Analyzing Features that Drive a Quality of Answer on a Question Answering webpage Stats.StackExchange

Miroslav Král
MSc Business Analytics
University College London
Miroslav.kral.15@ucl.ac.uk

Agus Nur Hidayat
MSc Business Analytics
University College London
Agus.hidayat.15@ucl.ac.uk

Abstract

This paper presents study on the question answering webpage Stats.StackExchange. This is a webpage where users can ask a question on topic from statistic or machine learning and other members of the community can answer the question.

In this paper we analyze how to recognize experts or influential users in the Stats.StackExchange network. We compared different methods on how to identify these experts and we conclude that number of accepted answers owned by users is a better measurement to determine whether the user is an expert or not. We used this knowledge to help users becoming experts in the community by analyzing what makes an answer to be more likely accepted. The result is a list of features with their strength of significance in affecting the probability of answer to be more likely accepted. Our analysis can help users to reach high reputation in the Stats.StackExchange network.

1 Introduction

The question answering (Q&A) webpages have become widely popular in recent years. These pages are the place where users can ask question and other users answer the question. There exist many different subtypes of Q&A webpages. It is possible to distinguish between pages on specific topic (for instance statistics or computer science) and the webpages on general topic (example of such a webpage is Yahoo!). Some authors also divide Q&A webpages based on the fact who answer the question on *Digital reference services* Q&A pages, *Ask an expert services* Q&A pages and *Community* Q&A pages [1]. In this paper we work with Community Q&A webpage that is on statistics topic and it is called Stats.StackExchange (www.stats.stackexchange.com).

Stats.StackExchange is a Q&A online community webpage that will mainly talk about topics in Data Science. According to an article from Harvard Business Review, Data Science is considered to be the sexiest job of the 21st century [2]. Being an expert in such a prospective domain as Data Science can boost the

employability of the user. Within the webpage, a user can ask a question and other ones can answer it. User who asks the question can mark one of the answers as an accepted answer. Questions and answers are stored and can be accessed by anyone who browse the webpage. Therefore, being a user in Stats.StackExchange that can provide a large number of accepted answers can have positive impact as more people will recognize the expertise of such a user in the field of Data Science.

Since Stats.StackExchange is a technical Q&A webpage, many of the answers require deep knowledge of the topic. To distinguish users with high level of expertise, Stats.StackExchange implemented ranking system that enables the marking of users with deep domain knowledge. Rank of users is based on their activities on the webpage and how their posts being helpful to other members of community (other users can up-vote or down-vote post of the user).

In this paper, we firstly look at who are the most valuable users of the community by analyzing their interaction within the network. For the evaluation we use PageRank algorithm and Out-Degree measurement from social network analysis and we compare our analysis with the ranking system that is already implemented in the webpage. Secondly, in this paper, we have the main goal to understand what are the features that drive an answer to be more likely accepted. Result of this paper could help users in Stats.StackExchange to provide an answer that has higher probability to be accepted.

2 Related Work

Social networks and also question answering webpage networks are subject of research of many scientists. Due to the fact that probably the two biggest Q&A pages are Yahoo! and StackOverflow, most of the papers are related to these two pages.

Common topic of most of the papers is how to identify experts and influential people in the network. Movshovitz-Attias et al. (2013) analyzed build-in reputation system of StackOverflow. They talked about how can users up-vote and down-vote posts of different users and another detail of the reputation system. They also implemented PageRank algorithm to discover important users in the network. During their analysis, they discover that there exists a correlation between user's activity in first few months since registering at StackOverflow and the expertise of user. Therefore, they build a model that can predict whether a user will be recognized as an expert in network based on his initial activity [3]. Jurczyk and Agichtein (2007) presented a study where they used link analysis to discover authorities in Yahoo! Q&A webpage. They compare two methods, HITS and Degree analysis in identifying the authorities and they concluded that HITS is better method. They discovered authorities across different topics [4]. Ray, Dey and Gaonkar (2011) in their paper compared StackOverflow webpage with Enterprise Social Network Platforms in terms of behaviour of users and presented framework how to recognize experts in both networks [5].

Few studies were also focused on quality of answers at Q&A pages. Harper, Raban and Rafaeli (2008) performed an experiment where they compared quality

of answers at free webpages such as StackOverflow and paid Q&A webpages. They conclude that free webpages where every community member can answer the question has better quality answers [1].

