Identifying Expert Users in *Zhihu* Dataset Using Activities Data: A Chinese Online Question & Answer Community

Andi Liao

Abstract: Expert users are extremely essential to the thrive of online Q&A communities as they are willing to share knowledge and make content contribution. Using profile information and activity summary data, this research build several random forest classifiers to identify expert users in *Zhihu* user dataset. The final model shows that Popularity and Recognition and Content Contribution features are more important in classifying expert users, which can help online Q&A communities to success by keeping expert users.

Keywords: Random Forest Classifier; Expert Users; Q&A Communities; Content Contribution.

I. Introduction

Online Question & Answer communities, which allow users to post questions and obtain answers, are becoming more and more popular these days. Typical examples include *Quora*, *Yahoo! Answer* and *Stack Overflow*. In this research, *Zhihu* is chosen as the target community because it is the largest social based online Q&A community in China with 100 million registered individual users.

Knowledge sharing behaviors of users are crucial online communities, since it largely determines the survival of websites. As the research focuses on identifying expert users inside *Zhihu* users, the following section reviews previous researches about user interaction and user expertise in online Q&A communities.

1.1 User Interaction

Researchers study user interaction behaviors from different perspectives, and here it is discussed from the angle of user participation and user role.

1.1.1 User Participation

There are many factors contributing to participation behavior in online Q&A communities, including personal features, network features and mental motivation. Focusing on mental motivation, researchers (Guan, Wang, Jin, & Song, 2018) collect user activities data from *Zhihu* to build a hierarchical regression model. The result shows that identity-based trust, positive feedback, social exposure, norms of reciprocity and self-presentation information all facilitate continuous knowledge contribution behavior inside *Zhihu* community.

Researchers also believe that user participation is the most important factor contributing to the success of online Q&A communities. Researchers (Shah, Oh, & Oh, 2008) choose *Yahoo! Answers*, a social Q&A site, and *Google Answers*, a paid expert Q&A site, to compare. They figure out that the number of contributors and consumers are well balanced in *Yahoo! Answers* but not in *Google Answers*. By encouraging user participation, *Yahoo! Answers* develops as a responsive community and attracts users to repeatedly raise

questions or provide answers, while *Google Answers* has a high percentage of one-time consumers because of restricting user participation.

1.1.2 User Role

Basically, user roles in online Q&A communities can be divided into:

- contributors and consumers (Shah et al., 2008);
- administrators, content contributors and marginal roles (Wang & Zhang, 2016).

A recent research classifies *Zhihu* users as starters, answerers, technical editors and followers in health-related topics (Wang & Zhang, 2016). Researchers discover that men are more likely to be answerers and technical editors, while women prefer to be starters and followers. Most users are in IT industry, followed by public health and social work. In terms of content contribution, technical editors and starters post more questions, and answerers answer post more answers. As for social connection, followers follow more topics and users, and answerers have more followers. Speaking of popularity and recognition, technical editors have more pageviews, votes and likes. These activity measurements are not isolated, as they are all significantly and positively related to each other.

1.2 User Expertise

Researches about user expertise can be divided into two categories. One aspect is user expertise estimation, utilizing methods like interaction graph analysis and interest modelling; the other aspect is user quality prediction, including questions quality and

answers quality based on historical data (Xiao, Zhao, Wang, & Xiao, 2014). The following section reviews mainstream methods of detecting and predicting user expertise in online Q&A communities.

1.2.1 Detecting User Expertise

The simplest way to measure user expertise to use existing metrics extracted from the community (Shah et al., 2008). Researchers utilize the points and levels system embedded in *Yahoo! Answers* to cluster users, where the level of a user is largely determined by his or her contribution of answers. The result demonstrates that answers from users in higher levels are routinely selected as best answers and reflects user expertise knowledge in specific areas.

There are more complicated methods, including link analysis, clustering and unsupervised learning techniques applied in detecting expert users.

Researchers (Bouguessa, Dumoulin, & Wang, 2008) model Indegree, a measurement of user authority considering the sum of weights of edges that point to the current node, as a combination of two gamma components, and then use this technique to automatically discriminate expert users and non-expert users in *Yahoo! Answers*.

Researchers (Furtado, Oliveira, & Andrade, 2014) also use Wald algorithm and k-means cluster method to classify contributors in *Super User* community into four categories: no marked skill contributors, unskilled answerers, experts and activists. The key measurement is the combination of motivation metrics and ability metrics based on the quantity and quality of user contributions.

