Expert Identification in Community Question Answering: Exploring Question Selection Bias

Aditya Pal, Joseph A. Konstan
GroupLens Research
University of Minnesota
Minneapolis, MN 55455, USA
{apal, konstan}@cs.umn.edu

ABSTRACT

Community Question Answering (CQA) services enables users to ask and answer questions. In these communities, there are typically a small number of experts amongst the large population of users. We study which questions a user select for answering and show that experts prefer answering questions where they have a higher chance of making a valuable contribution. We term this preferential selection as *question selection bias* and propose a mathematical model to estimate it. Our results show that using Gaussian classification models we can effectively distinguish experts from ordinary users over their selection biases. In order to estimate these biases, only a small amount of data per user is required, which makes an early identification of expertise a possibility. Further, our study of bias evolution reveals that they do not show significant changes over time indicating that they emanates from the intrinsic characteristics of users.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: User/Machine Systems - theory and models; H.3.3 [**Information Search and Retrieval**]: Information Filtering - General

General Terms

Algorithms, Experimentation, Human Factors

Keywords

Expert Identification, Selection Bias, Question Answering

1. INTRODUCTION

Community Question Answering (CQA) services provide a platform for Internet users to form social communities and exchange knowledge in the form of questions and answers. In these communities, there are typically a small number of highly active users, and an even smaller number of domain experts - users who provide a large number of technically correct, complete and reliable answers. Identifying these experts during initial phase of their engagement can lead to approaches to retain, mentor these users.

In this paper, we examine how users select questions for answering and propose the notion of *question selection bias*. We hypothesize that experts who aim to provide answers which would be perceived by the community as valuable, tend to prefer answering questions which don't already have good answers. We term this preference as question selection bias and use it to identify experts. The main contributions of this paper are as follows: First we propose a mathematical model to capture this bias in CQA. Second, we propose using a simple machine learning model that uses this bias for the task of expert identification. Third, we show that these biases can be effectively estimated using user's initial interactions making early identification of experts a possibility. Fourth, we show that selection biases do not show significant changes over time, leading us to conclude that these biases stem from the intrinsic characteristics of the users.

2. RELATED WORK

Expert identification approaches employ a graph analysis in conjunction with text analysis. Graph based approaches model CQA as a graph induced as a result of a users' interactions. Zhang [4] modeled CQA as an expertise graph and proposed *Expertise Ranking* algorithm. Jurczyk [3] identified authorities using link analysis of the underlying graph.

Other approaches have looked at overall interaction characteristics of user. Bouguessa [1] proposed a model to identify authoritative actors based on the number of best answers provided by them. Zhang [4] proposed a measure called *Z-score* which combines the number of answers and questions given by a user to a single value in order to measure the relative expertise of a user. Topic-based models to identify appropriate users to answer a question have recently been proposed by Jinwen [2].

3. DATASET

We collected data from *TurboTax Live Community*[1] (TurboTax) which is a Q&A service related to preparation of tax returns. The dataset consists of 633,112 questions provided by 525,143 users and 688,390 answers provided by 130,770 users over the years 2007-2009. TurboTax has employees that manually evaluate an expert candidate on factors, such as correctness and completeness of answers, politeness in responses, language and choice of words used. As of now, they

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.
Copyright 2010 ACM 978-1-4503-0099-5/10/10 ...$10.00.

[1] https://ttlc.intuit.com/app/full_page

have labeled 83 experts and intend to label many more. For our experiment, we selected users who have provided 10 or more answers (U_{10} users). There are 1,367 such users in the dataset. Table 1 summarizes interaction characteristics of the U_{10} users and experts amongst them.

$\lvert U_{10}\rvert$	$\lvert A_{10}\rvert$	$\lvert BA_{10}\rvert$	$\lvert U_{10e}\rvert$	$\lvert A_{10e}\rvert$	$\lvert BA_{10e}\rvert$
1367	226539	23662	83	177426	20731

Table 1: Dataset description. U_{10} are users who provided 10 or more answers. U_{10e} are labeled experts amongst them. A, BA stands for answers and best answers, respectively, provided by them.

