copied directly to the next generation without crossover. Other individuals undergo crossover with probability p_c to produce two new individuals. Similarly, the optimal individual is copied directly

without mutation. Mutation uses uniform mutation, with other individuals mutated with probability p_m to produce new individuals. Current population fitness values are then calculated to identify the optimal individual, iterating until conditions are met.

3.2.3 GA-BP Neural Network Training Process The GA-BP neural network training process for social Q&A community user-generated answer quality evaluation involves comparing actual output answer quality level y with expected quality level Y. If they differ, errors are calculated using relevant error formulas, then error signals are backpropagated along original paths. Different sample data are used for learning and training to obtain weight coefficients between input-hidden and hidden-output layers, making MSE values progressively smaller until errors fall below the set threshold or maximum training iterations are reached. The GA-BP neural network obtains initial weights and thresholds through training, forming the evaluation model. Test set sample data are then input into this evaluation model for automated evaluation. The output layer produces actual utility value y, which is restored to real values using the postmnmx function to obtain answer quality evaluation results, completing the answer quality evaluation.

4. Applied Research—Taking the "Zhihu" Website as an Example

4.1 Data Collection and Preprocessing

This study selected answers to the question "How to evaluate Huawei Mate 10 & Mate 10 Pro?" on Zhihu as the application object for quality evaluation methods. As of January 20, 2018, this question had 494 answer texts. GoSeeker software was used to collect relevant data on question askers/viewers, respondents, and answer texts. The quantification methods described in Section 3.1 were applied to each indicator. During quantification, since Zhihu lacks explicit user levels and authority, follower count was used for quantification, assuming more followers indicate higher authority. Since Zhihu lacks explicit best answer designation, the number of answers featured in Zhihu Daily and Zhihu Roundtable was used to quantify best answer count, assuming featured answers possess authority and representativeness.

Considering different user needs' impact on answer text quality, 10 Zhihu APP users aged 18-35 were selected as research subjects (User 1-10). From a user perception perspective, answer quality levels were manually annotated using a 10-point scale without predetermined evaluation standards, relying solely on subjective judgment. Since only one question was selected for this applied research, user question count was identical across users and had no impact on output, thus only used for comparative analysis among multiple users. Following the general 80-20 rule for neural network model construction, 400 answers were selected as training samples (answers 1-400) and 94 answers (401-494) as

test samples.

4.2 Analysis of Answer Quality Evaluation Method Application

4.2.1 Comparative Analysis of Different Algorithms The collected data were first analyzed using four methods: standard BP neural network, SVM, maximum entropy, and GA-BP neural network. Matlab 2015a served as the software platform, with neural network toolbox functions and genetic algorithm toolbox programming implementing BP neural network and GA-BP neural network evaluation methods, along with SVM and maximum entropy algorithms. Indicators from text characteristics, respondent characteristics, and timeliness dimensions were selected as baseline features, tested using User 1's data sample. Accuracy (P) and average error (M) measured algorithm accuracy and performance. Accuracy refers to the proportion of correctly evaluated samples in the test set, with classification considered accurate when absolute difference between actual and expected values is within 0.3. Higher accuracy indicates better method accuracy. Average error is represented by the mean of absolute errors across test samples, with smaller values indicating stronger model precision and rationality. Results are shown in Table 6 , demonstrating that GA-BP neural network outperformed other classification algorithms with higher accuracy and lower error, making it applicable for user-generated answer quality evaluation.