To detect expert users on *Quora*, researchers collect profiles of users who follow a certain set of topics, and label them as either expert or non-expert (Patil & Lee, 2016). Researchers find out that expert users and non-experts behave differently: expert users have more followers, make more edits, generally post longer answers, post more questions and prefer lexical words compared to non-expert users. Then researchers build J48, SVM and Random Forest classifier using activity features, quality of answer features and linguistic features. Random Forest classifier outperforms other models by achieving 95.94% accuracy in the general-topic dataset, and even better in specific-topic dataset.

1.2.2 Predicting User Expertise

Apart from identifying expert users, researchers also predict expert users using different approaches, such as link analysis and machine learning methods.

Comparing the HITS algorithm with Degree algorithm, researchers (Jurczyk & Agichtein, 2007) discover that the HITS model is more robust when predicting experts in *Yahoo! Answers*, and it can successfully predict experts in a specific domain and the percentage of best answers generated by expert users in the next 5 months.

Researchers (Xiao et al., 2014) also predict the performance of new-coming *Zhihu* users using PageRank algorithms and related social media features. Researchers choose new *Zhihu* users who logged in using the social media account *Sina Weibo*, and test combinations of unbiased / biased PageRank algorithms and *Weibo*'s prestige/relevance features to predict best answers and top experts. The result implies that using biased PageRank algorithms and prestige feature can greatly improve the prediction performance, especially when the history time window is small.

To find potential experts in *TurboTax Live* Community, researchers (Pal, Farzan, Konstan, & Kraut, 2011) utilize SVM and Decision Tree over motivation and ability features of users. The conclusion is that SVM can reach 0.89 precision when combing two features, consisting of the number of answers/votes/best answers, frequency of login, average time gap between answers and usage of pronoun.

1.3 Research Interest

Based on the literature mentioned above and data availability, this research aims at detecting expert users using random forest classification using demographical information and account activities data, as it performs the best among previous classification models. The goal is to identify which factors are most important in detecting expert users and provide theorical explanation for why these features are unique for expert users, including Social Cognitive Theory and Social Capital Theory (Chiu, C. M., Hsu, M. H., & Wang, E. T., 2006).

II. Methods

2.1 Models

Random forest classification model is built using Python 3.6 and scikit-learn package 0.19.1. Data is split into training set and testing set, and performance of models are

evaluated by accuracy score and cross validation score. To train the best classifier, four models are built:

- Default Model: Use default setting of random forest classifiers;
- Weighted Model: Adjust sample weights for expert users and non-expert users;
- GridSearchCV Model: Change max depth, min samples split and min samples leaf parameters for random forest classifiers;
- Final Model: Combine best parameters of previous models.

2.2 Data

2.2.1 Data Collection

Zhihu user dataset comes from a public GitHub repository *https://github.com/MatrixSeven/ZhihuSpider*, where the owner kindly shared the web-scraping data generated by his own *ZhihuSpider*. The spider is written in Java, and the original dataset of 420962 users is in MySQL format.

Only the users_info table is used in the analysis, and Table 2.2.1 is the summary of users_info table:

Table 2.2.1 Summary of Users_Info Table

Variable	Description
name	The user name
address	The location/base
education	The education level
company	The working company
job	The occupation

headline	The motto of user
answer	Number of answers the user provided
question	Number of questions the user asked
article	Number of articles the user wrote
favorite	Number of answers the user starred
agree	Number of upvotes the user received for answers
thanked	Number of thanks the user received for answers
following	Number of users the user is following
followers	Number of followers the user has
topic	Number of topics the user is following
columns	Number of columns the user is following
sex	Gender
weibo	The weibo address of the user
index_url	The profile link of the user

2.2.2 Data Cleaning

There are four main steps in data cleaning:

- Transforming data from MySQL database to csv format;
- Reserving selected variables;
- Handling missing data;
- Converting text data into categorical type.

It is worth noticing that there are lots of missing data in selected variables. Some of them are missing when scraping the profile, so this subset of records is filtered due to inactivity. However, the other part of the missing data is because users prefer not to demonstrate personal information, so this subset of entries is treated as NA values, a special category, rather than ignoring them.

The purpose of converting text data is to utilize available data to the maximum extent. As natural language processing method is not utilized in classifying expert users, it would be better to convert them into categorical variables so that they can be useful features in modelling.

