Expert finding in community question answering websites

Xiao Zhang
zx1209084495@163.com

Abstract

Community question answering sites are becoming more important in our work and lives. People are always having questions and seeking answers. Zhihu is a site that provides a platform for people to share knowledge. It quickly attracted many users. People need to get the knowledge what they need from the site. Hence, high-quality answers are what it needs. The paper aims to find experts which are also the users to answer the questions. The right users can provide reliable answers to the questions they are good at. To solve the problem, we propose a framework, which considers the feature of the questions, the users, and the answers fully. In particular, we split a suitable train set by the invitation time, which has a distribution much closer to the test set. Meanwhile, the method also utilizes historical information fully. The results show that our model performs better than the benchmark models.

Expert finding, Question encoder, User encoder,LightGBM

ACM Reference Format:
Xiao Zhang. 2019. Expert finding in community question answering websites. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

Community question answering sites are becoming increasingly important in people's lives. This is because that we inevitably encounter questions in both life and work, and these community question answering sites provide a way for us to seek answers. At the same time, the high quality of the answers to a certain extent to ensure the vitality of the site. Zhihu is a well-known comprehensive community platform of the Chinese Internet. Users can ask questions, answer questions, and share their opinions on the website. Founded in 2011, Zhihu has become a website with 220 million users and hundreds of thousands of new questions and UGC generated every day. Meanwhile, it is important for Zhihu to have high-quality answers. In the light of this, how to effectively invite other users with the ability and interest to answer the newly raised questions, improve the problem-solving rate and answer production, has become an urgent requirement of Zhihu website. In this paper, we mainly explore the task to predict the probability between the given question and user. It can find the right users (experts) for the new questions. There is a lot of complicated information in Zhihu, as shown in Figure 1. The questions have question title, question description, related topic, and many answers. Meanwhile, the answers contain much information, such as the answer content, the number of thumbs-ups and so on. Thereafter, how to utilize this information to represent the question is a challenge. The users contain some basic information like gender and activity. Meanwhile, users are divided into different categories in different concepts and follow some topics. The answers they gave earlier are also part of the users' information. Hence, how to represent the users by this related information is also a challenge.

To address the above challenge, we propose a model, which can represent the questions and users, as shown in Figure 2. The model consists of three modules: question encoder, user encoder, and prediction module. The first module utilizes the length of question title and description, and historical answer information to represent the question. The second module employs the basic information of users, their categories in many concepts, and their historical answers to represent users. In the last module, we add some interactive information between the questions and users. Finally, we predict the result by the LightGBM method. There are two contribution as follow:

- We present a framework, which can predict whether the users will answer the given question well.
- It not only utilizes the related information belong to the users and the questions but also applies some interactive information.

The remainder of this paper is structured as follows. Section 2 briefly reviews the related work. The proposed model is introduced in Section 3. Section 4 detail the dataset and the performance, followed by our concluding remarks in Section 5.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