4. EXISTING VALUE ON A QUESTION

Answers to a question contain several attributes that reflect their relative value from the perspective of the community such as votes given by the community members, answer status (e.g. best answer, helpful answer). Based on this, we can define the value on a question in terms of the value of answers (V_a) provided on it:

$$V_q = \sum_a V_a = \sum_a w_0 \cdot \text{votes}(a) + w_1 \cdot \text{status}(a) \qquad (1)$$

where status$(a) \in \{1, 2\}$ is 2 if answer has a special status (best answer, helpful answer, etc) otherwise 1. Note that the value on a question increases as more answers are provided but the value can decrease if answers get negatively voted. The weight parameters w_0, w_1 are scaling constants which are used to adjust the relative priority of one attribute over other. For our experiments, we choose $w_0 = w_1 = 1$ (as this works best). Value is discretized into six buckets using the following strategy: If the value on a question is less than 1 it is assigned bucket 0, if it is larger than 4 it is assigned bucket 5 else it is assigned the respective bucket number from 1-4.

The value as defined here is a simple approximation of the true value on a question. Additionally this value is consistent across all users in order to keep the model simple. Our results indicate that the model performance is surprisingly promising even with this simplistic assumption.

5. QUESTION SELECTION BIAS

We explore question selection bias by measuring the degree to which a user prefers answering a question with a given existing value. The existing value gives a clue as to whether the user intends to make a valuable contribution or not. For example, if the user intends to make a valuable contribution, then with high chances user will choose a question with low existing value. Our hypothesis suggests that this tendency of picking a question with low existing value is prominent in experts. Ordinary users either do not have enough expertise to answer a question, or are not well motivated to put effort in answering the question at length. So it might be the case that some of them prefer to pitch in only when question has received few good answers.

In order to measure user's selection bias, we consider all the answers provided by that user. Let a user u choose to answer a question q with existing value v_0. Then, we get evidence that u prefers to answer question with value v_0:

$$P(V_{uq}|A_{uq}{=}1) = \begin{cases} 1 & \text{if } V_{uq} = v_0 \\ 0 & \text{otherwise} \end{cases}$$

where V_{uq} is a discrete random variable indicating the existing value of the question q just before u posted his or her answer. A_{uq} is a random variable that takes values $\{0, 1\}$ indicating whether user u answered question q. Using Bayes rule, we estimate the user's selection bias as follows:

$$P(V_u|A_u{=}1) = \sum_q P(V_{uq}|A_{uq}{=}1) \cdot P(A_{uq}{=}1|A_u{=}1) \qquad (2)$$

where $P(A_{uq}{=}1|A_u{=}1)$ can be considered as the prior probability that u selects q for answering. The prior is considered to be uniform over all the questions answered by u, hence it has an averaging effect. The bias distribution provides insight into the user's selection process: if $P(V{=}0|A{=}1)$ is high then the user prefers to answer questions with no value; if $P(V{=}1|A{=}1)$ is high then the user prefers answering questions with some existing value. Choosing a discrete distribution for bias rather than a continuous one has the advantage that the data required to learn a user's bias is not very large.

5.1 Machine Learning Models

We use biases to define feature vector for each user:

$$x_u = [P(V_u = v|A_u) : \forall v, v \neq max(V_u)]^T \qquad (3)$$

Based on a user's feature vector, that user is classified as either an expert or an ordinary user. We use *Ridge regression* and *Logistic regression* for the task of binary classification. We also use Generative Model based on Gaussian distribution as described below.

5.1.1 Inference using Gaussian Model

This model assumes that users are Gaussian distributed in terms of their selection biases. Figure 1 shows that the scatter plot of first two features for the dataset and the contours of the Gaussian distribution fitted based on the MLE estimates. The smaller contours captures contains majority of the experts and some of the ordinary users as well. It also shows that biases of experts are concentrated in a small region and for other users it is more widespread. The model parameters of Gaussian distribution are $\theta = \{\mu, \Sigma\}$.

$$P(x|\theta) = \frac{1}{(2\pi)^{\frac{|x|-1}{2}} |\Sigma|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\} \qquad (4)$$

The model parameters for the two classes are $\theta_E = \{\mu_E, \Sigma_E\}$ and $\theta_O = \{\mu_O, \Sigma_O\}$. We assume that a user's bias is i.i.d which simplifies Maximum Likelihood Estimation (MLE):

$$\theta^{MLE} = argmax_\theta\{P(D|\theta)\} = argmax_\theta\{\prod_u P(x_u|\theta)\} \qquad (5)$$

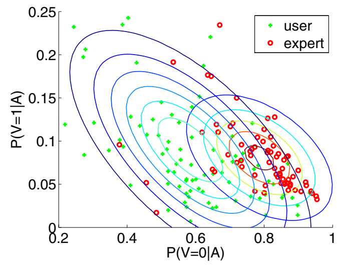

Figure 1: Plot of first two features of biases and the contours of Gaussian distribution using MLE.

