StackOverflow Data Analytics

Ashwin Bhide
Georgia Institute of Technology
Atlanta, Georgia
ashwin.bhide@gatech.edu

Chaitanya Bapat
Georgia Institute of Technology
Atlanta, Georgia
chai.bapat@gatech.edu

Nidhi Menon
Georgia Institute of Technology
Atlanta, Georgia
nmenon34@gatech.edu

Sneha Venkatachalam
Georgia Institute of Technology
Atlanta, Georgia
sneha30@gatech.edu

Vaibhav Tendulkar
Georgia Institute of Technology
Atlanta, Georgia
vtendulkar@gatech.edu

1 INTRODUCTION

StackOverflow, a flagship site of StackExchange network is an online question-answer community providing rapid access to knowledge and expert peers. Research on StackOverflow shows certain shortcomings in terms of user engagement and question answering.

There is an increasing trend in number of unanswered questions on StackOverflow, and percent distribution of answered questions is skewed according to question domain and other factors. Currently, StackOverflow displays unanswered questions by tags instead of routing questions to specific users. The limitation of this approach is that there is no intelligent algorithm that makes use of extracted features to identify experts and route the pertinent questions to them.

2 PROBLEM DEFINITION

(1) To analyze StackOverflow to find trends in data, and understand community user engagement and dynamics
(2) To design a question routing and expert-recommendation system that will enable experts to provide satisfactory answers to questioners

3 SURVEY

StackOverflow's growing prominence is due to the active involvement of software developers in content moderation[1] and presence of game mechanics[6] to make the forum competitive. Bogdan et al[2] tried to explore an interplay between StackOverflow and software development reflected by code changes in Github. Octay et al[3] suggest using Quasi-experimental Designs to study human behaviour on Stack Overflow for analyzing long-lasting value of questions. Laura et al.[4] establish a positive correlation between user reputation scores and their contribution to a diverse tags while Bogdan et al [5] assessed the representation and social impact of gender in Stack Overflow. Gharibi et al. [6] talk about a system that recommends unachieved badges to users based on their behavior. In [7], the authors used unsupervised learning to categorize mined Stack Overflow questions and defined a ranking algorithm to help understand developer issues. Asaduzzaman et al. [15] analyze StackOverflow data to see why unanswered questions remain unanswered.

Morakot et al. [8] propose a feature based prediction approach to predict answerers for a given question using Random Forests classification technique. Thiago et al in [9] study the behavior of the experts in the online communities using TF-IDF measure to rank topics of expertise and introduces a recommendation approach based on trending topics. Zhou et al. in [11] consider the question routing problem as a classification task, deriving a variety of features about the question, user and the relationship between them. Guo et al. in [12] suggest a generative model to discover latent topics and interests to recommend answer providers. Ashton et al in [16] talk about relationships and temporal characteristics to calculate long-lasting value of answers and identifying questions without satisfactory answers. In [10] Chang et al. propose a linear-regression based recommendation model that finds domain experts. Jun et al. in [13] analyzed the network representing the asker-helper interactions in an online community and concluded that it produced a different bow-tie structure than what is associated with the graph of World Wide Web. Pedro et al. [14] come up with a rankSLDA algorithm which combines supervised ranking for recommending questions based on their domain-based ranking.

4 PROPOSED METHOD

4.1 Intuition

We propose an approach to extract features to identify experts in particular domains and route relevant questions to them. To understand the user engagement and dynamics of the community based question answering platforms, we plan to use graph visualization softwares. This approach will be successful as routing the question to experts will help the questioner get the correct answer quickly.

4.2 Description

Data extraction and cleaning:

We used the raw data from StackExchange Data Archive[17]. Since the available data was huge, we wrote SQL queries to extract required datafields like tags, view-counts, bounties, etc. in CSV format. The cleaning process involved filling in the missing values using OpenRefine, removing HTML tags using Python to ensure that cleaned text was passed on to the analysis phase. Fields such as the number of lines of code and the readability score of the post were calculated by using nltk library in Python. Higher readability score indicated greater likelihood of users understanding and answering the question. Using the readability library, we calculated text readability using algorithms like Flesch-Kincaid, Coleman-Liau, Dale-Chall, SMOG, Automated Readability Index, Flesch Reading Ease and averaged outputs to find the final readability score of the post.