After cleaning, there remains 420949 unique users for analysis. Figure 2.2.2 is a screenshot of the dataset after cleaning, including all the variables that are used in the analysis.

user_id	name	headline	answer	question	article	favorite	agree	thanked	following	followers	topic	columns	sex	weibo
430741	李开复	0	107	6	1	0	96117	22401	201	981917	28	0	0	1
339335	黄继新	1	782	1334	95	44	75274	20039	9608	789897	135	635	1	1
392321	周源	1	341	612	8	7	42553	10132	1876	752113	160	154	1	0
675267	yolfilm	1	1509	106	2	10	835981	198641	226	732463	134	59	1	1
337598	张亮	1	1407	1711	98	4	187148	39908	2218	697974	104	88	1	1
392163	李淼	0	1157	47	121	5	347455	67016	756	623385	196	55	1	1
420717	采铜	1	981	101	75	11	569696	134148	1050	580736	26	94	1	1
384961	葛巾	1	34	1	14	0	168827	47313	312	580650	11	6	1	1
367420	朱炫	1	196	4	47	2	1128626	245011	204	579459	37	12	1	1
392249	maggie	1	591	84	14	13	168648	63614	593	552459	32	43	1	1

Figure 2.2.2 Snapshot of Dataset After Cleaning

2.2.3 Adding Labels

To perform random forest classification, an expert label for each user is needed.

Zhihu adopts an unconventional verified-user policy: every user can provide identity proof materials and apply for the verified symbol. Some of the expert users haven't applied for the verified symbol yet, still, they are recognized as expert users. And some of the non-expert users obtained the verified symbol, but clearly, they are far away from expert users. Therefore, the verified symbol is not equivalent to the user's expertise, and other data sources are necessary for mapping labels to users.

A h-index is introduced instead of official personal verified symbol for measuring user expertise. The h-index is defined as: *at least h number of answers of this user received at least h number of agreed*, which is similar to the h-index in academics. More importantly, h-index is considered as a more accurate measurement for user contribution inside *Zhihu* community unofficially.

Although h-index is calculated using *answer* and *agreed*, the indirect impact is acceptable as h-index focuses more on the popularity and quality of each answer, and the *answer* and *agreed* used as features are the total number of *answer* and *agreed* of each user. Taking into consideration that writing answers and receiving agreed as complicated social interaction process, the h-index is qualified as an external data source for expert labels.

A top 1000 user name list based on h-index, which can be found via *https://www.zhihu.com/question/31273136/answer/106466841*, is modified, discussed and released by excellent users in the programming area under this topic. Only the intersection beteen *Zhihu* users_info table and top h-index list in the training dataset receive expert labels, other users in the training set are treated as non-expert users.

III. Result

3.1 Summary Statistics

3.1.1 Demographical Information

Gender, headline and *Weibo* are viewed as demographical information in this study. Headline and *Weibo* are chosen here as they can be treated as binary variables, which are

ideal for modelling. 83.31% of users claim as male; 60.60% of users have headline; 21.79% of users display their *Weibo* account url.

Due to the massive amount of missing data, address, education, company and job are not considered in the following analysis process. Still, a summary table 3.1.1 is provided here to see the typical user profile of this *Zhihu* dataset in terms of address, education, company and job.

Table 3.1.1 Top item of address, education, company and job

Variable	Item	Count
address	Beijing	18751
education	Huazhong University of Science and Technology	947
company	Student	757
job	Product Manager	941

3.1.2 Account Activities

The account activities of users are divided into three dimensions (Richardson, J., & Swan, K., 2003):

- Content Contribution: *answer*, *question*, *article*;
- Social Connection: *following*, *followers*, *favorite*, *topic*, *columns*;
- Popularity and Recognition: *agree*, *thanked*.

As shown in Table 3.1.2, every variable has extremely long tails, which implies that users play different roles in contributing to the community. Most of them are quiet, but those in minority are active. It is clearer when *following* and *followers* are zoomed in. In Figure 3.1.2, the left two plots indicate that most users have less than 20 *followers* and less than 400 *following*. But in the right two plots, there are some super users who have plenty of

following and *followers*, which drags the mean value to numbers significantly different from zero.