Overall, online Q&A pages offer wide range of possibilities of research. For instance, Adamic et al. explored Yahoo! page and explore which topics are related based on interaction users. They clustered categories based on patterns of interaction of the users and they found out that some users are focused only on specific topic while others are active across many topics [6]. On the other hand, Zhong et al. (2013) look at dynamic of the social network and included also Q&A webpages [7].

An important study for this paper is study of Bosu, Corley, Heaton, Chatterji, Carver and Craft (2013) who explore how should users behave on StackOverflow in order to reach high reputation [8]. In their paper they also explain how user of Q&A network can reach high reputation. While they look on this problem in terms of users behaviour (for which type of question should users write answers), our approach is different and we also identify how answers should look like.

In this paper we first identify expert users in the Stats.StackExchage using two methods, that are PageRank algorithm and Out-Degree measurement from social network analysis. We compare our results with the Reputation implemented on Stats.StackExchange webpage. Comparing our results with the one on the webpage allows us to decide which one of these two methods works better. In the second part of this paper, we examine what makes an answer to be more likely accepted. Our analysis can help users to reach better reputation.

3 Dataset description

In this paper we use data from Q&A webpage Stats.StackExchange (can be reached at https://archive.org/download/stackexchange/stats.stackexchange.com.7z). Stats.StackExchange is a Q&A page on the topic of Data Science. In our analysis we used dataset of user interaction on this webpage from time period since 19-07-2010 until 06-03-2016. In other words, we used complete history data of Stats.StackExchange.

Stats.StackExchange data set contains 74,089 questions and 73,540 answers. Out of this 73,540 answers, there are 23,441 accepted ones, that means that these answers were marked by users who asked the questions as the most helpful ones. Altogether, the dataset contains 78,003 users. However, only 28,134 (36%) of them are active on the webpage and wrote at least one post.

4 Identifying experts

We used data about interaction between users to perform link analysis and identify the experts in the network. We compare our results with reputation system implemented on the Stats.StackExchange network.

First link analysis method we use is PageRank. PageRank is an algorithm developed by Larry Page and Sergey Brin at Stanford University in 1996. Formally it can be defined as a long term probability that randomly selected post is written by a particular user. Page rank assign to each active member of

community a number (probability) between 0 and 1. Sum of this probabilities has to be equal to 1 [9].

