Method & Result: Exploring user behavior patterns based on Zhihu dataset - a Chinese online Q&A community

Andi Liao

1. Data Section

1.1 Data Collection

The Zhihu user dataset used in this research 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.

The data collection process is relatively simple for this study - I directly downloaded the dataset from the repository, and star the contributor as he hoped.

1.2 Data Structure

The dataset contains three tables: follower, users, and users_info. After omitting repeated and irrelevant variables, only part of these three tables are used in this research. Below is the summary:

Table 1.2 Summary of variables in dataset

Table	Variable	Variable Description
follower	user	The user name
	follower	The user's follower name
users	id	The unique id of user
	from_id	The user id which the current user is following
users_info	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

1.3 Data Preprocessing

There are four main steps in data preprocessing: transform data from MySQL to csv format; reserve selected variables; handle missing data; convert 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, therefore, I filter this subset of records as missing indicate inactivity in this case. However, the other part of the missing data is because that user prefer not to demonstrate personal information, so I treat NA values as a special category rather than ignoring them.

The purpose of converting text data is to utilize available data to the maximum extent. As I won't use natural language processing methods in classifying expert users, it would be better to convert them into categorical variable so that they can be useful features in modelling.

After preprocessing, there remains 420949 unique user ids for analysis.

2. Method

2.1 K-Means Clustering

To observe the general pattern of user groups, I first implement a unsupervised learning method, K-Means clustering in scikit-learn package. I expected to see users can be divided into at least two meaningful clusters, expert user and non-expert users.

2.2 Adding Labels

In order to perform random forest classification, a supervised learning method, the next step is to add "expert" label for each user.

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.

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" will be 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.

Inside Zhihu community, the *h-index* is considered as a more accurate measurement for user contribution unofficially. Additionally, a top 1000 user name list based on *h-index*, calculating at 2016, is modified, discussed and released by *excellent users* in the programming area under this topic. The top *h-index* list is the basis of "expert" label adding process.

In this study, I divide the dataset into training, validation and test set, and only the record in training and validation set will be labeled.

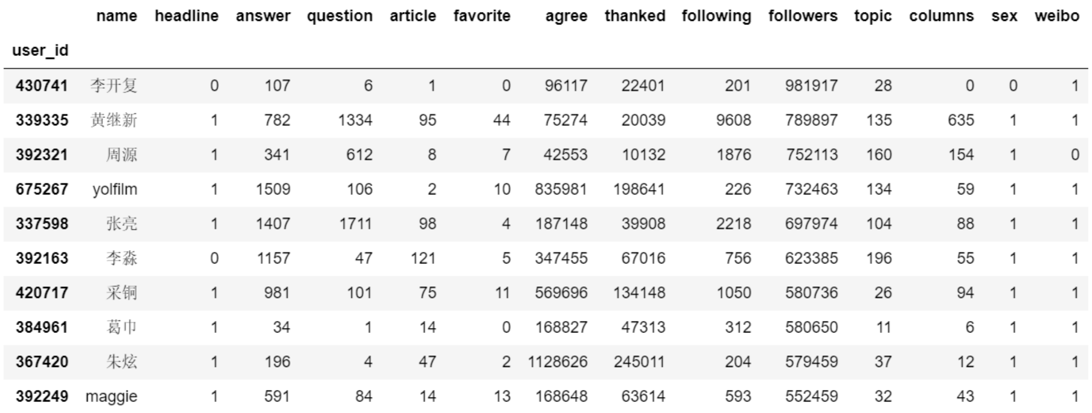
Refernece:https://www.zhihu.com/question/31273136/answer/106466841

2.3 Random Forest Classification and Prediction

The final part of this research is perform random forest classification using training and validation set, and predict whether a user is expert or not in the test set. The model will use the scikit-learn package.

3. Result

Below is a screenshot of the dataset after cleaning, including all the variables that will be used in the analysis.



	name	headline	answer	question	article	favorite	agree	thanked	following	followers	topic	columns	sex	weibo
user_id														
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

3.1 Summary Statistics for Key Variables

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. Among the dataset, 83.31% of users claim as "male", and rest of them are "female"; 60.60% of users have headline, but others don't; 21.79% of users display their weibo account url, while others don't.

