Expert Mining in Stack Overflow

Ankit Joshi, Tanvi Mainkar
Department of Computer Science
North Carolina State University
Raleigh, NC 27695
ajoshi5@ncsu.edu, tmainka@ncsu.edu

***Abstract*—On Stack Overflow (SO)[1], which is one of the most popular technical Q&A forums [1], users may sometimes face a problem of whether or not they can trust a solution posted by other users. A user can safely trust a solution only if it has been posted by an expert in a particular field. Reputation of a user on SO is a good indicator of his expertise. However, the reputation of a user is calculated based on his contribution across all technologies.[2] Therefore, it is not a clear indicator of his expertise in a particular technology. We needed a mechanism that would help us in determining whether or not a user is an expert in a specific technology. This project explores a method, which uses the net number of votes a user has gained, by answering questions related to a technology, to determine his expertise. We assumed that the users above a particular score are experts in a specific technology. We then supported this assumption, by showing that these users demonstrate behaviors, that are usually the characteristics of experts. Now, a user would be able to look at the technology score of a particular responder and decide whether or not he should place his trust on the responders solution.**

I. Introduction

Stack Overflow (SO) provides a platform for programmers to discuss and seek solutions to various problems in a variety of technologies.[1] Since SO has a very active user base, majority of the problems receive multiple solutions. One of the concerns, especially for new users on SO, is whether or not they can trust the solutions provided by other users. A new user can safely trust a solution only if it has been posted by someone, who can be thought of as an expert in the field. Thus, determining the expertise of a user in a specific technology is very important, especially from the point of view of a new user to SO.

On SO, the **Reputation** of a particular user serves as a measure of how much the community trusts the user. The primary way for any user to gain reputation on SO is by posting good questions and useful answers.[2] Depending on whether or not the community members are convinced with the user's questions and answers, they either vote in favor or against his questions and answers, thus affecting the user's reputation. However, the reputation of a user is calculated across all the technologies, for which the user has asked questions or given answers.[2] Therefore, it is not a clear indicator of expertise in a particular technology and we need a way to calculate a user's expertise in a specific technology.

Posts on SO are classified by technology and the net number of votes that a user receives for a particular post play a major role in deciding his/her reputation.[2] Therefore, calculating the total number of votes for all the posts made by a user in a particular technology, may help us in determining the expertise of the user in that particular technology.

To support the above mentioned hypothesis, we assume that users above a particular technology score are experts. We then short-list a set of users, based on certain parameters mentioned below, that can help differentiate between expert users and average users, and then compare this set with our assumed set of expert users. These parameters are as follows:

Average Response Time: One quality of an expert is that he can provide correct answers in short span of time, thus helping the asker to quickly resolve his queries.

Number of Accepted Answers: This parameter gives an idea about the number of answers provided by a user which the askers found the most convincing. More the number of accepted answers given by a user, more is the trust that can be put on his expertise in given subject.

Number of Answers that received the Highest Up-Votes for a particular question: This parameter can be used to find out the expertise of a user on SO, since one of the quality of an expert is his ability to express his answers in a way that appeals to the community.

Number of Answers that received the Highest Up-Votes, but were not accepted: There are times when an answer given by a user may not be marked as the accepted answer but it still receives the maximum number of up-votes for that particular question. Such instances showcase the ability of that user to explain an answer in a way that appeals to most of the users and hence should be considered as a quality of an expert. [2]

Number of Badges Earned: Users are awarded with different badges based on their activeness on the forum and the usefulness of their contribution[3]. Hence, the number of badges earned by a user serves as a good indicator for determining his expertise.

[1] http://stackoverflow.com/

[2] http://stackoverflow.com/help/whats-reputation

[3] http://stackoverflow.com/help/badges

II. Research Questions

In this paper, we have tried to answer the below question:

Above what score in a particular technology can a user be considered to be an expert?

To answer this question, we need to first determine answers to the below mentioned sub-questions for a specific technology.

Q1 - Which users have the highest number of accepted answers?

Q2 - Which users have the most number of highest up-voted answers?

Q3 - Which users have the most number of highest up-voted, but not accepted answers?

Q4 - Which users have earned highest number of badges?

Q5 - Who are the most active users and which users among them, have the fastest average response time?

III. Approach

In this project, the activity of a set of users who have answered questions related to a particular technology was collected from the SO data dump[4] and the Stack Exchange Data Explorer[5]. Users were assumed to be experts in a particular technology, if they have earned a score above a certain threshold in that technology. This data set was also used to determine the top performers in that technology in terms of the number of answers which got accepted, the number of answers which received the highest points, the number of answers which received the highest points but were not marked as accepted, the number of badges earned in that technology and the average response times to the questions posted. The overlap between the expert user set and the sets of top performers obtained from the above analysis was used to prove that assumption made about the user being an expert in that field was correct. Since, our assumption proved to be true, a user can now determine which other users on SO are experts in a particular technology.