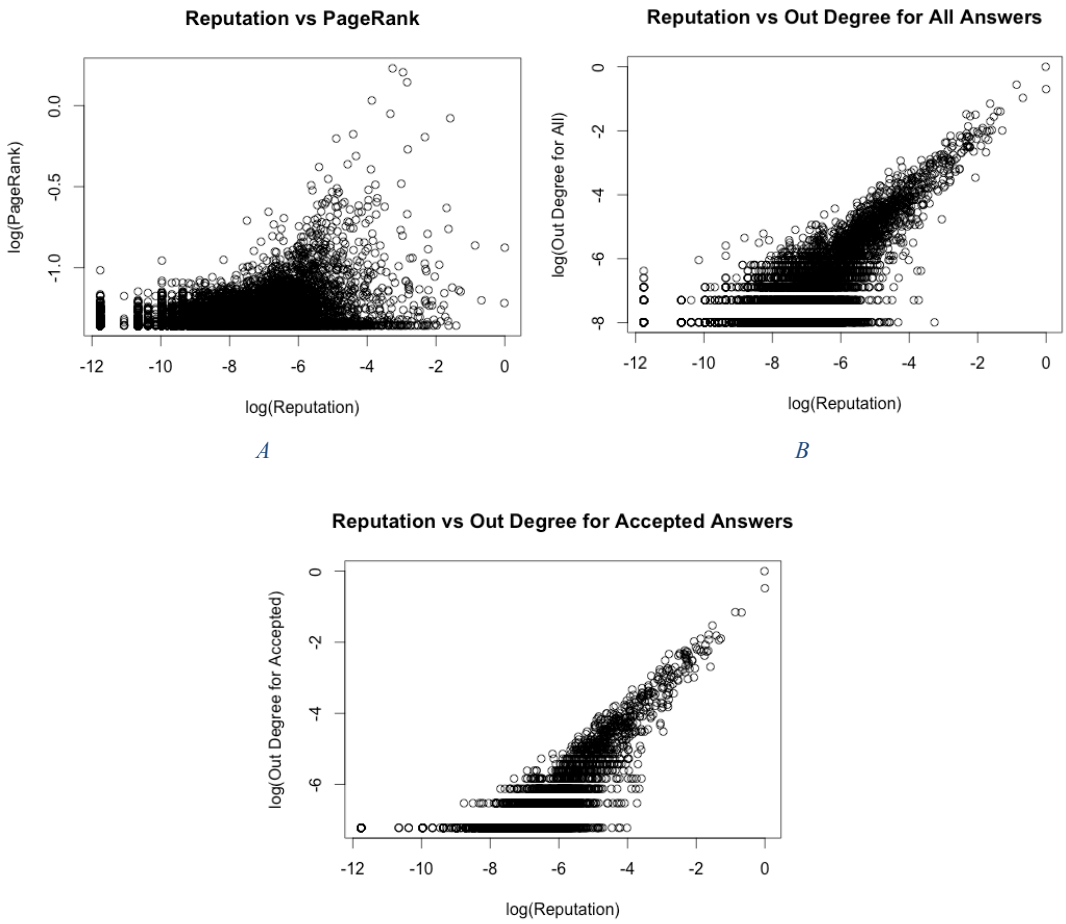

Figure 1 A: Plot captures PageRank of user against Reputation of user; B: Plot captures out-degree of user based on all answers against reputation of users; C: Plot shows out-degree of users based on accepted answers against reputation. All the plots are on log-log scale.

We calculate PageRank for network the 28,134 (number of users who wrote at least one post) users who actively participate on the webpage. In this network nodes are active user of Q&A page (users who wrote at least one post) and oriented edges shows who answer whose question. Figure 1 part A shows how PageRank score we calculated correlate with the reputation of users on Stats.StackExchange. There is not a strong correlation between these two rankings, Pearson's correlation coefficient between these two rankings is **0.1695**.

We selected out-degree of user, that is how many questions has the user answered, as a second method how to determine who are the experts in community. We decided to calculate out-degree based on the paper from Movshovitz-Attias [3] who showed showed that out degree analysis works good in identifying expert users on StackOverflow Q&A page. We calculate two types of out-degrees for every user:

- out-degree based on how many answers user provided
- out-degree based on how many accepted answers user provided

Figure 1, part B shows the relationship between out-degree of user based on all answers and reputation of the users. Part C shows scatter plot of out-degree of users based on accepted answers that users wrote and reputation of the users. There is strong correlation between both out-degree values and reputation. However, in terms of values of Pearson's correlation coefficients out-degree based only on accepted answers is correlated strongly with users' reputation (**0.9626**) than out degree based on all answers provided **(0.9303**).

By comparing the plots and Pearson's correlation coefficients we conclude that best estimator of users' expertise is how many accepted answer user wrote. There is a strongest correlation between users who wrote large number of answers that were marked as most helpful by the user who ask the question and users that has high reputation in reputation system implemented on the Stats.StackExchange website. This metrics is better estimator than number of answers written by user and PageRank of user.

5 What makes an answer to be an accepted one

Being an expert on Data Science is a valuable and a desirable prospect. Having high reputation on Stats.StackExchange Q&A webpage can be seen as a proof that a person has a deep understanding and knowledge about Data Science and it can boost employability of person significantly.

In section 4 – Identifying experts is shown that there exists strong positive correlation between being a high ranked user of Stats.StackExchange and having large number of accepted answers. Therefore, in this section, we present an overview of which features make answer the accepted answer.

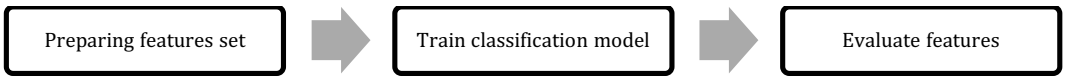

Figure 1 Process diagram of process of determining feature importance

We formulated the task of identifying drivers of accepted answers as a task of building a binary classification model. Given the set of the features we classify the answers into two classes: accepted answer and not accepted one. However, we are not interested in actual performance of the model, we are interested in the relative importance of the features, which features cause that the answer is classified as accepted. Process diagram in Figure 2 illustrates whole process.

