Shi, J., Shen, H. & Ma, Q. (2018). What kind of answer will be better: Exploring the features of high-quality answer contents in social Q&A community. In *Proceedings of The 19th International Conference on Electronic Business* (pp. 558-562). ICEB, Newcastle upon Tyne, UK, December 8-12.

What Kind of Answer will be Better: Exploring the Features of High-quality Answer Contents in Social Q&A Community

(*Work in Progress*)

Junpeng Shi, Nanjing University of Posts and Telecommunications, Nanjing, China, njuptsjp@163.com
Hongzhou Shen*, Nanjing University of Posts and Telecommunications, Nanjing, China, shzsys@126.com
Qiaohui Ma, Nanjing University of Posts and Telecommunications, Nanjing, China, 1198354664@qq.com

ABSTRACT

The rigorousness, professionalism and seriousness of answer contents in social Q&A communities have declined sharply. It is of great significance to tell users how to provide excellent answer contents. The purpose of this research is to explore what features of answer content are more likely to make a high-quality answer. For this purpose, the research, taking "Zhihu" as an example, collects the data of answer contents with a crawler, combines initial analysis of the answer contents with related research literature, and selects 9 features that may have impacts on the quality of answer contents. Then the supervised machine learning method will be used to explore the features that can really affect the quality of answer contents, so as to provide reference for guiding users to provide high-quality answer contents.

Keywords: social Q&A community, high-quality answer contents, features, supervised machine learning

*Corresponding author

INTRODUCTION

With the continuous development and wide application of information and communication technology, the traditional mode of knowledge dissemination has undergone tremendous changes. Users are no longer confined to the keyword-based search engines to acquire knowledge, but also can ask questions and seek for answers in the social Q&A communities. A social Q&A community is the product of the combination of traditional knowledge dissemination network and Internet technology. Users can use this kind of community to ask questions that need to be solved and get answers from other users to solve the problems. Users can also answer questions on the community and view the best questions, answers or respondents in the community. At present, the representative social Q&A communities are "Yahoo! Answers", "Quora", "Zhihu" and so on.

Social Q&A community creates value through the connection of questions and answers that provided by users. On the one hand, there are no entry conditions for users who want to answer questions. Users from all walks of life with different identities and backgrounds can answer questions. On the other hand, there are huge differences in knowledge reserve, motivation and information literacy among different users. These factors make the quality of answers imbalanced, which is embodied in the simple and random form of answers. As a result, the rigorousness, professionalism and seriousness of the whole social Q&A community have declined dramatically. In this context, how to enhance the rigor of social Q&A community and guide users to provide higher quality answers is worthy of in-depth discussion.

In order to improve the professionalism of social Q&A community and guide ordinary users to provide high-quality answers, we focus on the answer content itself, without considering the differences in respondents' abilities and other users' evaluations, try to address this research question: What features of answer content are more likely to make a high-quality answer? Answering this question can help guide users of social Q&A community to give higher quality contents when answering questions. To this end, this research, taking China's most well-known social Q&A community "Zhihu" as the research platform, collects user's answer contents with a crawler program, and explores various features that can affect the quality of answers by using the method of supervised machine learning. Findings of this research could provide suggestions for users to make higher quality answers from the perspective of content form, and provide strategies for the functional optimization of social Q&A communities from the perspective of user requirements.

The rest of this paper is organized as follows. Firstly, we review the related literature from 2 aspects, including evaluation of the quality of answer content and features of answer content. We then introduce the research method, and give the process of feature selection. Research data collection and preprocess will be discussed later. Finally, we conclude the current research work and present the plan of the future work.