4.2.2 Applied Research on GA-BP Neural Network By setting parameters and functions, a GA-BP neural network-based social Q&A community user-generated answer quality evaluation model was constructed to analyze the impact of adding user characteristics and social-emotional features to baseline features. The GA-BP neural network adopted a 3-layer structure. Excluding user question count, the input layer had 15 neurons. Hidden layer node count was determined through trial-and-error, finding MSE minimized with 10 hidden nodes. Learning rate $= 0.01$, maximum training iterations $= 100$, target error $= 0.01$. Answers 1-400 served as training samples, 401-494 as test samples. Genetic algorithm optimization obtained optimal initial weights and thresholds, with parameters: population size $= 40$, maximum generations MAXGEN $= 80$, real-number encoding chromosome length $= 121$, crossover probability px $= 0.2$, mutation probability pm $= 0.1$. The optimized weights and thresholds were then input into the BP neural network for retraining to evaluate test samples for Users 1-10. All 10 users' training samples stopped iterating within 100 steps when target error 0.01 was reached.

Taking User 1 as an example, using GA-BP neural network with baseline + user + emotional features as input, the genetic algorithm found optimal values within 40 generations (Figure 4 [Figure 4: see original paper]). The BP neural network stopped iterating after 11 steps when target error 0.01 was reached (Figure 5 [Figure 5: see original paper]). Figure 6 [Figure 6: see original paper] shows expected versus actual output values for User 1's test samples.

Using GA-BP neural network evaluation with input features of baseline only,

baseline + user features, baseline + social-emotional features, and baseline + user + social-emotional features, accuracy P and average error M for Users 1-10 are shown in Table 7 .

5. Discussion and Analysis

Findings reveal:

(1) Comparative analysis of various evaluation methods demonstrates that GA-BP neural network can be applied to social Q&A community user-generated answer quality evaluation. While its accuracy didn't reach the highest reported in existing research, it significantly outperformed SVM and maximum entropy methods when using our designed features. Figure 5 shows GA-BP neural network converges much faster than standard BP neural network, achieving rapid sample training within 11 steps and reaching target error 100% of the time without falling into local minima or infinite loops, enabling rapid construction of evaluation models. This confirms the method's rationality and scientific validity for social Q&A community user-generated answer quality evaluation, warranting further application and promotion.

(2) Table 5 shows that when using baseline + user features or baseline + social-emotional features, although some user data samples showed only slight accuracy P improvement, average error M decreased substantially, indicating that adding user features brings evaluation values closer to target values. When using baseline + user + social-emotional features, accuracy P improved significantly, with average evaluation accuracy reaching 76.38% and low average error, demonstrating that GA-BP neural network quality evaluation values are closer to user-annotated real values, showing strong simulation capability and practicality. The introduction of social-emotional features and user features improves evaluation accuracy, confirming the rationality and effectiveness of our designed evaluation index system. Moreover, the GA-BP neural network-based user needs-oriented answer quality evaluation method can learn and train according to different user information needs and characteristics, finding intrinsic relationships between inputs and outputs saved as weights in the neural network for continuous self-adaptation and adjustment. This enables personalized quality evaluation system design based on different user information needs, increasing model adaptability and generalizability, forming a flexible, personalized user-generated answer quality evaluation method.

(3) For social Q&A communities, ensuring platform user-generated content quality and providing high-quality knowledge services are driving forces for platform development. Communities should evaluate and screen newly generated answer quality according to different user needs and characteristics, visually presenting high-quality answers to users to promote dissemination of quality answer content. Based on our findings, social Q&A

communities should mine and evaluate quality answer content from perspectives of answer text characteristics, respondent characteristics, timeliness, user characteristics, and social-emotional characteristics. Machine learning methods such as artificial neural network models (e.g., BP neural network) can be employed to evaluate and screen quality content, recommending and presenting quality content to users to control and optimize platform answer quality, attract new users, and build community identity among existing users, thereby promoting sustainable social Q&A community development.

