Data-driven Approach for Quality Evaluation on Knowledge Sharing Platform

Abstract

In recent years, voice knowledge sharing and question answering (Q&A) platforms have attracted much attention, which greatly facilitate the knowledge acquisition for people. However, little research has evaluated on the quality evaluation on voice knowledge sharing. This paper presents a data-driven approach to automatically evaluate the quality of a specific Q&A platform (Zhihu Live). Extensive experiments demonstrate the effectiveness of the proposed method. Furthermore, we introduce a dataset of Zhihu Live as an open resource for researchers in related areas. This dataset will facilitate the development of new methods on knowledge sharing services quality evaluation.

Introduction

Knowledge sharing platforms such as Quora and Zhihu emerge as very convenient tools for acquiring knowledge. These question and answer (Q&A) platforms are newly emerged communities about knowledge acquisition, experience sharing and social networks services (SNS).

Unlike many other Q&A platforms, Zhihu platform resembles a social network community. Users can follow other people, post ideas, up-vote or down-vote answers, and write their own answers. Zhihu allows users to keep track of specific fields by following related topics, such as “Education”, “Movie”, “Technology” and “Music”. Once a Zhihu user starts to follow a specific topic or a person, the related updates are automatically pushed to the user's feed timeline.

Although these platforms have exploded in popularity, they face some potential problems. The key problem is that as the number of users grows, a large volume of low-quality questions and answers

emerge and overwhelm users, which make users hard to find relevant and helpful information.

Zhihu Live is a real-time voice-answering product on the Zhihu platform, which enables the speakers to share knowledge, experience, and opinions on a subject. The audience can ask questions and get answers from the speakers as well. It allows communication with the speakers easily and efficiently through the Internet. Zhihu Live provides an extremely useful reward mechanism (like up-votes, following growth and economic returns), to encourage high-quality content providers to generate high-level information on Zhihu platform.

However, due to the lack of efficient filter mechanism and evaluation schemes, many users suffer from lots of low-quality contents, which affects the service negatively. Recently, studies on social Q&A platforms and knowledge sharing are rising and have achieved many promising results. Shah et al. BIBREF0 propose a data-driven approach with logistic regression and carefully designed hand-crafted features to predict the answer quality on Yahoo! Answers. Wang et al. BIBREF1 illustrate that heterogeneity in the user and question graphs are important contributors to the quality of Quora's knowledge base. Paul et al. BIBREF2 explore reputation mechanism in quora through detailed data analysis, their experiments indicate that social voting helps users identify and promote good content but is prone to preferential attachment. Patil et al. BIBREF3 propose a method to detect experts on Quora by their activity, quality of answers, linguistic characteristics and temporal behaviors, and achieves 97% accuracy and 0.987 AUC. Rughinis et al. BIBREF4 indicate that there are different regimes of engagement at the intersection of the technological infrastructure and users' participation in Quora.

All of these works are mainly focused on answer ranking and answer quality evaluation. But there is little research achievement about quality evaluation in voice-answering areas. In this work, we present a data-driven approach for quality evaluation about Zhihu Live, by consuming the dataset we collected to gather knowledge and insightful conclusion. The proposed data-driven approach includes data collection,

storage, preprocessing, data analysis, and predictive analysis via machine learning. The architecture of our data-driven method is shown in Fig. FIGREF3 . The records are crawled from Zhihu Live official website and stored in MongoDB. Data preprocessing methods include cleaning and data normalization to make the dataset satisfy our target problem. Descriptive data analysis and predictive analysis are also conducted for deeper analysis about this dataset.

The main contributions of this paper are as follows: (1) We release a public benchmark dataset which contains 7242 records and 286,938 text comments about Zhihu Live. Detailed analysis about the dataset is also discussed in this paper. This dataset could help researchers verify their ideas in related fields. (2) By analyzing this dataset, we gain several insightful conclusion about Zhihu Live. (3) We also propose a multi-branched neural network (MTNet) to evaluate Zhihu Lives' scores. The superiority of our proposed model is demonstrated by comparing performance with other mainstream regressors.

The rest of this paper is organized as follows: Section 2 describes detailed procedures of ZhihuLive-DB collection, and descriptive analysis. Section 3 illustrates our proposed MTNet. In section 4, we give a detailed description of experiments, and the last section discusses the conclusion of this paper and future work.

Data Collection

In order to make a detailed analysis about Zhihu Live with data-driven approach, the first step is to collect Zhihu Live data. Since there is no public dataset available for research and no official APIs, we develop a web spider with python requests library to crawl data from Zhihu Live official website. Our crawling strategy is breadth-first traverse (we crawl the records one by one from the given URLs, and then extract more detailed information from sub URLs). We follow the crawler-etiquette defined in Zhihu's robots.txt. So we randomly set 2 to 5 seconds pause after per crawling to prevent from being banned by Zhihu, and

avoid generating abnormal traffic as well. Our spider crawls 7242 records in total. Majority of the data are embedded in Ajax calls. In addition, we also crawl 286,938 comments of these Zhihu Lives. All of the datasets are stored in MongoDB, a widely-used NoSQL database.