RELATED LITERATURE

Evaluation of the Quality of Answer Content

Because of the uneven quality of user-generated answer content in the social Q&A community, it has become a hot topic for researchers to evaluate and rank the quality of answer content from the perspective of readers. The key to the evaluation of the

quality of answer content is to determine the evaluation criteria. Through content analysis and iterative induction of the types of comments, seven criteria for choosing the best answers were deduced: content value, cognitive value, social emotional value and information source value, external value, utility and general statement (Kim et al., 2008). Shah et al. (2010) took Yahoo! Answers as a research object. By comparing the criteria of selecting the best answer between expert group and questioner, 13 criteria affecting users' choice of the best answer were determined. Based on these criteria, a classifier was trained to classify the quality of answers. The number of points praised by answers can be the criterion of high-quality answer. On this basis, some researchers studied the specific features that may affect the quality of the answers. Li et al. (2015) found that respondents' credibility and timeliness of answers, as well as the length of answers, were positively correlated with the quality of peer-judged answers. The quality of peer-judged answers was affected by the inclusion of social elements in the answers. Jeon et al. (2006) evaluated the quality of answers from 12 non-textual features, such as number of clicks, length of answers, and active level of respondents. Empirical data show that the model can improve the accuracy of evaluating the quality of answers. Li and Zhang (2018) directly used the number of praise points obtained by the answers as the criteria for high-quality answers, and explored the influencing factors of perceived usefulness of knowledge sharing users. The results showed that the features of answers (timeliness, pictures or citations), the quality of answers (answer-centeredness, emotional support), and the features of respondents (social network-centeredness, credibility) have a positive effect on the voting of answering usefulness, and the linguistic diversity of answers has a negative effect on the voting of answering usefulness.

To figure out what feature may increase the probability of the answer being cited, Calefato et al. (2015) conducted a research on the user answer content of Stack Overflow and found that technical-style has a positive impact. User's reputation may be a major influencing factor, Lin and Shen (2015) propose a SmartQ model, which uses a reputation management system based on categories and topics to evaluate users' willingness and ability to answer various questions. However, Liu et al. (2015) believed that previous works analyzed the answer quality (AQ) based on answer-related features, but neglect the question-related features on AQ. Previous work analyzed how asker- and question-related features affect the question quality (QQ) regarding the amount of attention from users, the number of answers and the question solving latency, but neglect the correlation between QQ and AQ (measured by the rating of the best answer), which is critical to quality of service (QoS). Aiming at the unbalanced quality of answer content, Jin and Li (2016) constructed a real-time monitoring framework of user generated content quality based on SPC from the perspective of UGC creation process.

Features of Answer Content

In the study of the features of the answer content, following the early exploratory study from the answer content itself (S. Oh et al. 2008), researchers pay more attention to a specific subject area and study the features of the answer content in specific areas. Jeng et al. (2017) explored the features of the answers on the "ResearchGate" by combining qualitative content analysis and quantitative statistical analysis methods, and whether these features exist in different disciplines. The conclusion indicates that the type of problem and the difference in the subject will have an impact on the features of the answer. Oh et al. (2013) studied the linguistic features of Yahoo! Answers's eating disorders from three dimensions: answer content, answer style and sentiment analysis. By comparing the responses of the two categories of social emotional problems and information problems, it was found that the needs of different questions constituted different features of the answer content. Negative emotional words are used extensively in social emotional problems, while the answer to information problems uses more complex, precise and objective feature words.

Some researchers have carried out in-depth research on the demonstration mode of the answer content. Ruan and Xia (2018) applied topic models to the recognition semantics of high-quality user-generated content to mine the features of high-quality user-generated content. Based on the analysis of the features of social platform users, the user-generated content theme model based on LDA was constructed, which demonstrated the relationship between high-quality user reviews and topic distribution. The difference in the subject will have an impact on the features of the answer. From the perspective of discussion patterns, Savolainen (2012) used five main discussion mode which Toulmin (2003) proposed to analyze the answers to the 100 global warming questions on Yahoo! Answers which found that the most commonly used discussion was statement, followed by rebuttal and non-rebuttal, and a mixture of multiple discussion modes. In most cases, respondents will support their opinions with personal opinions and experience. At the same time, the research results show that the use of rebuttal and mixed discussion mode will make the answer have better quality and credibility than the statement and non-rebuttal argumentation model. Furthermore, Savolainen (2013) analyzed the answers to the rebuttals and mixed discussion models obtained in the above study. By exploring what kind of information resources the respondents use to support refutation, it is found that questioning the rationality of other answers and their background assumptions is the most frequently used method, and the respondents will also question the motivation of other respondents. At the same time, the study found that respondents mainly use information resources on the Internet as resources to support refutation. Then Savolainen (2014) studied the features of answers from the perspective of rhetorical strategies. Based on the rhetoric theory put forward by Borchers (2018), it is believed that the rhetoric strategies of explaining reasons, results and authority are most commonly used. At the same time, it is found that respondents will use scientific information resources to support their rhetorical strategies, such as research institutions' websites or persuasive video resources.

