StackExchange 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 PROBLEM DEFINITION

1. Given a question, find expert users to answer
2. User engagement driven ML solution
a. Visualizing StackExchange as a Graph
b. Applying ML algorithms

2 HEILMEIER QUESTIONS

(1) What are you trying to do? Articulate your objectives using absolutely no jargon.
1. Given a question, find expert users to answer
2. Understand user engagement on Stack Overflow

(2) How is it done today; what are the limits of current practice?
Currently, Stack Overflow does not route questions to specific users. The user selects relevant tags which displays the latest unanswered questions in the concerned domain.
The limitation of the current practice is that there is no intelligent algorithm that makes use of extracted features to identify experts and route the pertinent questions to them

(3) What's new in your approach? Why will it be successful?
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) Who cares?
1. Stackexchange users
2. Development community

(5) If you're successful, what difference and impact will it make, and how do you measure them (e.g., via user studies, experiments, groundtruth data, etc.)?
Measure:
Ratio of correctly predicted experts to the ground truth (users with most votes)

Impact:
a. Reduce wait-time for an answer
b. Prevent question starvation
c. Maximize use of resources (experts on StackOverflow)

(6) What are the risks and payoffs?
Risks: Accuracy of answers Time to get an answer to a question
Payoffs: Building a better knowledge base with accurate answers Raise overall participation rate with more user satisfaction Reduce time lag between asking and answering a question

(7) How much will it cost?
No cost since we use open-source data set , software applications and surveys

(8) How long will it take?
45 days (20+25)

(9) What are the midterm and final exams to check for success? How will progress be measured?
Midterm - Data Cleaning, Feature Extraction, Preliminary data analysis + Visualization
Final - Fine-tuning data analysis + Visualizing

3 INNOVATION

A research on StackOverflow exposed certain shortcomings in terms of user engagement and question answering. We propose a question-routing and expert-recommendation system that will enable experts to provide satisfactory answers to questioners while ensuring valuable user contribution.

4 PLAN OF ACTIVITIES

Groundwork (done so far): Literature Survey :
Understanding StackOverflow Design - Chaitanya Bapat
Studying Expert Recommendation - Vaibhav Tendulkar, Nidhi Menon
Question Routing - Sneha Venkat

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

User Engagement - Ashwin Bhide

We intend to finish the following activities:
a)By 10th November:
Data Cleaning: Chaitanya Bapat
Feature Extraction: Sneha Venkat
Preliminary data analysis model: Vaibhav Tendulkar, Nidhi Menon
Visualization : Ashwin Bhide

b)By 1st December:
Fine-tuning features: Ashwin Bhide, Sneha Venkat
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

5 LITERATURE SURVEY

StackOverflow a flagship site of the StackExchange network is an online question and answer community providing rapid access to knowledge and expert peers. To understand the reason StackExchange is the most-used website for QA, we referred to the paper[1] by Lena et al. The reason behind StackOverflows prominence is the active involvement of software developers in content moderation and presence of game mechanics[6] to make the forum competitive. Moreover, the use of StackOverflow has close relation with the the time software developers spend on actually answering questions versus coding. Bogdan et al[2] tried to explore such an interplay between StackOverflow and software development reflected by code changes in Github. The results highlighted that the more active Github committers provided more answers while asking fewer questions. Similarly, Octay et al [3] suggest using Quasi-experimental Designs to analyse human behaviour on Stack Overflow and determine the cause-and-effect relationships, involving approaches which we could incorporate for analyzing long-lasting value of questions.

User engagement plays a major role in the success of online communities. 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, which can be used to target sections of the community for increased user participation. Gharibi et al. [6] talk about a system that recommends unachieved badges to users based on their behavior, to increase user engagement by calculating the correlation between unachieved badges and users previously awarded badges. In [7], the authors used unsupervised learning to categorize mined Stack Overflow questions and also defined a ranking algorithm to rank questions to understand issues faced by web developers. Asaduzzaman et al. [15] analyze stackoverflow data to see why unanswered questions remain unanswered. These methods could be incorporated into our project in order to extract more valuable information helping the software community to understand and address such issues.

Another aspect contributing to the growth of StackOverflow is its expert finding mechanism. Morakot et al. [8] propose a feature based prediction approach to predict who will answer a given question using Random Forests classification technique, suggesting a social-network based approach to exploit the relational attributes of the community. 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. These can be combined with the linear-regression based recommendation model in [10] proposed by Chang et al. that finds domain experts and detects low quality answers to be routed to groups of users rather than a single expert, for improved performance. 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. Whereas, Pedro et al. [14] come up with a rankSLDA algorithm which combines supervised ranking with topic modeling for recommending questions to users based on their domain-based ranking.

6 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!

[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

[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