Table 3.1.2 Summary of Account Activities Variables

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

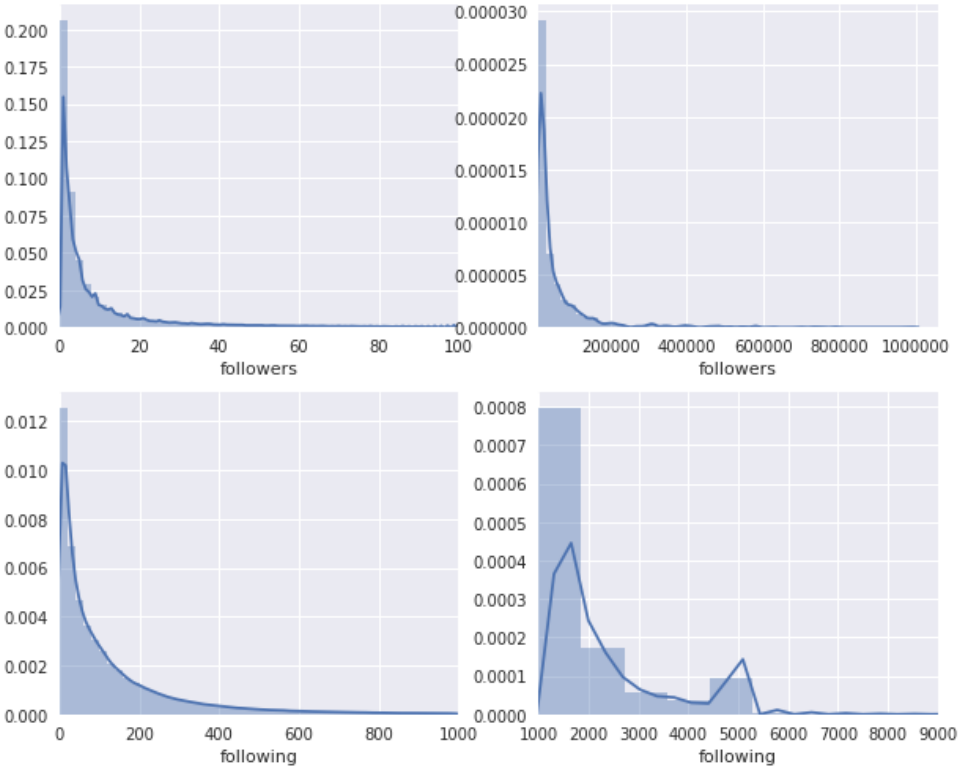

Figure 3.1.2 Density Plots of *Followers* and *Following*

3.2 Correlation Analysis

Figure 3.2 is the correlation matrix of variables involved in this study. In general, variables are not correlated with each other, however, some of them are relevant, which is in accordance with expectation:

- *thanked* and *agree* are highly correlated (*r = 0.9628*);
- *followers* and *agree*, *followers* and *thanked* are moderately correlated (*r = 0.6382, 0.6576*);
- *sex* and *answer*, *columns* and *following*, *columns* and *topic* are slightly correlated (*r = 0.3173, 0.3848, 0.3639*).

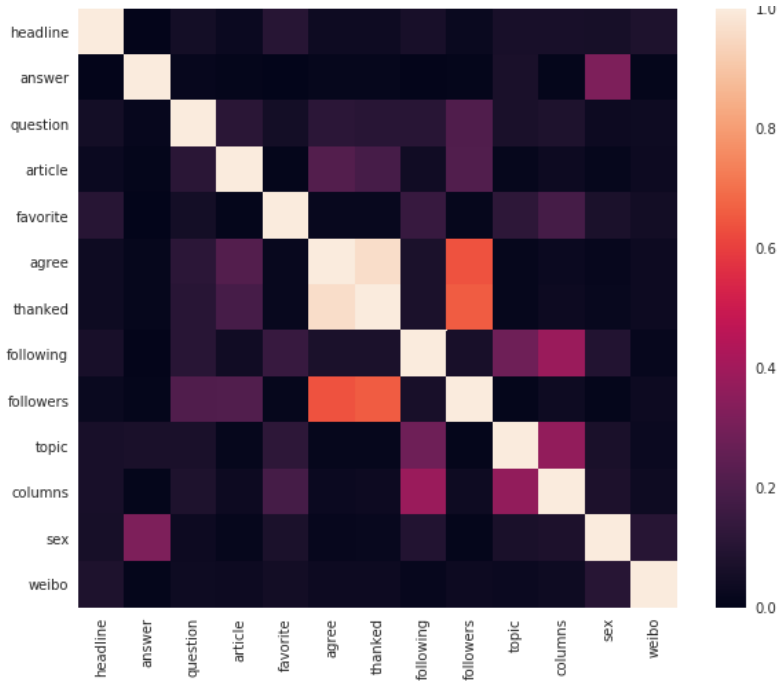

Figure 3.2 Correlation Matrix of Variables Involved

The *thanked* button and the *agree* button are located closely to each other in the website design, and users tend to click *thanked* and *agree* at the same time. The difference is that

the individual number of *agree* of each answer is displayed, while the total number of *thanked* is only visible via visiting user profile page.

