Identifying Expert Users in Zhihu Dataset Using Profile Data – A Chinese Q&A Community

Andi Liao
University of Chicago, Master of Computational Social Science Program

Introduction

Users have their unique behavior patterns in **online Q&A communities.** Though mental motivation elicits expert user participation in knowledge sharing, it is essential for communities to identify and keep expert users.

Detecting user expertise

- *Wald algorithm & k-means cluster*:
  - combination of motivation metrics and ability metrics.
- *J48, SVM and Random Forest classifier*:
  - activity features, quality of answer features and linguistic features. Random Forest classifier outperforms by achieving 95.94% accuracy.

Predicting user expertise

- *PageRank algorithm*:
  - predict performance of new Zhihu users using biased PageRank algorithms and prestige feature extracted from social media – Weibo.
- *SVM/Decision Tree*:
  - find potential experts using answers/votes/best answers, frequency of login, average time gap between answers and usage of pronoun.

Table 1. Summary of Account Activities Variables.

Features	Mean	Median	Max
answer	28	1	669118
question	2	1	3181
article	0.25	0	1344
favorite	5	2	239
following	156	67	43932
followers	212	3	981917
topic	46	21	22122
columns	11	3	4042
agree	413	0	1218509
thanked	90	0	304153

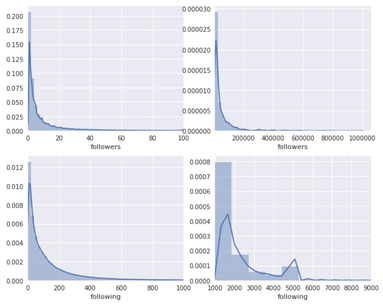


Figure 1. Density Plot of Followers and Following.

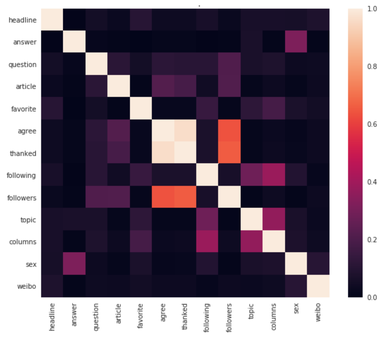


Figure 2. Correlation Matrix between Features.

Data

Data Collection

Zhihu user dataset comes from an open Github repository[1] , where the owner shared **web-scraping data** generated by Java ZhihuSpider in MySQL format.

Cleaning

There are four steps in data cleaning: transforming data from MySQL to csv format; reserving selected variables; handling missing data; converting text data into categorical type. After cleaning, there remains **420949** users.

Preprocessing

To perform supervised learning methods - random forest classification, an **expert label** for each user is needed. As Zhihu allows users to apply for verified symbols, the verified symbol is not equivalent to users expertise.

H-index, defined as: *at least h number of answers of this user received at least h number of agreed*, is considered as a more accurate measurement for user contribution unofficially. A **top 1000 user h-index list[2]**, which released by top users in programming area, served as the label dataset in this study.

Method

The goal is to perform **random forest classification** using demographical information and account activities features to classify users into expert and non-expert with **K-fold cross-validation**.

Performance of models are evaluated according to the number of experts correctly predicted. Accuracy score and cross validation score are not emphasized due to imbalance of expert and nonexpert users in dataset.

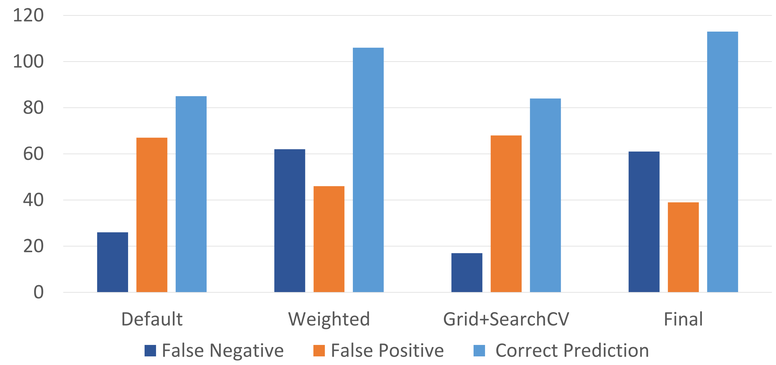
Splitting into training and test with ratio 3:1, four models are calculated:

- *Default*: use default setting of Random Forest Classifiers.
- *Weighted*: adjust weights for class 0 and class 1.
- *GridSearchCV*: change max depth, min samples split and min samples leaf
- *Final*: best parameters of previous models.