Statistical Analysis

The rating scores are within a range of INLINEFORM0 . We calculate min, Q1, median, Q3, max, mean, and mode about review count (see Table TABREF8). Because the number of received review may greatly influence the reliability of the review score. From Table TABREF8 we can see that many responses on Zhihu Live receive no review at all, which are useless for quality evaluation.

One of the most challenging problems is no unique standard to evaluate a Zhihu Live as a low-quality or high-quality one. A collection of people may highly praise a Zhihu Live while others may not. In order to remove the sample bias, we delete those records whose review count is less than Q1 (11). So we get 5477 records which belong to 18 different fields.

The statistics of review scores after deletion are shown in Table TABREF9 . The mean score of 5477 records is 4.51, and the variance is 0.16. It indicates that the majority of Zhihu Lives are of high quality, and the users' scores are relatively stable.

Badge in Zhihu represents identity authentication of public figures and high-quality answerers. Only those who hold a Ph.D. degree or experts in a specific domain can be granted a badge. Hence, these speakers tend to host high-quality Zhihu Lives theoretically. Table TABREF10 shows that 3286 speakers hold no badge, 1475 speakers hold 1 badge, and 446 speakers hold 2 badges, respectively. The average score of Zhihu Lives given by two badges holders is slightly higher than others. We can conclude that whether the speaker holds badges does have slightly influence on the Zhihu Live quality ratings, which is

consistent with our supposition.

Furthermore, we calculate the average scores of different Zhihu Live types (See Table TABREF11). We find that Others, Art and Sports fields contain more high-quality Zhihu Lives, while Delicacy, Business and Psychology fields contain more low-quality Lives. We can conclude that the topics related to self-improvement tend to receive more positive comments.

There are two types of Zhihu accounts: personal and organization. From Table TABREF12 , we can see that the majority of the Zhihu Live speakers are men with personal accounts. Organizations are less likely to give presentation and share ideas upon Zhihu Live platform.

Comments Text Analysis

Apart from analyzing Zhihu Live dataset, we also adopt TextRank BIBREF5 algorithm to calculate TOP-50 hot words with wordcloud visualization (see Fig. FIGREF14). Bigger font denotes higher weight of the word, we can see clearly that the majority of the comments show contentment about Zhihu Lives, and the audience care more about “content”, “knowledge” and “speaker”.

Performance Metric

We define the quality evaluation problem as a standard regression task since the scores we aim to predict are continuous values. Hence we use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to estimate the performance of diverse learning algorithms. MAE and RMSE are used to evaluate the fit quality of the learning algorithms, if they are close to zero, it means the learning algorithm fits the dataset well. DISPLAYFORM0 DISPLAYFORM1

where INLINEFORM0 denotes the number of samples, INLINEFORM1 denotes the input feature vector of a sample INLINEFORM2 , INLINEFORM3 denotes the learning algorithm, INLINEFORM4 denotes the groundtruth score of a Zhihu Live response INLINEFORM5 .

The results are calculated by randomly selecting 80% in the dataset as training set, and the remaining records as test set.

MTNet

In this section, we first give a brief introduction of the neural network and then present a description of our proposed MTNet to predict the quality of responses in detail.

Deep Neural Network

Deep neural network (DNN) has aroused dramatically attention due to their extraordinary performance in computer vision BIBREF6 , BIBREF7 , speech recognition BIBREF8 and natural language processing (NLP) BIBREF9 tasks. We apply DNN to our Zhihu Live quality evaluation problem aiming to approximate a function INLINEFORM0 which can accurately predict a Zhihu Live's score.

In our quality evaluation task, we take multiple layer perception BIBREF8 as the basic composition block of MTNet. Since we treat the Zhihu Live quality evaluation problem as a regression task, we set the output neuron equal to 1. DNNs are trained by backpropagation algorithm BIBREF8 .

The calculation details of neural network can be illustrated as: DISPLAYFORM0

where INLINEFORM0 represents output of a neuron, INLINEFORM1 represents weights of the

connections, INLINEFORM2 represents bias, INLINEFORM3 represents nonlinear activation function (sigmoid, tanh and ReLU are often used in practice).

MTNet Architecture

The architecture of our proposed MTNet is shown in Fig. FIGREF24 . It includes 4 parts: an input layer for receiving raw data; shared layers for general feature extraction through stacked layers and non-linear transformation; branched layers for specific feature extraction; and the output layer with one neuron. The output of the last shared layer is fed into different branches. These branches are trained jointly. In the last shared layer, the information flow is split into many branches BIBREF7 , which enables feature sharing and reuse. Finally, the output result is calculated in the output layer by averaging outputs from these branches BIBREF10 . The overall neural network with different branches is trained in parallel. The detailed configuration of MTNet is listed in Tabel TABREF21 .