Summary of Literature Review

Through the review of relevant literature, it can be found that researchers have made in-depth research on the content and quality of answers in social Q&A communities, especially the evaluation of the quality of answers, which provides us with

ideas for further research on the features that can affect the quality of answers. Currently, most of the research focuses on how to evaluate the quality of an answer and how to rank the answers, aiming at helping find out better answers. However, this research from the perspective of users, studies what features of answer will lead to a higher quality answer, so as to help users input better answers and help social Q&A communities improve their question answering functions.

RESEARCH METHOD

This research will be based on a large amount of answer content data on "Zhihu", using machine learning method to explore answers with which features are more likely to be high-quality answers. With the development of computer technology and the reduction of the cost of collecting and storing data, a large amount of data has been accumulated in various fields of people's lives, thus providing a wider range of uses for machine learning. Machine learning methods are usually divided into supervised learning, unsupervised learning, and reinforcement learning. This classification is based on the fact that if the training data entered by the learning system contains existing experience. The training data for supervised learning includes real results with known tasks. These results can be used to evaluate the pros and cons of learning outcomes. Common supervised learning includes regression analysis and statistical classification. There is no real result known in the unsupervised learning, which means it does not contain the existing empirical knowledge. In the learning process, it is necessary to establish an evaluation method for the learning result. The common unsupervised learning is clustering. The empirical knowledge of reinforcement learning comes from the feedback of the environment in the learning process, sometimes there will be delays.

Specifically, this research will use the regression analysis algorithm in supervised learning classification. The experimental data source is the answers under the topic of "Computer Science" on the "Zhihu" social Q&A community. According to the initial analysis of the collected answers and the existing related research on the quality of answer content, we filter out all the possible features that may affect the quality of the answer. When collecting the answers of each question, the original order of answers is retained. The order is determined by the ranking mechanism of "Zhihu" that follows the principle of weighted voting. The top-ranked answer endorsed by other users reflects the high quality of the answer content. Therefore, we take the top 10% of the answers as high-quality answers and mark them as positive samples; the last 90% of the answers are normal answers and are marked as negative samples. Under the premise of determining the specific features and sample values, the appropriate regression analysis algorithm will be used to train and analyze the data set, and feature evaluation is used to explore which features can improve the quality of the answer content.

FEATURE SELECTION

By observing and analyzing the high-quality answers on "Zhihu", and combining with the existing related research on the answer content of social Q&A community, we filter out all the possible features that may have an impact on the quality of the answers. The selected features are shown in Table 1.

Text Features of Answer Content

When users read the answers, the length of text and the number of paragraphs might be the factors that directly affect users' acquisition of key knowledge points. To some extent, the longer text and the more paragraphs will show a more detailed answer, which is helpful to users' in-depth understanding. However, it's relative low efficiency to acquire key knowledge points. Therefore, in this research, the length of the text and the number of paragraphs are taken as two text features of the answer content. It happens frequently that quoting massive data to demonstrate the viewpoint can increase the credibility of the answer. Meanwhile, marking key sentences can help users quickly locate the key points of the answer. Also, the number of data (times of data-quoting) contained in the text and marked sentences are also taken into account in this research. The different social Q&A community platforms will form some unique style of user's answer content, which is reflected in some unique words in the platform, such as "Thanks for inviting", "Respondent lives in the United States", "Just got off the plane" and so on in "Zhihu". Therefore, this study regards the frequency of these unique words as a textual feature.