Question Recommendation:

A lot of users ask questions on StackOverflow every day, in all these questions, the answerers may get lost in the haystack of questions while looking for questions they would likely know the answer to, and would want to answer. To help them navigate through this haystack, we propose a solution in which we recommend questions to users based on their past activity of what questions they have given answers to. Rather than have them navigate through the haystack looking for the proverbial needle, we try to present them with the needle without them having to put in too much effort. We are currently using K-means to cluster questions based on the tags they have been given. K in our model is the unique number of tags in our dataset. We keep track of the number of answers a user has given in a particular domain and the average answer score per domain. When a new question comes in, we query to see what cluster the question belongs to and then check the top 5 answerers for that cluster and recommend the question to them. For the above approach we have used user features such as user id, question id, tags, answer score.

Having implemented one algorithm, our next step involves using ensemble learning techniques to improve the results. We will then compare the results of the algorithm and pick the one that performs the best.

We plan to use Random Forest classification algorithm to make a prediction based on a set of features of questions and users. The feature vector for questions will include tags, length, readability score and number of code fragments. The feature vector for the users will include reputation, length of membership, age and number of upvotes he / she has received votes and number of accepted answers.

Our approach will be divided into two phases :

- Learning phase : In this phase, we will train the random forest classifier by feeding it a set of known question-answerer pairs. This will be derived from historical data.
- Testing phase : We will set aside a fraction of our data which will not be used for training the algorithm. Once the learning phase is complete, we will pair each question in the test set with an answerer and ask the model to predict the chances of the answerer answering the question. We will pick the top n (this number might change) users predicted by our algorithm who are most likely to answer our question and compare these with the users who actually answered the question. This will help us determine the accuracy of our algorithm.

Data analysis and visualization:

We divided the data analysis process into two parts. In the first part, we analyzed user engagement on StackOverflow by monitoring the number of questions asked, and the number of upvotes and downvotes they received. A similar study was conducted on the answers posted to these questions. In the second part, we focussed on the major issue faced by StackOverflow users- unanswered questions. The data we scraped indicated an increasing trend in the number of unanswered questions over the past few years which defeats the core purpose of the question-answer community. Another study revealed the increase in the number of bounties that users place on their questions in hopes of receiving immediate answers. These analyses laid the foundation of our project, giving us the objective of building a system that could help improve StackOverflow. All visualizations included in this report were generated using Tableau.

5 INNOVATION

(1) Inclusion of Readability score : Intuition behind including readability score was that a higher readability score implies easier understanding of question. It in turn enables more users to answer the question with greater accuracy.
(2) Intelligent question routing mechanism driven by question tags & user skill match.
(3) Expert-recommendation system that enables experts to provide satisfactory answers to questioners while ensuring valuable user contribution.

6 EXPERIMENTS AND EVALUATION

6.1 Description of testbed :

(1) Via Visualizations
  (a) What is the trend in unanswered questions?
  (b) How increasingly important is it becoming for users to place bounties on questions?
  (c) What is the trend in user engagement in terms of question-answers and upvotes-downvotes?
(2) Via Machine Learning Model
  (a) Which user is most likely to answer a given new question correctly?
  (b) Which users are the top answerers for a given domain?

6.2 Observation & Analysis

Model driven analysis using k-means clustering accomplished grouping of questions based on the tags and the score. We evaluated the model by testing it with outsample data to confirm that the tags retrieved from the given unknown question helped to find top users related to the particular tag. Here for example, we considered top 100 tags and were able to train the model with 250,000 questions based on the features - question id, tag id, answerer id and answer score.

User engagement analysis performed on data revealed that despite receiving a huge number of views, questions on StackOverflow remain unanswered for long periods of time (see Fig. 1). Fig. 2 shows an increasing trend, both in creation of questions and in closing of unanswered questions by users, due to which, in recent years, more users are forced to place huge bounties on questions requiring immediate answerers, as depicted in Fig. 3 and Fig. 4. These results drew our attention to the need for expert-recommendation and question-routing systems. Drilling down to a period of one week, we also visually analysed user activity in terms of upvotes and downvotes for questions and answers on StackOverflow in Fig. 5 and Fig. 6.