Table 2. Summary of Feature Importance in Each Model.

Original		Weighted		GridSearchCV		Final	
Feature	*Weight*	*Feature*	*Weight*	*Feature*	*Weight*	*Feature*	*Weight*
agree	0.2665	following	0.2378	agree	0.3870	agree	0.4404
thanked	0.1769	topic	0.1473	thanked	0.2671	thanked	0.2350
followers	0.1702	favorite	0.1106	followers	0.1497	followers	0.1437
answer	0.1017	followers	0.1019	answer	0.1002	article	0.0819
following	0.0579	columns	0.0903	article	0.0397	answer	0.0661

Chart1: Correct and Wrong Predictions of Each Model.



Results

Demographical Information & Account Activities

- 83.31% of users claim as "male"; 60.60% of users have headline; 21.79% of users display their Weibo account URL.
- Every variable has extremely long tails, minority of users are active.
  - *Content Contribution*: *answer, question, article*.
  - *Social Connection*: *following, followers, favorite, topic, columns*.
  - *Popularity/Recognition*: *agree, thanked*.

Random Forest Classifiers

- *Default*:
  - it performs fairly with low false negative and reasonable features.
- *Weighted*:
  - with weighted ratio 1:10000000, it improves when finding experts, but important features change dramatically.
- *GridSearchCV*:
  - it improves based on the default models judging from top features.
- *Final Model*:
  - using the best parameters from GridSearchCV and weighted ratio 1:5, it keeps a balance of correct prediction and meaningful features.

Conclusions

From the random forest classifier model, the most important features of expert user are **Popularity and Recognition** and **Content Contribution**, and **Social Connection** doesn't contribute much in identifying expert users. This finding is reasonable, as some popular users who always provide "clever" answers will not be viewed as expert users in the model.

As for the application of this research, the model can help communities to thrive by encouraging expert users to contribute more, promoting answers of expert users and even finding potential expert users when they first join.

Contact

Andi Liao
University of Chicago
Email: liaoad17@uchicago.edu
Website: www.github.com/liaoandi
Phone: 7732195749

References & Website

1. Bouguessa, M., Dumoulin, B., & Wang, S. (2008). Identifying authoritative actors in question-answering forums: The case of yahoo! Answers. In Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining (pp. 866–874). ACM.
2. Furtado, A., Oliveira, N., & Andrade, N. (2014). A case study of contributor behavior in q&A site and tags: The importance of prominent profiles in community productivity. Journal of the Brazilian Computer Society , 20 (1), 5.
3. Guan, T., Wang, L., Jin, J., & Song, X. (2018). Knowledge contribution behavior in online q&A communities: An empirical investigation. Computers in Human Behavior , 81 , 137–147. Elsevier.
4. Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. In Proceedings of the sixteenth acm conference on conference on information and knowledge management (pp. 919–922). ACM.
5. Pal, A., Farzan, R., Konstan, J. A., & Kraut, R. E. (2011). Early detection of potential experts in question answering communities. In International conference on user modeling, adaptation, and personalization (pp. 231–242). Springer.
6. Patil, S., & Lee, K. (2016). Detecting experts on quora: By their activity, quality of answers, linguistic characteristics and temporal behaviors. Social network analysis and mining , 6 (1), 5. Springer.
7. Shah, C., Oh, J. S., & Oh, S. (2008). Exploring characteristics and effects of user participation in online social q&A sites. First Monday , 13 (9).
8. Wang, Z., & Zhang, P. (2016). Examining user roles in social q&A: The case of health topics in zhihu. Com. Proceedings of the Association for Information Science and Technology , 53 (1), 1–6. Wiley Online Library.
9. Xiao, Y., Zhao, W. X., Wang, K., & Xiao, Z. (2014). Knowledge sharing via social login: Exploiting microblogging service for warming up social question answering websites. In Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers (pp. 656–666).

[1]https://github.com/MatrixSeven/ZhihuSpider
[2]https://www.zhihu.com/question/31273136/answer/106466841