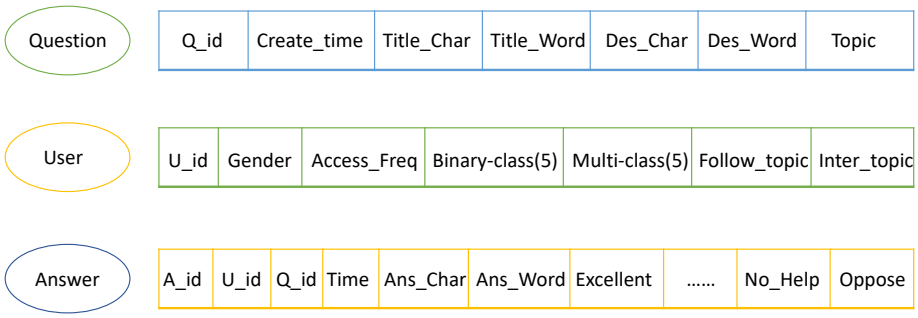

Figure 1. Dataset Information

2 Related work

Expert finding has attracted much attention in the information retrieval (IR) community[2][12][7], and become a well-studied field. The task of expert finding is defined as detecting a set of persons with relevant expertise for the given question. Meanwhile, there are some researches about Expert Finding in many field, such as organizations[4][8], social networks[1][3]. There are some studies of Expert finding in the CQA field. Riahi et al.[6] investigate the suitability of two statistical topic models for solving this issue. Zhao et al. [10] propose a novel ranking metric network learning framework for expert finding by exploiting both users' relative quality rank to given questions and their social relations. Zhou et al.[11] propose a topic-sensitive probabilistic model, which is an extension of PageRank algorithm to find experts in CQA. Liu et al.[5] propose a topic-sensitive probabilistic model to estimate the user authority ranking for each question, which is based on the link analysis technique and topical similarities between users and questions. Zhao et al.[9] think that the past question-answering activities of most users in real CQA systems are rather limited. So, they propose the Graph Regularized Latent Model (GRLM) to infer the expertise of users based on both past question-answering activities and an inferred user-to-user graph. However, they all ignore the time information, we will fully consider it and interactive information between the questions and the users.

3 Model

In this section, we first formulate the task, then we detail the model.

3.1 Problem formulation

Let $\mathcal{S} = \{(\mathcal{Q}_i, \mathcal{U}_i, \mathcal{T}_i, \mathcal{Y}_i)\}_{i=1}^{m}$ denote the set of training instances, $\mathcal{Q}_i$ is the i-th question, $\mathcal{U}_i$ is the i-th user, $\mathcal{T}_i$ is the time of invitation, $\mathcal{Y}_i$ is 0 or 1, and M is the number of training set. Hereafter, $\mathcal{Q}_i$ has $\{(\mathcal{A}_j)\}_{l=1}^{n}$, n is the number of answer about the i-th question. Meanwhile, we observed that the invitation time for the test set is one week in the future. To get better performance, we trained the model with a week's data, and the historical information as the complementary feature for the training data and test data.

3.2 Question encoder

Given a question q_i, we represent it based on its related information. Firstly, each question has the question title and question description. We utilize the length of them as the part of the feature. Meanwhile, the historical answer of question as complementary information. The answers contain much information on different concepts. We also make statistics on past invitations to the question. We encode the question as follows:

$$\begin{cases} \mathbf{q}_i^1 = q\_id, \\ \mathbf{q}_i^2 = sum \oplus std \oplus count \oplus mean(historical\_label), \\ \mathbf{q}_i^3 = sum(historical\_answer), \\ \mathbf{q}_i^4 = sum \oplus max \oplus mean(historical\_answer\_feature), \\ \mathbf{q}_i = [\mathbf{q}_i^1; \mathbf{q}_i^2; \mathbf{q}_i^3; \mathbf{q}_i^4] \end{cases} \tag{1}$$

where the second equation indicates that the statistics information of the historical invitation. $\mathbf{q}_i^3$ is the number of answers to this question. $\mathbf{q}_i^4$ is the statistics information of the historical answer, it is obtained by splicing the result of these operations for historical information together. The similar equations do the same thing. Finally, the concatenation of these features represents the question.

3.3 User encoder

Given a user u_i, it contains 21 attributes, of which 5 attributes are the same for all users. Hence, we deleted them. The gender, activity, binary classification, multi-classification, and score all can be mapped into a number. We also add the length of following topics , the length of interesting topics, and the interest-value information. Meanwhile, the historical

answer information of the user is counted.

$$\begin{cases} \mathbf{u}_i^1 = u\_id, \\ \mathbf{u}_i^2 = map(gender, activity, binary-class, multi-class, score), \\ \mathbf{u}_i^3 = length(follow\_topic, interesting\_topic), \\ \mathbf{u}_i^4 = min \oplus max \oplus mean \oplus std(interest\_value), \\ \mathbf{u}_i^5 = sum(historical\_answer), \\ \mathbf{u}_i^6 = sum \oplus mean \oplus std \oplus count(historical\_label), \\ \mathbf{u}_i^7 = sum \oplus max \oplus mean(historical\_answer\_feature), \\ \mathbf{u}_i = [\mathbf{u}_i^1; \mathbf{u}_i^2; \mathbf{u}_i^3; \mathbf{u}_i^4; \mathbf{u}_i^5; \mathbf{u}_i^6; \mathbf{u}_i^7], \end{cases} \tag{2}$$

in which $\mathbf{u}_i^4$ is the statistics from different angles about the interest-value of interesting topics, $\mathbf{u}_i^5$ is the number of the user's historical answer, the follow two-equation is same as the question encoder but the body is the user.