For classification, Bayes rule is used to generate the posterior distribution of class conditioned on feature vector:

$$P(E|x_u) \propto P(x_u|\theta_E^{MLE}) \cdot P(E) \tag{6}$$

where $P(E)$ is prior probability of a user being an expert. Prior probability is the ratio of experts in the training data. We also compute posterior probability of user belonging to the other class, whichever class has a higher probability.

6. EXPERIMENTS

We use 10-fold cross-validation to run the learning models. To compare model performance, we report on the precision, recall and $F1$ score of models in *prediction of experts*.

6.1 Bias Analysis

Figure 2 compares the average biases of experts with those of other users. The biases of experts are significantly different for $v = 0, 5$ (t-test, $p \approx 0$). Similarly, for $v = 1$ bias means are different (95%CI, $p < 0.04$). It shows that experts are more selective in picking questions with zero existing value (no answers or bad answers on the question).

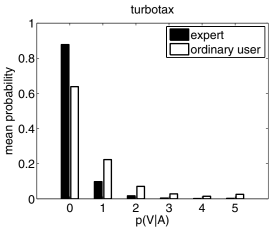

Figure 2: Average selection bias of users.

6.2 Model Comparison

We consider a few other models that extract different features of users and compare their performance with our model. Following is a brief description of the other models:

Z: This model is based on Z-score [4] ($Z(a, q) = \frac{a-q}{\sqrt{a+q}}$, where a and q are the number of answers and questions given by a user). The feature vector for user is $[a\ q\ Z(a, q)]^T$.

T: A model based on text analysis of answers. It includes features such as the use of positive/negative emotions, self-reference, big words, and words referring to categories such as religion, gender, etc. We also add other features such as the number of answers and the votes received.

We use the Gaussian model over features extracted from all the different models for classification. Table 2 shows the model performance in predicting experts. We observe that the recall of B is significantly higher than both T and Z, indicating that the experts are tightly clustered over their biases, and our formalism of bias is extremely effective in retrieving them. Z+B model performs better than most models, indicating that biases along with simple features can boost predictive performance significantly.

	T	Z	B	Z+B
Precision	**0.81**	0.60	0.28	0.62
Recall	0.27	0.47	**0.92**	0.87
F1 score	0.40	0.53	0.43	**0.72**

Table 2: Model performance. B is bias-based model.

Low precision of B indicates that there are many other users in the community who show similar selection tendencies as experts. On a deeper analysis, we find that several of the TurboTax users have given 10-15 answers which could have led to an inaccurate estimation of biases for these users. If we restrict our focus to users with a higher number of answers (Section 5.3), we see that the model performance improves. Another perspective is that TurboTax has a manual expert identification process and they have not yet completely labeled all the experts in the community, so several of the false positive can turn out to be experts in future.

6.3 Considering Answer Threshold

In the previous experiment, we established that experts exhibit similar biases, and leveraged that fact to identify experts. The precision of the bias-based model did not turn out to be significantly high. The main reason is that several of the U_{10} users have given only 10-15 answers. Bias estimation for such small number of answers might not be accurate. In this experiment, we run our model over users who have provided N answers or more. We call N the answer threshold. The answer threshold serves two purposes: First, as we increase N, the ratio of experts to ordinary users in the pool of selected users increases. This is because experts typically give many more answers. Second, inaccurate biases are eliminated from the data. Figure 3 shows that the model performance improves with increase in answer threshold. At high answer threshold, model precision improves by 250% (0.7 vs 0.28). The Gaussian model beats the two regression models indicating that the users are indeed Gaussian distributed over their biases.

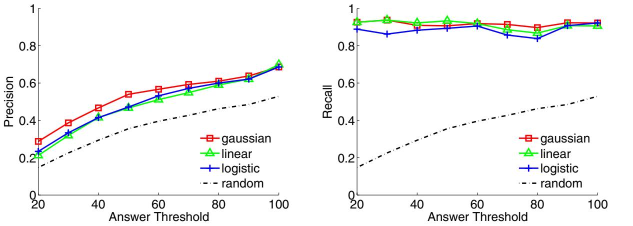

Figure 3: Performance of bias-based learning model.

Additionally, it shows that by carefully choosing the answer threshold, we can strike a balance between model performance and amount of data required before predicting a user as an expert or not. Since recall doesn't consider those experts who were discarded due to threshold, an effective tradeoff can be devised by maximize the product of F1 score and the number of experts amongst selected users. This leads us to the answer threshold of 30-40 answers as best.