Followers are moderately correlated with *agree* and *thanked*, as contents generated by users with more *followers* have higher chances of being promoted to other users' timelines, especially to their followers first pages. When users come across updates of their *following*, they are more likely to express *agree* and *thanked* because of the *following- followers* relationship.

3.3 Random Forest Classifiers

All four models have similar performances due to the imbalance sample size of expert users and non-expert users at first glance. However, the final random forest classifier model does successfully outperform other models.

3.3.1 Model Parameters

Table 3.3.1 shows the comparison between four models, and the final classifier makes full use of adjustable parameters, including class weight, max depth, min samples split and min samples leaf. The optimization process of classifier parameters using GridSearchCV function is crucial, as it returns the best combinations of parameters given a certain search range. Therefore, trees in the last two models are under more control and grow in a more meaningful way.

Table 3.3.1 Comparison of Model Parameters

Parameters	Default	Weighted	Grid+SearchCV	Final
class weight	None	1:10000000	None	1:4
max depth	None	None	5	5
min samples split	2	2	2	2
min samples leaf	1	1	2	2

3.3.2 k-fold Cross Validation and Prediction

Table 3.3.2 shows the mean score for 10-fold cross validation. Because of extreme unbalance between expert users and nonexpert users, these scores are quite similar. Though the default model is slightly better than all other models, the final model outperform the weighted model and the Grid+SearchCV model in terms of 10-fold cross validation mean score.

Table 3.3.2 Mean Score of 10-fold Cross Validation

	Default	Weighted	Grid+SearchCV	Final
score	0.9991466004610275	0.9990454572663492	0.9992034941698982	0.9990644221023931

Figure 3.3.2 displays the correct and wrong prediction rates of four models. The default model has relatively low false positive rate, but extremely high false positive rate, which implies that the default parameters are not effective in detecting expert users. Using an exaggerated class weight, the weighted model reduces the false positive rate, increase the false positive rate and the correction prediction rate at the same time. However, it should be noticed that the cost of reaching the higher prediction accuracy is unaffordable. The Grid+SearchCV model slightly improves prediction performance of default model. When combing the weighted model and the Grid+SearchCV model together, the final model

manages to increase the correct prediction rate by roughly 25% while keeping the balance of low false positive rate and false negative rate and reasonable class weight.