7 PLAN OF ACTIVITIES

(1) Ongoing:
  - Data Cleaning: Vaibhav Tendulkar

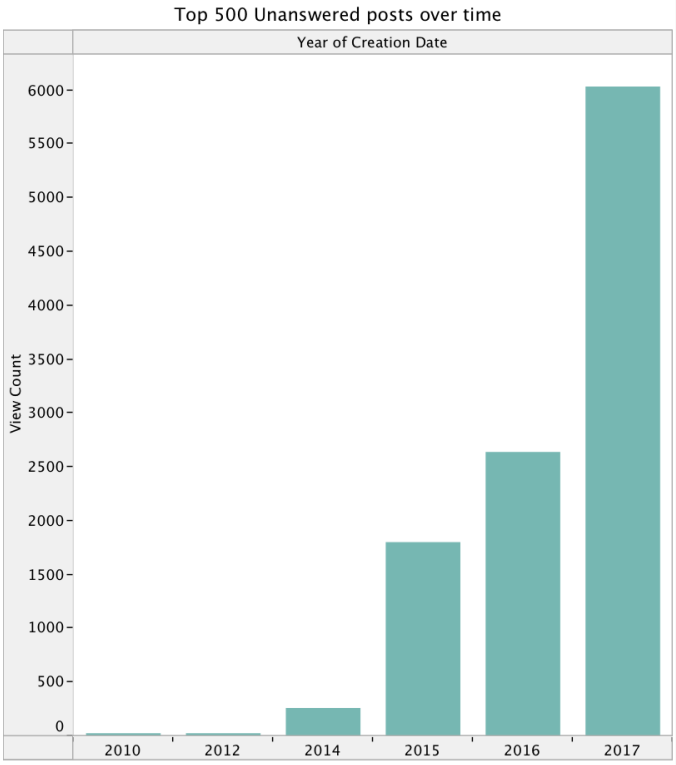

Figure 1: Top 500 Unanswered posts over time

- Data Extraction & Querying: Sneha Venkatachalam & Nidhi Menon
- Visualization and Analysis: Sneha Venkatachalam & Nidhi Menon
- Model Development: Chaitanya Bapat & Ashwin Bhide

(2) By 1st December:

- Tuning features: Ashwin Bhide, Sneha Venkatachalam
- Analyzing user engagement on StackOverflow: Chaitanya Bapat
- Proposing an expert recommendation system: Vaibhav Tendulkar, Nidhi Menon

All team members have contributed similar amount of effort.

8 REFERENCES

[1] Lena Mamykina, Bella Manoim, Manas Mittal, George Hripcsak, BjoÌĹrn Hartmann, Design Lessons from Fastest QA Site in West

[2] Bogdan Vasilescu,Vladimir Filkov, Alexander Serebrenik, StackOverflow and GitHub: Associations Between Software Development and Crowdsourced Knowledge

[3] H. Oktay, B. J. Taylor, D. D. Jensen, Causal Discovery in Social Media Using Quasi-Experimental Designs

[4] Laura MacLeod, Reputation on Stack Exchange: Tag, You are It!

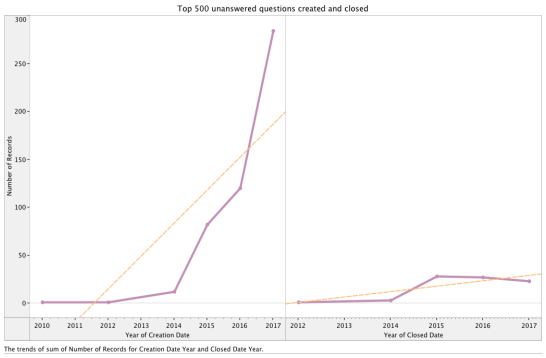

Figure 2: Top 500 unanswered questions created and closed

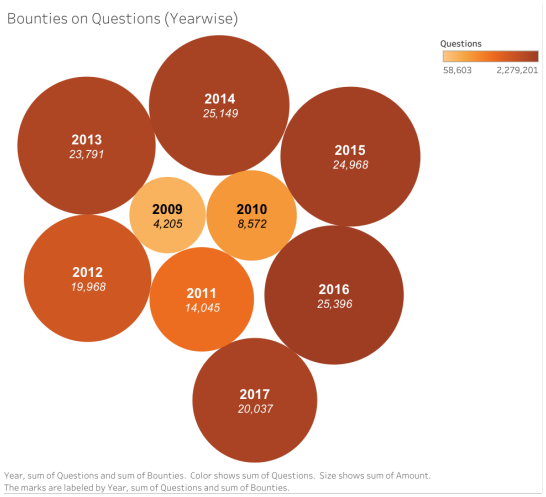

Figure 3: Bounties on Questions(Yearwise)