This study addresses automated evaluation of user-generated answer quality in social Q&A communities, introducing social-emotional features and user characteristic dimensions to construct a user-generated answer quality evaluation index system through factor analysis and structural equation modeling, addressing issues of incompleteness, ambiguity, and lack of personalization. A GA-BP neural network-based automated evaluation method was designed. Zhihu website data was selected for applied research on the index system and automated evaluation method. Application results demonstrate the rationality and effectiveness of the constructed index system and evaluation method. However, limitations remain. First, the application sample selection has limitations, using only partial Zhihu data to verify method effectiveness, with single-topic content not extended to various fields and Q&A community types. Sampling limitations may cause research conclusion bias. Future research will expand application object selection and method application scope. Second, evaluation indicators were only selected and quantified from external feature levels such as text and respondent characteristics, without deep semantic content analysis. Future research should combine semantic web and machine learning technologies to strengthen semantic-level evaluation of user-generated answer quality. Additionally, context is an important factor influencing user answer quality evaluation and selection, warranting further exploration of different dimensional factors' impact on answer quality in subsequent research.

References

[1] KIM S, OH J S, OH S. Best-answer selection criteria in a social Q&A site from a user-oriented relevance perspective[J]. Proceedings of the Association for Information Science and Technology, 2007, 44(1): 1-15.

[2] ISHIKAWA D, KANDO N, SAKAI T. What makes a good answer in community Q&A? An analysis of assessors' criteria[EB/OL]. [2018-12-26]. https://www.researchgate.net/publication/228449185.

[3] OH S, WORRALL A, YI Y J. Quality evaluation of health answers in community question answering? An analysis of assessors' criteria[J]. Proceedings of the Association for Information Science and Technology, 2011, 48(1): 1-3.

[4] FICHMAN P. A comparative assessment of answer quality on four question answering sites[J]. Journal of information science, 2011, 37(5): 476-486.

[5] CHUA A Y K, BANERJEE S. So fast so good: an analysis of answer quality and answer speed in community question-answering sites[J]. Journal of the Association for Information Science and Technology, 2013, 64(10): 2058-2068.

[6] Sun Xiaoning, Zhao Yuxiang, Zhu Qinghua. Construction of social search answer quality evaluation indicators based on SQA system[J]. Journal of Library Science in China, 2015, 41(4): 65-82.

[7] Li Xiangyu, Chen Kun, Luo Lin. Application research of FWG1 method in social Q&A platform answer quality evaluation system construction[J]. Library and Information Service, 2016, 60(1): 74-80.

[8] Zhang Yuxuan. Research on answer information quality perception in social Q&A platforms based on external clues[D]. Wuhan: Central China Normal University, 2016.

[9] Jiang Wen, Xu Xin, Wu Gaofeng. Automated evaluation of online Q&A community information quality with additional emotional features[J]. Library and Information Service, 2015, 59(4): 100-105.

[10] Yuan Hong, Zhang Ying. Quality evaluation of questions and answers in Q&A communities—A comparative study based on Baidu Knows and Zhihu[J]. Digital Library Forum, 2014(9): 43-49.

[11] Kong Weize, Liu Yiqun, Zhang Min, et al. Research on answer quality evaluation methods in Q&A communities[J]. Journal of Chinese Information Processing, 2011, 25(1): 3-8.

[12] Luo Yi, Cao Qian. Research on answer quality in social Q&A platforms based on RIPA method[J]. Library and Information Service, 2015, 59(3): 126-133, 25.

[13] Jiang Wen, Xu Xin. Review of online Q&A community information quality evaluation research[J]. New Technology of Library and Information Service, 2014(6): 41-50.

[14] JEON J, CROFT W B, LEE J H, et al. A framework to predict the quality of answers with non-textual features[C]//Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2006: 228-235.

[15] SHAH C, POMERANTZ J. Evaluating and predicting answer quality in community QA[C]//Proceedings of the 33rd international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2010: 411-418.

[16] Li Chen, Chao Wenhan, Chen Xiaoming, et al. Quality evaluation of questions and answers in Chinese community Q&A[J]. Computer Science, 2011, 38(6): 230-236.

[17] Wang Wei, Ji Yuqiang, Wang Hongwei, et al. Research on answer quality evaluation in Chinese Q&A communities: Taking Zhihu as an example[J].

Library and Information Service, 2017, 61(22): 36-44.