Rhetorical Features of Answer Content

Compared with the pure text content, users prefer the photo-illustrated content, which can not only make the answer more vivid and interesting, but also make the answer more comprehensive and clearer. Borchers' (2018) research indicates that when explaining the reasons and results and demonstrating authority, respondents prefer rhetorical strategies and use scientific information resources to support their strategies, such as citing external pictures and links. In addition, when answering some professional questions, there are often some proper nouns with a hyperlink, which can be linked to help users quickly understand domain knowledge. These words are called keywords. In addition to the answer content itself, the external information resources cited by the answer content are described as the rhetorical features of the answer content. Consequently, we propose three rhetorical features: the number of pictures, the number of external links, and the number of keywords.

Emotional Features of Answer Content

Xu et al. (2016) aimed at user-generated content of information in text form under Web 2.0 environment, used fuzzy statistical method to determine index weight and membership degree of fuzzy comprehensive evaluation, selected appropriate fuzzy synthesis operator through comparative analysis, and established user-generated content fuzzy comprehensive evaluation model FCE based on emotional analysis. Jin (2015) found that emotional support had a positive impact on information adoption. Joyce et al. (2006) found that information with positive emotions helps to enhance user identification. Therefore, this

study regards the emotional value of the answer content as a feature. The emotional value will be calculated by text sentiment analysis method.

Table 1: Selected Features

Feature ID	Feature classes	Feature name	Introduction
F1	Text feature	Text length	The length of answer content text
F2		paragraphs	Number of text paragraphs in answer content
F3		Number of data	Times of data-quoting in answer content
F4		Number of marked	Number of marked sentences in answer content
F5		Number of unique words	Number of unique words in answer content
F6	Rhetorical Feature	Number of pictures	Number of pictures in answer content
F7		Number of links	Number of links in answer content
F8		Number of key words	Number of key words that link to other pages
F9	Emotional feature	Emotional value	Emotional value calculated from the answer content

Source: This study.

DATA COLLECTION AND PREPROCESS

In this research, the social Q&A community "Zhihu" was taken as the research platform, and a total of 87,637 answers were collected under the topic of "Computer Science". The data set contains all the answers from the questions created until 24 o'clock on September 7, 2019. Each of the collected answers is stored in JSON format and contains question, answer content, and respondent information. According to the total of 9 features in the three dimensions that have been determined, feature values of each answer will be processed and calculated. Currently, the calculation of the eight values under the text features and rhetorical features of the answer content has been completed.

CONCLUSION AND FUTURE WORK

This research takes the answer contents in social Q&A community as research object, and hopes to figure out the features of high-quality answer contents against the common ones. By the relevant literature review, we have established three dimensions: text features, rhetorical features and emotional features of answer content, and then identified 9 features that might affect the quality of answer contents. The study took the social Q&A community "Zhihu" as the research platform, and collected a total of 87,637 answers. At present, the calculation of a total of 8 values under the text and rhetorical features has been completed. In the future research, the emotional value of answer content will be calculated through sentiment analysis. We will use the machine learning method to analyze the data set, and feature evaluation will be used to explore which features can improve the quality of the answer contents. Finally, suggestions to guide ordinary users to provide high-quality answers will be proposed and strategies for the functional optimization will be provided.

ACKNOWLEDGMENT

This research is sponsored by National Natural Science Foundation of China (71974102, 71403134).

REFERENCE

Borchers, T., & Hundley, H. (2018). *Rhetorical theory: An introduction.* Waveland Press.

Calefato, F., Lanubile, F., Marasciulo, M.C., & Novielli, N. (2015). Mining successful answers in stack overflow. *In Proceedings of the 12th Working Conference on Mining Software Repositories* (pp. 430-433). Florence, Italy, May 16-24.