3.4 prediction module

In this section, give a question q_i and a user u_i, we first obtain the intersections between the topics of the question and the follow (interest) topics of the user. Hereafter, we add the length of the two intersection and the statistics information of the interest-value of the corresponding intersection. The difference between the question creation time and the invitation time is also applied. Finally, we can get the interactive feature $\mathbf{qu}_i$.

Thereafter, we apply the LightGBM classification method to learn the relationship between the questions and the users. Based on the derivation of the GBDT and XGBoost models, the algorithm is improved to form LightGBM, which is a faster model with a lower memory footprint.

$$\mathbf{p}_i = LightGBM([\mathbf{q}_i; \mathbf{u}_i; \mathbf{qu}_i]) \tag{3}$$

where $\mathbf{q}_i$ represents the question vector, $\mathbf{u}_i$ is the user representation encoded by the second module, $\mathbf{qu}_i$ is the interactive information between the question and the user, and $\mathbf{p}_i$ is the probability that the user answers the question.

4 Experiment

In this section, we first introduce the dataset of the competition. Hereafter, we detail the experiment setting and the result of our model.

4.1 Dataset

The train set contains over 9 million data and the test set has 1141718 samples. It totally has 1931654 users and 1829900 questions. The number of user's features is 21, and the number of the multi-categories is 2561, 291, 428, 1556, 2 respectively. The questions contain 7 features. Meanwhile, the answers have 17 features except for the question id, user id, and its id. In particular, the questions, the answers, and the invitation all have creation time. These times have a big effect on the model. The invitation time of the train set starts at 3838 and ends at 3867. The invitation time of the test set starts at 3868 and ends 3874. At the same time, the creation time of the answer is between 3807 and 3874.

Table 1. The results of our model and the baselines. The best result is shown in bold.

Model	AUC
Baseline_1	0.702
Baseline_2	0.807
Our model	**0.823**

4.2 Experiment setting

4.2.1 implement detail

After observation, we can find that the time of the test set is a week in the future. In order to make the training scenario closer to the test set, we choose the final week's data of the train set as the new train set. Meanwhile, the data before the final week can be the historical feature, such as the number of the historical answer for the question. The answers also are selected on time. In particular, it is different for the train set and the test set about the historical data. The origin train set and its related question (answer) information are the complementary features for the test set.

4.2.2 Evaluation Metrics

We use the evaluation metrics AUC (Area Under the ROC cureve) provided by the competition official.

$$AUC = \frac{\sum_{iePostiveClass} \mathbf{rank}_i - \frac{M(M+1)}{2}}{MXN} \tag{4}$$

As shown in the formula above, M is the positive sample number, N is the negative sample number, and $\mathbf{rank}_i$ is the location of the i-th positive sample.

4.3 Baselines

- Baseline_1: It utilizes a lot of statistics about the questions, the users, and the answers. Meanwhile, it also applies the interactive information between the questions and the users. But the model doesn't pay attention to the time information.
- Baseline_2: Our model is based on this model. The model considers the time information, but not all historical information. Meanwhile, it ignores some important features, such as interactive information.

4.4 Performance Comparison

The performance of our model and the baselines are shown in table 1. After observation, we can find that the Baseline_1 performs worst. This is because it ignores the time information and it keeps the data distribution from being close