[18] Cui Minjun, Duan Liguo, Li Aiping. Research on multi-feature hierarchical answer quality evaluation method[J]. Computer Science, 2016, 43(1): 94-97, 102.

[19] Hu Haifeng. Research on feature representation and fusion in user-generated answer quality evaluation[D]. Harbin: Harbin Institute of Technology, 2013.

[20] WANG R Y, STRONG D M. Beyond accuracy: what data quality means to data consumers[J]. Journal of management information systems, 1996, 12(4): 5-33.

[21] JOHN B M, CHUA A Y K, GOH D H L. What makes a high-quality user-generated answer?[J]. IEEE Internet computing, 2011, 15(1): 66-71.

[22] LIU B, FENG J, LIU M, et al. Predicting the quality of user-generated answers using co-training in community-based question answering portals[J]. Pattern recognition letters, 2015, 3(58): 29.

[23] Xu Anying, Ji Zongcheng, Wang Bin. Research on answer quality prediction based on user response order in community Q&A[J]. Journal of Chinese Information Processing, 2017, 31(2): 132-138.

[24] HONG L, LEE J T, SONG Y I, et al. A model for evaluating the quality of user-created documents[C]//Asia information retrieval symposium. Berlin: Springer, 2008: 496-501.

[25] LIU Y, BIAN J, AGICHTEIN E. Predicting information seeker satisfaction in community question answering[C]//Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval. New York: ACM, 2008: 483-490.

[26] TIAN Q, ZHANG P, LI B. Towards predicting the best answers in community-based question-answering services[EB/OL]. [2018-12-26]. http://www.public.asu.edu/~bli24/Papers/ICWSM2013.pdf.

[27] Liu Gaojun, Ma Yanzhong, Duan Jianyong. Quality evaluation of "question-answer pairs" in community Q&A systems[J]. Journal of North China University of Technology, 2012, 24(3): 31-36.

[28] Lai She'an, Cai Zhongmin. Q&A quality evaluation method based on similarity in Q&A communities[J]. Computer Applications and Software, 2013, 30(2): 266-269.

[29] CAI Y, CHAKRAVARTHY S. Answer quality prediction in Q/A social networks by leveraging temporal features[J]. International journal of next-generation computing, 2013, 4(1): 1-27.

[30] LI B, JIN T, LYU M R, et al. Analyzing and predicting question quality in community question answering services[C]//Proceedings of the 21st international conference on World Wide Web. New York: ACM, 2012: 775-782.

[31] Yuan Jian, Liu Yu. Hybrid-based community Q&A answer quality evaluation model[J]. Application Research of Computers, 2017, 34(6): 1708-1712.

[32] ANAND D, VAHABZADEH F A. Predicting post importance in question answer forums based on topic-wise user expertise[C]//International conference on distributed computing and Internet technology. Berlin: Springer, 2015: 365-376.

[33] ARAI K, HANDAYANI A N. Predicting quality of answer in collaborative Q/A community[J]. Society and culture, 2013, 2(3): 21-25.

[34] Wu Minglong. Structural Equation Modeling—AMOS Operation and Application[M]. Chongqing: Chongqing Publishing House, 2009: 52-53.

[35] Zhu Shuangdong. Neural Network Application Foundation[M]. Shenyang: Northeastern University Press, 2000.

[36] TANG H, WU E X, MA Q Y, et al. MRI brain image segmentation by multi-resolution edge detection and region selection[J]. Computerized medical imaging and graphics, 2000, 24(6): 349-357.

[37] JEMEL S, HISSEL D, PERA M C, et al. On-board fuel cell power supply modeling on the basis of neural network methodology[J]. Journal of power sources, 2003, 124(2): 479-486.

Author Contributions

Guo Shunli: Initial draft writing, data collection and processing;
Zhang Xiangxian: Framework development, revision and final approval;
Tao Xing: Data collection and processing;
Zhang Liman: Format revision, Chinese and English abstract writing.

Note: Figure translations are in progress. See original paper for figures.