6.4 Performance over First Month data

Here we consider the answers given by users within the month of their first answers to compute their biases. Figure 4 shows the performance of model over this data. It shows that the model performance has improved considerably over Figure 3. This presents evidence that biases can be effectively used to identify experts while their association with the community is in its early stages. This can be pretty useful in identifying potential experts by providing them retentive incentives to avoid them from churning.

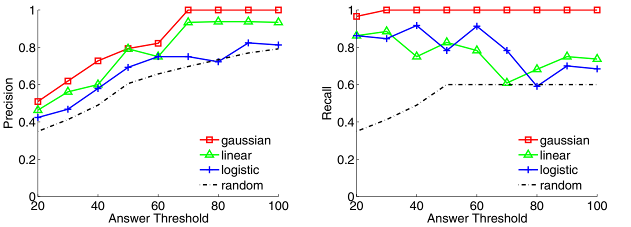

Figure 4: Performance of bias-based learning model using one-month data.

6.5 Bias Evolution

CQA systems are extremely dynamic in nature - several factors change over time, such as interfaces, functionality, users, user's interests, and activity patterns. Under such dynamics, the selection preferences of a user can change. We study these preferences by dividing the data into 5 equal time-slots and computing biases per slot. We plot the bias mean and deviations for all users (Figure 5). The biases do not show any noticeable changes over time (same result over all other dimensions). This indicates that biases are not influenced by dynamics of CQA even over longer period of time. This provides some evidence to our claim that these biases emanate from the intrinsic characteristics of the user.

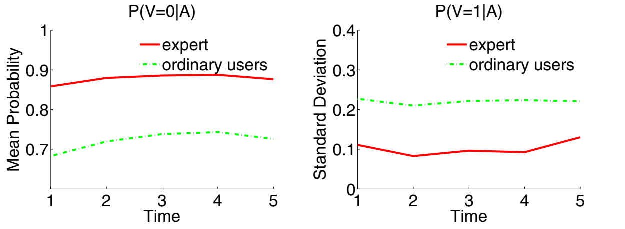

Figure 5: Bias mean and standard deviation (first dimension only) of users across five time windows.

6.6 Other Datasets

We also performed experiments on dataset collected from *Stackoverflow.com*, which lead to same conclusion as we get above with the TurboTax dataset. Due to space constraints we do not present our results on these other datasets.

7. CONCLUSION

In this paper, we present question selection bias as a new measure to study the behavior of users in CQA. In our setting, this bias indicates the degree to which users prefers to answer questions in different stages of answer completeness. This captures the intuition that some users prefer to answer questions where they can create the largest value - the questions with no good answers. On the other hand, some users prefer to answer questions with some existing value, perhaps because they do not have enough expertise to give a complete solution or because they do not want to put much effort into answering. Our results show that experts select with high probability questions with low existing value. We also show that these biases emanate from the intrinsic characteristics of user and do not get influenced by the dynamics of the underlying community.

This paper also establishes that a user's selection bias can be effectively used to identify users who have the potential of expertise in their early stages of engagement with the community. Identifying potential users can be extremely useful, as these users can be more effectively motivated or mentored to reach the expertise level faster.

We also show that bias can be mixed with other simple measures to improve the predictive power of expert identification models ($Z+B$ model). Though the labeling of experts in the TurboTax dataset is partial, we get significant predictive performance. It not only justifies this first formal step towards capturing user biases, but motivates us to explore biases of several types and their effects on the community.

Last but not the least, we show that employing a minimum threshold of 30-40 answers is optimal for measuring biases. This makes the task of expert identification computationally feasible for large datasets.

Future Work

Selection biases present an interesting metric in understanding the psyche of a user in the community. Q&A interfaces can be personalized to show questions conforming to each user's bias. Such steps can increase the participation of users. We would also like to remove several simplifying assumptions made in this paper, in order to model user's behavior more accurately.

Acknowledgements

We would like to thank F. Maxwell Harper for suggestions towards improving the paper. This work was supported by the National Science Foundation, under grants IIS 08-12148 and IIS 08-08692.

8. REFERENCES

[1] M. Bouguessa, B. Dumoulin, and S. Wang. Identifying authoritative actors in question-answering forums: the case of yahoo! answers. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 866–874, New York, NY, USA, 2008. ACM.

[2] J. Guo, S. Xu, S. Bao, and Y. Yu. Tapping on the potential of q&a community by recommending answer providers. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 921–930, New York, NY, USA, 2008. ACM.

[3] P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 919–922, New York, NY, USA, 2007. ACM.

[4] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230, New York, NY, USA, 2007. ACM.