Due to the massive amount of missing data, address, education, company and job will not enter the following data analysis process. Still, a brief summary is provided here to see the typical user profile of this Zhihu dataset.

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:

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

As shown in the table and figure below, every variable has extremely long tails, which implies that users play really different roles in contributing to the community. Most of them are quiet, but those in minority are active.

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

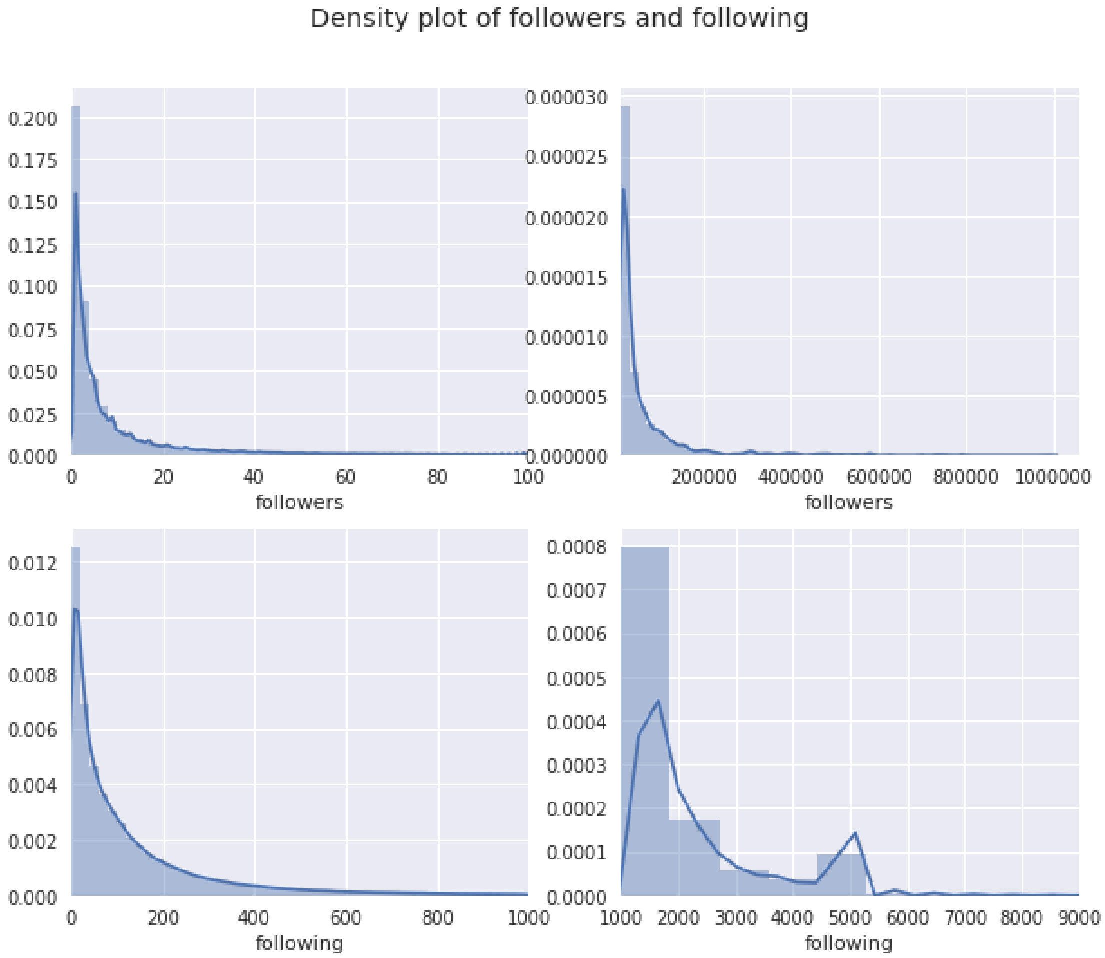


Figure 3.1

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

3.2 Correlation Analysis

Here is the correlation matrix figure of variables involved in this study. In general, variables are uncorrelated with each other, however, some of them are relevant:

- *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($r = 0.3173,0.3848,0.3639$).