The advantages of MTNet are as follows:

With multi-branched layers, different data under diverse levels can be fed into different branches, which enables MTNet extract multi-level features for later regression.

Multi-branched architecture in our MTNet can also act as an ensemble method BIBREF10 , which promotes the performance as well.

We use mean square error (MSE) with INLINEFORM0 regularization as the cost function.
DISPLAYFORM0

where INLINEFORM0 denotes the raw input of INLINEFORM1 -th data sample, INLINEFORM2 denotes

the capacity of dataset, INLINEFORM3 denotes groundtruth score of INLINEFORM4 -th Zhihu Live. INLINEFORM5 denotes INLINEFORM6 regularization to prevent from overfitting.

Experiments

We implement our method based on Scikit-Learn BIBREF11 and PyTorch , and the experiments are conducted on a server with NVIDIA Tesla K80 GPU.

Data Preprocessing

Several features' types in ZhihuLive-DB are not numerical, while machine learning predictor can only support numerical values as input. We clean the original dataset through the following preprocessing methods.

For categorical features, we replace them with one-hot-encoding BIBREF11 .

The missing data is filled with Median of each attribute.

We normalize the numerical values with minimum subtraction and range division to ensure values [0, 1] intervals.

The review scores are used as labels in our experiments, our task is to precisely estimate the scores with MTNet. Since the data-driven methods are based on crowd wisdom on Zhihu Live platform, they don't need any additional labeling work, and ensure the reliability of the scores of judgment as well.

Feature Selection

Since feature selection plays an import part in a data mining task, conventional feature extraction methods need domain knowledge BIBREF12 . Feature selection influences model's performance dramatically BIBREF13 .

For conventional regression algorithms, we conduct feature selection by adopting the best Top K features through univariate statistical tests. The hyper-parameter such as regularization item INLINEFORM0 is determined through cross validation. For each regression algorithm mentioned above, the hyper-parameters are carefully tuned, and the hyper-parameters with the best performance are denoted as the final comparison results. The details of f_regression BIBREF14 , BIBREF11 feature selection are as follows:

We calculate the correlation between each regressor and label as: INLINEFORM0 .

We convert the correlation into an F score and then to a p-value.

Finally, we get 15-dimension feature vector as the input for conventional (non-deep learning based) regressors.

Deep neural network can learn more abstract features via stacked layers. Deep learning has empowered many AI tasks (like computer vision BIBREF6 and natural language processing BIBREF9) in an end-to-end fashion. We apply deep learning to our Zhihu Live quality evaluation problem. Furthermore, we also compare our MTNet algorithm with baseline models with carefully designed features.

Experimental Results

We train our MTNet with Adam optimizer for 20 epochs. We set batch size as 8, and weight decay as

1e-5, we adopt 3 branched layers in MTNet. Detailed configuration is shown in Table TABREF21 . We use ReLU in shared layers, and relu6 in branched layers to prevent information loss. Our proposed MTNet achieves 0.2250 on MAE and 0.3216 on RMSE, respectively.

We compare MTNet with other mainstream regression algorithms BIBREF14 (linear regression, KNN, SVR, Random Forest and MLP). The architecture of MLP is 15-16-8-8-1, where each number represents the number of neurons in each layer. We try three kinds of kernels (RBF kernel, linear kernel, and poly kernel) with SVR in our experiments for fair comparison.

The results are listed in Table TABREF37 . Our method achieves the best performance in contrast to the compared baseline regressors.

Conclusion

In this paper, we adopt a data-driven approach which includes data collection, data cleaning, data normalization, descriptive analysis and predictive analysis, to evaluate the quality on Zhihu Live platform. To the best of our knowledge, we are the first to research quality evaluation of voice-answering products. We publicize a dataset named ZhihuLive-DB, which contains 7242 records and 286,938 comments text for researchers to evaluate Zhihu Lives' quality. We also make a detailed analysis to reveal inner insights about Zhihu Live. In addition, we propose MTNet to accurately predict Zhihu Lives' quality. Our proposed method achieves best performance compared with the baselines.

As knowledge sharing and Q&A platforms continue to gain a greater popularity, the released dataset ZhihuLive-DB could greatly help researchers in related fields. However, current data and attributes are relatively unitary in ZhihuLive-DB. The malicious comment and assessment on SNS platforms are also very important issues to be taken into consideration. In our future work, we will gather richer dataset, and

integrate malicious comments detector into our data-driven approach.

Acknowledgements

Supported by Foundation Research Funds for the Central Universities (Program No.2662017JC049) and State Scholarship Fund (NO.261606765054).