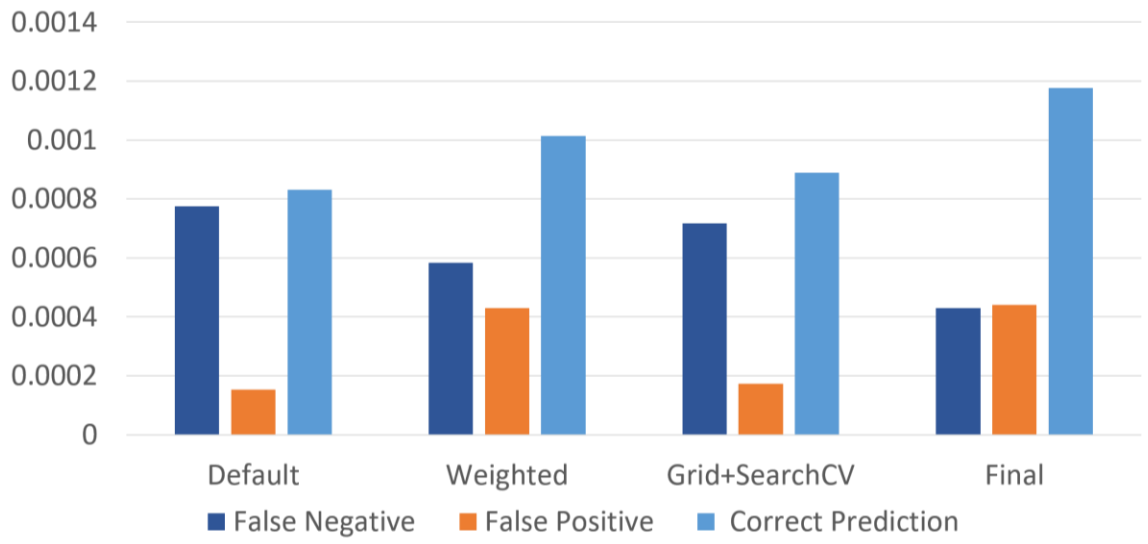

Figure 3.3.2 Correction and False Prediction Rates of Each Models

3.3.3 Features and Importance

Table 3.3.3 shows top 5 features and weights in each model. The default model considers all aspects of user account activities features, including Popularity and Recognition features, Social Connection features and Content Contribution features. When it comes to weighted model, the top features change dramatically along with the extreme class weights, as it concentrates on Social Connection features. The Grid+SearchCV model and the final model focus more on Content Contribution features rather than Social Connection features compared to the default model. Indeed, the most important characteristics of expert users is contributing high-quality contents into the community.

Table 3.3.3 Top 5 Features in Each Model

Default		Weighted		Grid+SearchCV		Final	
Features	**Weight**	**Features**	**Weight**	**Features**	**Weight**	**Features**	**Weight**
agree	0.245828	**following**	0.237009	**agree**	0.399177	**agree**	0.396559
thanked	0.206438	**topic**	0.156363	**thanked**	0.245432	**thanked**	0.260311
followers	0.149980	**favorite**	0.115668	**followers**	0.155634	**followers**	0.174633
answer	0.093799	**columns**	0.112455	**answer**	0.094975	**answer**	0.072039
following	0.062840	**followers**	0.084838	**article**	0.059767	**article**	0.070091

IV. Discussion

4.1 Conclusion

User behaviors have their unique patterns in online Q&A communities. Mental motivation elicits user participation in knowledge sharing, which leads to the success of online open Q&A communities (Guan et al., 2018), and users can be divided into different roles manually according to behavior data (Wang & Zhang, 2016). Being willing to share knowledge as contributors, expert users are extremely essential to the thrive of online Q&A communities. Therefore, it is important to identify expert users and make effort to keep them inside communities.

Previous research use profile information, account activity, linguistic features in questions and answers, content in social media accounts and other type of data to build models to classify expert users. Using profile information and activity summary data, this research successfully builds a random forest classifier detecting expert users in an unbalanced user dataset of *Zhihu*, a well-known Chinese online Q&A community.

By optimizing parameters in random forest classifiers, the final model combines advantages of the weighted model and the Grid+SearchCV model and reaches a relatively

robust state with reasonable parameter values. In the final classifier, both Popularity and Recognition and Content Contribution features are crucial in classifying expert users, but Social Connection doesn't contribute as much as common belief. This finding is useful, as some popular users who always provide "clever" answers are not viewed as expert users in the final model.

This research can help online Q&A 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. If *Zhihu* plan to add expert verified symbol in the future, this model can function as a supportive role in deciding which users are true expert users. The final model also provides evidence for excellent performances of random forest classifiers even in unbalanced dataset, which might widen the application of random forest method in this research area.

4.2 Limitation

Honestly, there are some limitations of this study which could be overcome by future research. The choice of using h-index as expert labels might introduce some confusion into classifiers. Though h-index is indirectly correlated with *answer* and *agreed*, it might have invisible influence on the classifier, especially given the current top 1 feature is *agreed*. Furthermore, using only user profile information might not be sufficient to identify expert users. If time permitted, social network analysis and natural language processing could be utilized to improve the random forest classifier model.

V. Reference

Bouguessa, M., Dumoulin, B., & Wang, S. (2008). Identifying authoritative actors in question-answering forums: The case of yahoo! Answers. *Proceedings of the 14th acm sigkdd international conference on knowledge discovery and data mining* (pp. 866–874). ACM.

Chiu, C. M., Hsu, M. H., & Wang, E. T. (2006). Understanding knowledge sharing in virtual communities: An integration of social capital and social cognitive theories. *Decision support systems*, *42* (3), 1872-1888.

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.

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.

Jurczyk, P., & Agichtein, E. (2007). Discovering authorities in question answer communities by using link analysis. *Proceedings of the sixteenth acm conference on conference on information and knowledge management* (pp. 919–922). ACM.

Pal, A., Farzan, R., Konstan, J. A., & Kraut, R. E. (2011). Early detection of potential experts in question answering communities. *International conference on user modeling, adaptation, and personalization* (pp. 231–242). Springer.

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.

Richardson, J., & Swan, K. (2003). Examining social presence in online courses in relation to students' perceived learning and satisfaction.

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).

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.

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. *Proceedings of coling 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 656–666).