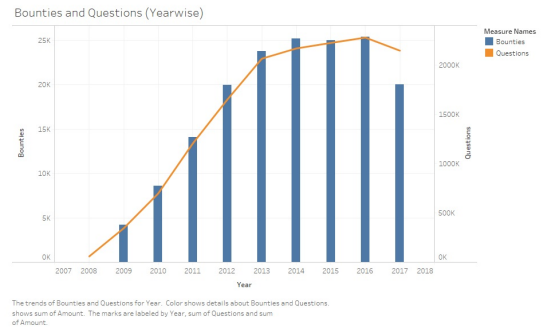

Figure 4: Bounties and Questions(Yearwise)

[5] Bogdan Vasilescu, Andrea Capiluppi, Alexander Serebrenik - Gender, Representation and Online Participation: A Quantitative Study of StackOverflow

[6] Gamified Incentives: A Badge Recommendation Model to Improve User Engagement in Social Networking Websites

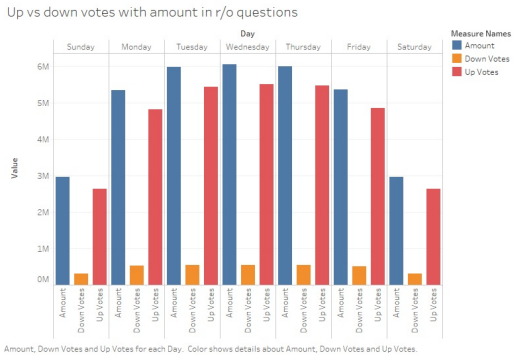

Figure 5: Ups vs down votes with amount inr/o questions

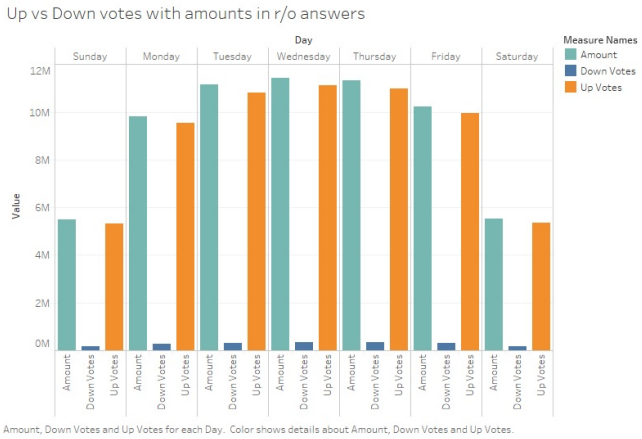

Figure 6: Ups vs down votes with amount in r/o answers

[7] K. Bajaj, K. Pattabiraman, A. Mesbah, Mining Questions Asked by Web Developers,

[8] Morakot Choetkiertikul, Daniel Avery, Hoa Khanh Dam, Truyen Tran and Aditya Ghose , Who will Answer my Question on Stack Overflow?

[9] Thiago B. Procaci, Bernardo Pereira Nunes, Terhi Nurmikko-Fuller, Sean W. M. Siqueira, Finding Topical Experts in Question & Answer Communities

[10] S. Chang and A. Pal, Routing Questions for Collaborative Answering in Community Question Answering, IEEE/ACM International Conference, 2013

[11] T. C. Zhou, M. R. Lyu, I. King, A Classification-based Approach to Question Routing in Community Question Answering,

[12] J. Guo, S. Xu, S. Bao, and Y. Yu, Tapping on the Potential of Q&A Community by Recommending Answer Providers,

[13] Jun Zhang, Mark S. Ackerman, Lada Adamic, Expertise Networks in Online Communities: Structure and Algorithms

[14] Jose San Pedro, Alexandros Karatzoglou, Question Recommendation for Collaborative Question Answering Systems with RankSLDA

[15] Muhammad Asaduzzaman, Ahmed Shah Mashiyat, Chanchal K. Roy, Kevin A. Schneider, Answering Questions about Unanswered Questions of Stack Overflow

[16] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, Jure Leskovec, Discovering Value from Community Activity on Focused Question Answering Sites: A Case Study of Stack Overflow

[17] StackExchange, 2017 Retrieved from https://data.stackexchange.com/