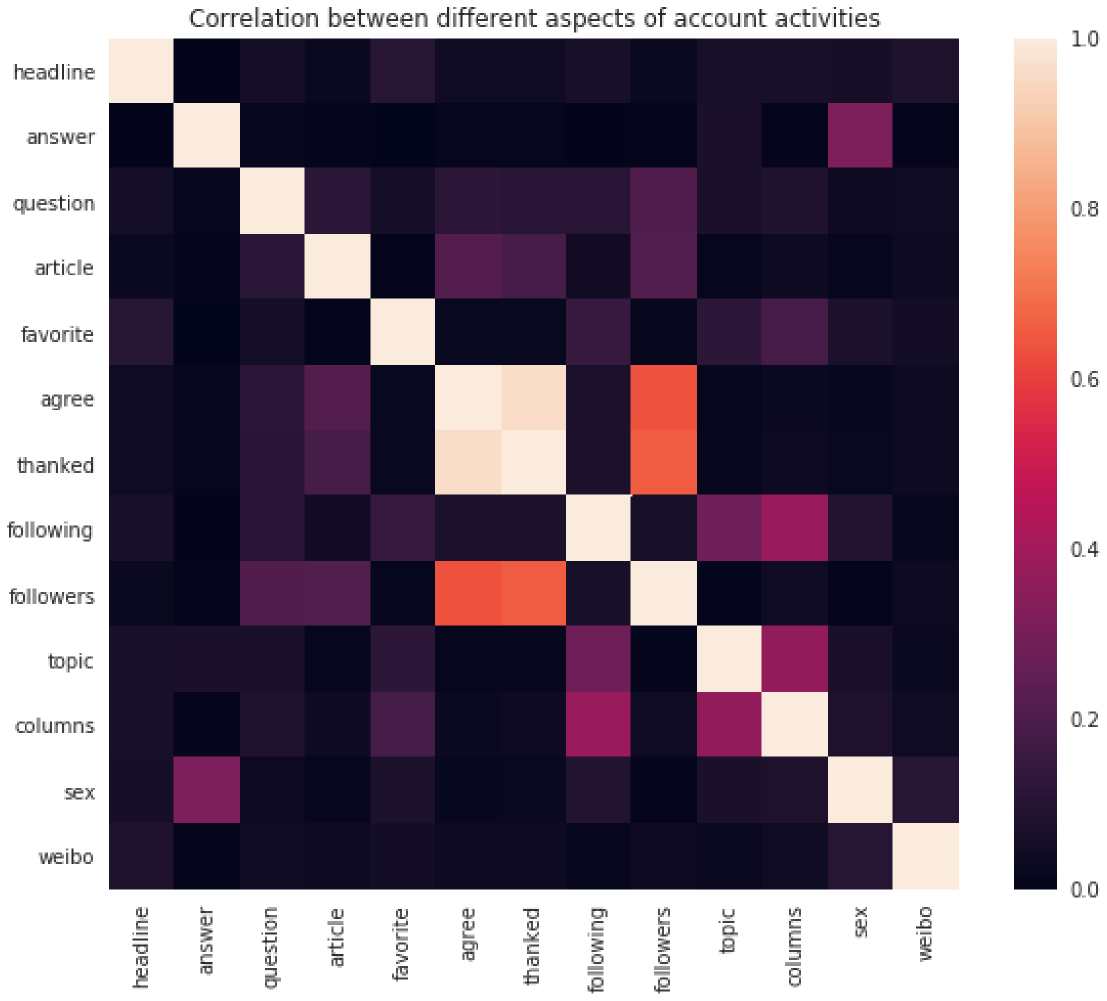


Figure 3.2

3.3 Things to do

Now I am working on adding labels to the training and validation set. A few things to expect: * Comparing expert users and non-expert in three account activities dimensions * K - Means clustering * Random Forest Classifiers