For the purpose of this project, we performed empirical analysis on two different technologies. The first one is **Java**, which is the most popular technology on SO and second one is **XSS**, which is among the less popular technologies, in terms of number of participating users and post count. The reason for selecting these two technologies was to observe the difference in the percentage of users who can be termed as experts in both a popular and a less popular technology.

IV. Data Collection and Processing

A. Obtaining the StackOverflow Data

Step 1: We started the project by exploring the SO schema using StackExchange Data Explorer.[5]

[4]http://meta.stackexchange.com/questions/198915/

[5]http://data.stackexchange.com/stackoverflow/query/new

Step 2: We downloaded the September 2013 data dump of SO from this[4] link.

Step 3: We imported data only from the **Posts** table of this dump, since our investigation only involved entries from the **Posts** table. In order to import the data from the **Posts** table into our local MySQL database we followed two steps:-

- We created the **Posts** table by referring to the schema given on the Stack Exchange Data Explorer[5] website.
- We then split the **Posts.xml** file present in the SO data dump into multiple XML's due to the large size of file and imported each XML individually in our local database.

Step 4: In order to make the computation less expensive, we extracted data related to the XSS technology and stored it in separate tables.

Step 5: The data for performing analysis on the Java technology was obtained by running queries directly on the Stack Exchange Data Explorer.[5]

B. Data Exploration

Step 1: We calculated the **Net Score** of all the users for a particular technology i.e. the net number of votes the users have received for all the answers that they have given for questions which belong to that particular technology.

Step 2: We calculated the **Average Response Time** of only the most active users of a particular technology. This is because, there might be certain users who have responded quickly but just to a few questions. Including the data about such users would give us misleading results. Hence, we filtered out the data related to such users.

Step 3: We calculated the number of answers given by each user which were **accepted** as the correct answers for questions belonging to the technology in consideration.

Step 4: We calculated the number of answers given by each user which received the **highest number of net votes** for a question belonging to the technology in consideration.

Step 5: We calculated the **number of answers given by users which were not accepted, but still received the maximum number of up-votes** for questions belonging to a specific technology.

Step 6: We then proceeded to calculate the **number of badges** that a user has earned by answering questions that belong to a particular technology. We chose to calculate the total number of **Enlightened, Nice Answer, Good Answer**

and Great Answer badges that the user has earned.[3]

- We first calculated the number of **Nice Answer** badges that a particular user has earned for a specific technology. A user obtains a Nice Answer badge, if he has obtained a net score of **10** or more for a particular answer that he has provided.[3]
- We then calculated the number of **Good Answer** badges that a particular user has earned for a specific technology. A user obtains a Good Answer badge, if he has obtained a net score of **25** or more for a particular answer that he has provided.[3]
- We then calculated the number of **Great Answer** badges that a particular user has earned for a specific technology. A user obtains a Great Answer badge, if he has obtained a net score of **100** or more for a particular answer that he has provided.[3]
- We then calculated the number of **Enlightened** badges that a particular user has earned for a specific technology. A user obtains an Enlightened badge, if the following conditions are satisfied:
  1) If he is the first person to answer that question.
  2) If his answer has received a score of more than 10.
  3) If his answer has been accepted by the inquirer.
  4) If his answer has not been accepted by himself.[3]
- We then created a table that combines the data obtained from the above analysis.
- We then calculated the total number of badges that each user has earned in a particular technology from the table created above.

V. Analysis of the Result Data Set

A. Data Analysis for the Java Tag

Step 1: Determining the Expert User Set

- We determined the set of expert users based on the scores that they have obtained by answering questions associated with the Java tag. We obtained a total of **185,762** users, who have achieved a net score of greater than or equal to zero. We found out that, on an average, a user has attained a score of **16** in this technology with a standard deviation of **290.70**.

- We then shortlisted the users who have obtained a score of **greater than or equal to 307** (which is the score value that is *one standard deviation above the mean*). The number of such users was found to be **1306**. These users are the top **0.7%** scorers for the Java tag. Therefore, we have assumed this set to be the Expert user set on SO for the Java technology.

Step 2: Determining the users having the most number of accepted answers

- We obtained a total of **74,491** users who have given at least 1 answer that got accepted. We found out that, on an average these users have given **5** answers which got accepted with standard deviation of **45.81**.

- We then shortlisted the users who have given **51 or more** accepted answers to questions asked about Java(which is the score value that *is one standard deviation above the mean*). The number of such users was found to be **999**. These users constituted the top **1.34%** of the result set.

- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **779**, which is **77.97%** of the total result set.

Step 3: Determining the users having the highest up-voted answers

- We obtained a total of **124,326** users who have given at least 1 answer that got the highest number of up-votes for a particular question. We found out that, on an average these users have given **6** answers which got the highest number of up-votes for a particular question with standard deviation of **58.73**.

- We then shortlisted the users who have given **124 or more** answers which received highest number of up-votes for questions asked about Java(which is the score value that is *two standard deviations above the mean*). The number of such users was found to be **688**. These users constituted the top **0.55%** of the result set.

- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **577**, which is **83.86%** of the total result set.

Step 4: Determining the users having the highest up-voted answers which were not accepted.

- We obtained a total of **95,230** users, who have given at least one answer that has the highest number of up-votes but is not marked as accepted. We found out that, on an average these users have given **4** answers which got the highest number of up-votes for a particular question but were not marked as accepted with standard deviation of **30.57**.

Table I
SUMMARY OF DATA ANALYSIS FOR THE JAVA TAG

	Total no. of users	No. of users shortlisted as the **Experts**	% of no. of users that are the **Experts**	No. of users overlapping with the **Expert** user set	% of overlap with the **Expert** user set
Number of Accepted Answers	74,491	999	1.34%	779	77.97%
Number of Highest Up-Voted Answers	124,326	688	0.55%	577	83.86%
Number of Highest Up-Voted, Unaccepted Answers	95,230	653	0.68%	517	79.17%
Number of Badges Earned	14,142	87	0.61%	87	100%
Average Response Time	191,774	840	0.43%	361	42.97%

- We then shortlisted the users who have given **66 or more** such answers to questions asked about Java(which is the score value that is *two standard deviation above the mean*). The number of such users was found to be **653**. These users constituted the top **0.68%** of the result set.

- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **517**, which is **79.17%** of the total result set.

Step 5: Determining users having the most number of badges

- We obtained a total of **14,142** users who have obtained 1 or more badges in Java. We found out that, on an average, these users have earned **4.78 (Approximately 5)** badges in this technology with a standard deviation of **32.12**.

- We then shortlisted the users who have obtained **70 or more** badges in Java(which is the *score value that is two standard deviations above the mean*). The number of such users was found to be **87**. These users are the top **0.61%** badge earners for the Java tag.

- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **87**, which is **100%** of the result set.

Step 6: Determining the users having the fastest average response time.

- We obtained a total of **191,774** users who have given at least 1 answer for a questions related to the Java tag. We found out that, on an average, a user has given **7** answers for the questions asked about this technology with a standard deviation of **69.74**.

- We then shortlisted the users who have given **77 or more** answers (which is the value that is approximately *one standard deviation above the mean*). The number of such users was found to be **2482**.

- Among the users selected above, we determined those users who had an average response time of **1 day or less**. The number of such users was found to be **840**.

- We then determined how many users in our assumed experts list are present in the list of the users obtained from the query above. The number of overlapping users was found to be **361**, which is approximately **42.97%** of the result set. The reason for a less overlap is that, it is quite possible that an expert user may not always be able to give the correct answers quickly, especially in case of complex problems.

B. Data analysis for the XSS Tag

Step 1: Determining the Expert User Set

- We determined the set of expert users based on the scores that they have obtained by answering questions associated with XSS tag. We obtained a total of **2328** users who have answered questions related to the *XSS* tag and have achieved a net score of greater than or equal to zero. We found out that, on an average, a user has attained a score of **3.30 (Approximately 4)** in this technology with a standard deviation of **14.06**.

- We then shortlisted the users who have obtained a score of greater than or equal to **18** (which is the score value that is *one standard deviation above the mean*). The number of such users was found to be **65**. These users

are the top **2.8%** scorers for the XSS tag. Therefore, we have assumed this set to be the Expert user set on StackOverflow for the XSS technology.

Step 2: Determining the users having the most number of accepted answers

- We obtained a total of **750** users who have given at least 1 answer that got accepted. We found out that, on an average these users have given **1.36 (Approximately 2)** answers which got accepted with standard deviation of **1.81**.

- We then shortlisted the users who have given **4 or more** accepted answers to questions asked about XSS(which is the score value that is *one standard deviation above the mean*). The number of such users was found to be **25**. These users constituted the top **3.33%** of the result set.
- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **20**, which is **80%** of the total result set.

Step 3: Determining the users having the highest up-voted answers

- We obtained a total of **1261** users who have given at least 1 answer that got the highest number of up-votes for a particular question. We found out that, on an average these users have given **1.41(Approximately 2)** answers which got the highest number of up-votes for a particular question with standard deviation of **2.56**.