I. Preparing feature set

In order to train the model, we prepared training dataset. Training dataset has 73,540 instances, 23,441 of them has class label accepted and rest not-accepted. We crafted 27 different features that can be divided into 3 different types. Although most of the features are self-explanatory, we present list of all features together with explanation of their meaning.

Post – based features:

- **Answer Score** – Number of users that up-vote the answer subtracted by number of users that down-vote the answer.
- **Answer Comment Count** – Number of comments of an answer.
- **Question Favorite Count** – Number of users that mark the question as favorite.
- **Question Total** Tags – Number of tags of the question.
- **Views** – Number of users that have viewed the user's profile.
- **Up-votes** – Number of up-votes given by the user.
- **Down-Votes** – Number of down-votes given by the user.
- **Question View Count** – Number of users that have viewed of the question.
- **Question Comment Count** – Number of comments of the question.
- **Time difference** – The difference of time between the answer and the question.
- **Question Answer Count** – Number of answers of the question.
- **Question Score** – Number of users that up-vote the answer subtracted by number of users that down-vote the question.

User – based features:

- **Out-degree Acc** - Number of accepted answers provided by the user.
- **PageRank** - Relative importance of a node in the network (in this case, a node is a user) based on link analysis algorithm by Google.
- **Time from Register** - Time difference between the registration time of the user and 22th of May 2016.
- **In-degree All** - Number of all questions asked by the user.
- **In-degree Acc** - Number of questions with an accepted answer asked by the user.
- **In-degree Ratio** - Number of questions with an accepted answer asked by the user divided by number of all questions asked by the user.
- **Profile Image** - Whether the user has a profile image or not.
- **Out-degree all** - Number of all answers provided by the user.
- **About Me Length** - Number of words of the user's description.
- **Website** - Whether the user provides a website information in the profile or not.

Text – based features:

- **Answer Body Length** – Number of words of the answer.
- **Question Body Length** – Number of words of the question.
- **Answer sentiment** – The sentiment score of the answer. In order to get sentiment analysis, we train Voting prediction algorithm that interoperate Naïve Bayes Classifier, Multinomial Naive Bayes, Logistic Regression and Support vector machines classifier.
- **Question sentiment** – The sentiment score of the question. The same classifier as in answer sentiment was used.
- **Question Title Length** – Number of words of the question's title.

Features from either the question or the user are for those that is related to the observed answer, that is, the question which the answer responded to and the user who provided the answer.

II. Train classifier

As a classifier we decided to implement Logistic Regression model. Logistic Regression is a commonly used binary classification model. This model is applied to situations where the targeted or dependent variable is dichotomous. Equation 1 below expresses the formula that can fit the model:

$$log_e\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \sum \beta_i x_i$$

Equation 1 Logistic Regression Formula

where i is the number of predictor variables and $\pi = \mu_y$ is the conditional mean of y. In the Logistic Regression model, it is important to know that the dependent variable is binary and follows a binomial distribution. Additionally, the Logistic Regression model also requires each observation or row to be independent [10].

III. Evaluate features

From the result of logistic regression, we are interested in p-value and t-value of every designed feature. P-value denotes whether a feature is considered to be a significant predictor or not. The closer the p-value to zero, the more important a feature is. T-value denotes whether a feature gives positive or negative effect toward the class labeling process and how strong that effect is. Both tables below, Table 1 and Table 2, shows result of p-value and t-value for most significant features, the most significant features have p-value **< 2e-16**. Both tables sort the features in descending order based on the strength of the effect. Table 1 below represents features with positive effect:

Features	**t-value**	**p-value**
Answer Score	32.042	< 2e-16
Answer Comment Count	22.461	< 2e-16
Answer Body Length	19.043	< 2e-16
Question Favorite Count	16.917	< 2e-16
Out-degree Acc	12.822	< 2e-16
In-degree Ratio	11.266	< 2e-16
Website	9.996	< 2e-16
Time From Register	7.707	1.30E-14
Question Body Length	5.579	2.43E-08
Question Total Tags	5.046	4.51E-07
Views	2.921	0.00349
Down Votes	1.867	0.06196
In-degree All	1.296	0.19509
In-degree Acc	0.286	0.77475