Jeng, W., DesAutels, S., He, D., & Li, L. (2017). Information exchange on an academic social networking site: A multidiscipline comparison on ResearchGate Q&A. *Journal of the Association for Information Science and Technology,* 68(3), 638-652. doi: 10.1002/asi.23692

Jeon, J., Croft, W.B., Lee, J.H., & Park, S. (2006). A framework to predict the quality of answers with non-textual features. *In Proceedings of the 29th International ACM SIGIR Conference on Research & Development in Information Retrieval.* Seattle, Washington, USA, August 06-11.

Jin, J.H. (2015). Research on the influencing factors of user's knowledge activities in online social Q&A communities. (Doctoral dissertation, Harbin Institute of Technology, Harbin, China, in Chinese).

Jin, Y., & Li, D. (2016). Quality monitoring of user-generated content based on SPC. *Information Science*, 34(5), 86-90. doi: 10.13833/j.cnki.is.2016.05.017

Joyce, E., & Kraut, R.E. (2006). Predicting continued participation in newsgroups. *Journal of Computer-Mediated Communication*, 11(3), 723-747. doi: 10.1111/j.1083-6101.2006.00033.x

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. doi: 10.1002/meet.1450440256

Li, J., & Zhang, T. (2018). Research on influencing factors of user perceived usefulness of knowledge sharing in social Q&A ———A case study of Zhihu. *Journal of Modern Information* (4), 20-28. doi: 10.3969 / j.issn.1008-0821.2018.04.003

Li, L., He, D., Jeng, W., Goodwin, S., & Zhang, C. (2015). Answer quality characteristics and prediction on an academic Q&A site: A case study on researchgate. *International Conference on World Wide Web Companion.* (pp. 1453-1458). ACM. Florence, Italy, May 18-22.

Lin, Y., & Shen, H. (2015). SmartQ: A question and answer system for supplying high-quality and trustworthy answers. *IEEE Transactions on Big Data,* 4(4), 600-613. doi: 10.1109/TBDATA.2017.2735442

Liu, J., Shen, H., & Yu, L. (2015). Question quality analysis and prediction in community question answering services with coupled mutual reinforcement. *IEEE Transactions on Services Computing,* 10(2), 286-301. doi: 10.1109/TSC.2015.2446991

Oh, J.S., He, D., Jeng, W., Mattern, E., & Bowler, L. (2013). Linguistic characteristics of eating disorder questions on Yahoo! Answers-content, style, and emotion. *In Proceedings of the 76th ASIS&T Annual Meeting: Beyond the Cloud: Rethinking Information Boundaries* (p. 87). Montreal, Quebec, Canada, November 01-05.

Oh, S., Oh, J.S., & Shah, C. (2008). The use of information sources by internet users in answering questions. *Proceedings of the American Society for Information Science and Technology, 45*(1), 1-13. doi: 10.1002/meet.2008.1450450279

Ruan, G., & Xia, L. (2018). Study of topic distribution features of user generated comments of high quality. *Library Journal,* 37(4), 95-101. doi: 10.13663/j.cnki.lj.2018.04.013

Savolainen, R. (2012). The structure of argument patterns on a social Q&A site. *Journal of the American Society for Information Science and Technology,* 63(12), 2536-2548. doi: 10.1002/asi.22722

Savolainen, R. (2013). Strategies for justifying counter-arguments in Q&A discussion. *Journal of Information Science,* 39(4), 544-556. doi: 10.1177/0165551513478892

Savolainen, R. (2014). The use of rhetorical strategies in Q&A discussion. *Journal of Documentation,* 70(1), 93-118. doi: 10.1108/JD-11-2012-0152

Shah, C., & Pomerantz, J. (2010). 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). Geneva, Switzerland, July 19-23.

Toulmin, S.E. (2003). *The Uses of Argument.* Cambridge, UK: Cambridge University Press.

Xu, Y., Zhang, H., & Chen, L. (2016). A fuzzy comprehensive evaluation method of UGC based on emotional analysis-Taking UGC as an example for Taobao comments on commodity texts. *Information Studies: Theory & Application,* 39(6), 64-69. doi: 10.16353/j.cnki.1000-7490.2016.06.013