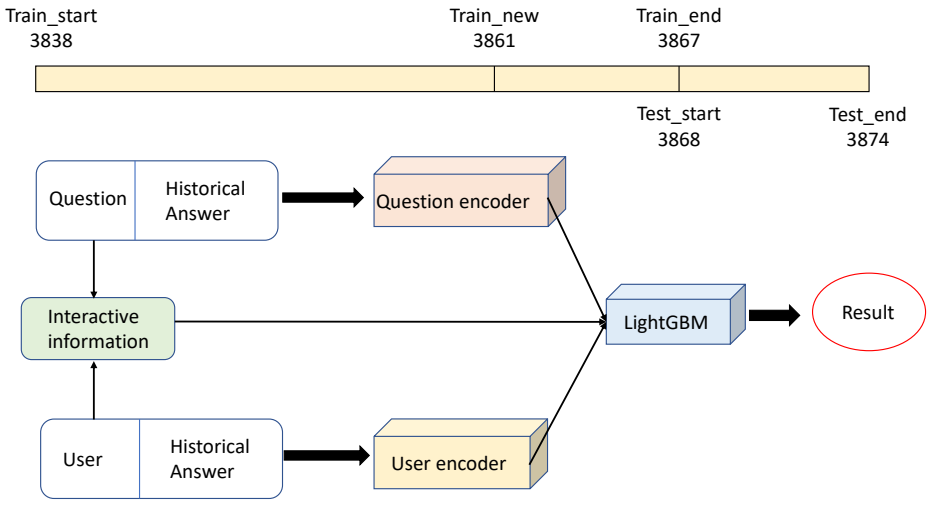

Figure 2. Model

to the test set. The result indicates that it is important to divide the data by time. Baseline_2 performs worse than our model. It doesn't consider the length of the question title and the question description. In particular, it doesn't utilize the interactive information between the questions and the users. Our model obtains the best result, because it considers the time information and more features in a comprehensive way.

5 Conclusion and future work

In this paper, we encode the questions and the user by utilizing the basic information and historical information. The results also show that our model obtains better performance. In the future, we will try to add the word embedding and more knowledge.

References

[1] Alessandro Bozzon, Marco Brambilla, Stefano Ceri, Matteo Silvestri, and Giuliano Vesci. 2013. Choosing the right crowd: expert finding in social networks. In *Proceedings of the 16th International Conference on Extending Database Technology*. ACM, 637–648.

[2] Arash Dargahi Nobari, Sajad Sotudeh Gharebagh, and Mahmood Neshati. 2017. Skill translation models in expert finding. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*. ACM, 1057–1060.

[3] Ahmad Kardan, Amin Omidvar, and Farzad Farahmandnia. 2011. Expert finding on social network with link analysis approach. In *2011 19th Iranian Conference on Electrical Engineering*. IEEE, 1–6.

[4] Maryam Karimzadehgan, Ryen W White, and Matthew Richardson. 2009. Enhancing expert finding using organizational hierarchies. In *European conference on information retrieval*. Springer, 177–188.

[5] Xuebo Liu, Shuang Ye, Xin Li, Yonghao Luo, and Yanghui Rao. 2015. Zhihurank: A topic-sensitive expert finding algorithm in community question answering websites. In *International Conference on Web-Based Learning*. Springer, 165–173.

[6] Fatemeh Riahi, Zainab Zolaktaf, Mahdi Shafiei, and Evangelos Milios. 2012. Finding expert users in community question answering. In *Proceedings of the 21st International Conference on World Wide Web*. ACM, 791–798.

[7] Wei Wei, Gao Cong, Chunyan Miao, Feida Zhu, and Guohui Li. 2016. Learning to find topic experts in Twitter via different relations. *IEEE Transactions on Knowledge and Data Engineering* 28, 7 (2016), 1764–1778.

[8] Dawit Yimam-Seid and Alfred Kobsa. 2003. Expert-finding systems for organizations: Problem and domain analysis and the DEMOIR approach. *Journal of Organizational Computing and Electronic Commerce* 13, 1 (2003), 1–24.

[9] Zhou Zhao, Furu Wei, Ming Zhou, and Wilfred Ng. 2015. Cold-start expert finding in community question answering via graph regularization. In *International conference on database systems for advanced applications*. Springer, 21–38.

[10] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. 2016. Expert Finding for Community-Based Question Answering via Ranking Metric Network Learning.. In *Ijcai*, Vol. 16. 3000–3006.

[11] Guangyou Zhou, Siwei Lai, Kang Liu, and Jun Zhao. 2012. Topic-sensitive probabilistic model for expert finding in question answer communities. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 1662–1666.

[12] Guangyou Zhou, Jun Zhao, Tingting He, and Wensheng Wu. 2014. An empirical study of topic-sensitive probabilistic model for expert finding in question answer communities. *Knowledge-Based Systems* 66 (2014), 136–145.