- We then shortlisted the users who have given **6 or more** answers which received highest number of up-votes for questions asked about XSS(which is the score value that is approximately *two standard deviations above the mean*). The number of such users was found to be **20**. These users constituted the top **1.58%** of the result set.

- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **19**, which is **95%** of the total result set.

Step 4: Determining the users having the highest up-voted answers which were not accepted

- We obtained a total of **712** users who have given at least 1 answer that got the highest number of up-votes for a particular question but was not marked as accepted. We found out that, on an average these users have given **1.26 (Approximately 2)** answers which got the highest number of up-votes for a particular question but were not marked as accepted with standard deviation of **1.74**.

- We then shortlisted the users who have given **4 or more** such answers to questions asked about XSS(which is the score value that is *one standard deviation above the mean*). The number of such users was found to be **14**. These users constituted the top **1.97%** of the result set.

- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **10**, which is **71.43%** of the total result set.

Step 5: Determining users having the most number of badges

- We obtained a total of **99** users who have obtained 1 or more badges in XSS. We found out that, on an average, these users have earned **1.81 (Approximately 2)** badges in this technology with a standard deviation of **1.35**.

- We then shortlisted the users who have obtained **2 or more** badges in XSS(which is the *score value that is above the mean*). The number of such users was found to be **51**. These users are the top **51.51%** badge earners for the XSS tag.

- We then determined how many users in our assumed experts list are present in the list of the users obtained above. The number of overlapping users was found to be **41**, which is **80.39%** of the result set.

Step 6: Determining the users having the fastest average response time

- We obtained a total of **2399** users who have given at least 1 answer for questions related to the XSS tag. We found out that, on an average, a user has given **1.43 (Approximately 2)** answers for the questions asked about this technology with a standard deviation of **2.74**.

- We then shortlisted the users who have given **4 or more** answers (which is the value that is approximately *one standard deviation above the mean*). The number of such users was found to be **75**.

- Among the users selected above, we determined those users who had an average response time of **1 day or less**. The number of such users was found to be **50**.

Table II
SUMMARY OF DATA ANALYSIS FOR THE XSS TAG

	Total no. of users	No. of users shortlisted as the **Experts**	% of no. of users that are the **Experts**	No. of users overlapping with the **Expert** user set	% of overlap with the **Expert** user set
Number of Accepted Answers	750	25	3.33%	20	80%
Number of Highest Up-Voted Answers	1261	20	1.58%	19	95%
Number of Highest Up-Voted, Unaccepted Answers	712	14	1.97%	10	71.43%
Number of Badges Earned	99	51	51.51%	41	80.39%
Average Response Time	2399	50	2.08%	18	36%

- We then determined how many users in our assumed experts list are present in the list of the users obtained from the query above. The number of overlapping users was found to be **18**, which is approximately **36%** of the result set. The reason for a less overlap is that, it is quite possible that an expert user may not always be able to give the correct answers quickly, especially in case of complex problems.

VI. CONCLUSION

It can be observed from the results that the assumed set of expert users have shown better performance than average users for both the XSS and Java tags. The results indicate that the assumed set of expert users have provided more number of accepted and up-voted answers and have earned more badges than average users, clearly indicating their proficiency in these technologies. Also we identified the instances where the answers which were not accepted, but still received highest up-votes. These set of users also matched with our assumed set of experts. The only parameter which is not a clear indicator of expertise is Response Time for which we observed that the assumed experts may not always be the quickest respondents.

Thus, we can conclude that users whom we assumed to be experts can actually be considered as experts and the inquirers can almost safely assume their answers to be true.

REFERENCES

[1] L. Mamykina, B. Manoim, M. Mittal, G. Hripcsak, and B. Hartmann, "Design lessons from the fastest q&a site in the west," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '11. New York, NY, USA: ACM, 2011, pp. 2857–2866. [Online]. Available: http://doi.acm.org/10.1145/1978942.1979366

[2] A. Bosu, C. S. Corley, D. Heaton, D. Chatterji, J. C. Carver, and N. A. Kraft, "Building reputation in stackoverflow: An empirical investigation," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 89–92. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487085.2487107

[3] V. S. Sinha, S. Mani, and M. Gupta, "Exploring activeness of users in qa forums," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 77–80. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487085.2487104

[4] S. Grant and B. Betts, "Encouraging user behaviour with achievements: An empirical study," in *Proceedings of the 10th Working Conference on Mining Software Repositories*, ser. MSR '13. Piscataway, NJ, USA: IEEE Press, 2013, pp. 65–68. [Online]. Available: http://dl.acm.org/citation.cfm?id=2487085.2487101