Table 1 Logistic Regression Result: Overview of features that affect whether the answer will be accepted in positive way. Table capture feature and corresponding t-statistics and p-value

Based on Table 1 above, we can infer that an answer that is more likely to be accepted has characteristics:

1. It has higher score, that is, the number of users who up-vote the answer subtracted by the number of users who down-vote the answer.
2. It has larger number of comments.
3. It contains larger number of words.
4. It is answering a question that is marked as favorite by larger number of users.
5. It is provided by a user that has more accepted answers
6. It is provided by a user who has ratio between number of his/her questions with an accepted answer and number of all his/her questions that is closer to 1.
7. It is provided by a user who has information about website in his/her profile

Table 2 below shows features with negative effect:

Features	**t-value**	**p-value**
Question Score	-22.739	< 2e-16
Time Difference	-21.12	< 2e-16
Out-degree All	-18.943	< 2e-16
Question Answer Count	-12.532	< 2e-16
Question Comment Count	-12.439	< 2e-16
About Me Length	-4.462	8.14E-06
Answer Sentiment	-3.282	0.00103
Question Sentiment	-3.152	0.00162
Reputation	-2.989	0.0028
Question Title Length	-2.891	0.00384
PageRank	-2.47	0.01353
Question View Count	-1.494	0.13517
UpVotes	-0.8	0.42351
Profile Image	-0.605	0.54535

Table 2 Logistic Regression Result: Overview of features that affect whether the answer will be accepted in negative way. Table capture feature and corresponding t-statistics and p-value

Based on Table 2 above, we can also infer that an answer that is less likely to be accepted has characteristics:

1. It is answering a question with higher score, that is, the number of users who up-vote the question subtracted by the number of users who down-vote the question.

2. It has larger time difference, that is, the difference of time between the answer and the question.
3. It is provided by a user who has a larger number of answers
4. It is answering a question with larger number of answers.
5. It is answering a question with larger number of comments.

6 Conclusion

Question answering webpages are source of valuable information. In this paper we presented analysis of one of the Q&A webpages – Stats.StackExchange.

In our study we focused on two goals. In the first part of the paper we compared 2 methods for identifying expert users in the network – PageRank and Degree analysis method. Our analysis showed that out degree of the user based on accepted answers (that is how many accepted answers user has) is best for determining influential users in the network.

Due to the fact that number of accepted answer is the best for predicting experts, we examined in the second part of this paper what makes an answer accepted answer. We trained logistic regression model and compare importance of the predictors and determine features of an answer that makes it more probably to be accepted.

Reference:

1. Harper, F. M., Raban, D., Rafaeli, S., & Konstan, J. A. (2008, April). Predictors of answer quality in online Q&A sites. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 865-874). ACM.
2. Davenport, T. H., & Patil, D. J. (2012). Data Scientist: The Sexiest Job of the 21st Century-A new breed of professional holds the key to capitalizing on big data opportunities. But these specialists aren't easy to find—And the competition for them is fierce. *Harvard Business Review*, 70.
3. Movshovitz-Attias, D., Movshovitz-Attias, Y., Steenkiste, P., & Faloutsos, C. (2013, August). Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on* (pp. 886-893). IEEE.
4. Jurczyk, P., & Agichtein, E. (2007, November). 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. Raj, N., Dey, L., & Gaonkar, B. (2011, August). Expertise prediction for social network platforms to encourage knowledge sharing. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on* (Vol. 1, pp. 380-383). IEEE.
6. Adamic, L. A., Zhang, J., Bakshy, E., & Ackerman, M. S. (2008, April). Knowledge sharing and yahoo answers: everyone knows something. In*Proceedings of the 17th international conference on World Wide Web* (pp. 665-674). ACM.
7. Zhong, E., Fan, W., Zhu, Y., & Yang, Q. (2013, August). Modeling the dynamics of composite social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*(pp. 937-945). ACM.
8. Bosu, A., Corley, C. S., Heaton, D., Chatterji, D., Carver, J. C., & Kraft, N. A. (2013, May). Building reputation in stackoverflow: an empirical investigation. In *Proceedings of the 10th Working Conference on Mining Software Repositories* (pp. 89-92). IEEE Press.
9. Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval* (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press.
10. Hosmer Jr, D. W., & Lemeshow, S. (2004). *Applied logistic regression*. John Wiley & Sons.