Learning from Gurus: Analysis and Modeling of Reopened Questions on Stack Overflow

Rishabh Gupta
International Institute of Information Technology
Hyderabad
rishabh.gupta@research.iiit.ac.in

P. Krishna Reddy
International Institute of Information Technology
Hyderabad
pkreddy@iiit.ac.in

ABSTRACT

Community-driven Question Answering (Q&A) platforms are gaining popularity now-a-days and the number of posts on such platforms are increasing tremendously. Thus, the challenge to keep these platforms noise-free is attracting the interest of research community. Stack Overflow is one such popular computer programming related Q&A platform. The established users on Stack Overflow have learnt the acceptable format and scope of questions in due course. Even if their questions get closed, they are aware of the required edits, therefore the chances of their questions being reopened increases. On the other hand, non-established users have not adapted to the Stack Overflow system and find difficulty in editing their closed questions. In this work, we aim to identify features which help differentiate editing approaches of established and non-established users, and motivate the need of recommendation model. Such a recommendation model can assist every user to edit their closed questions leveraging the edit-style of the established users of the platform.

CCS Concepts

•**Information systems** → **Reputation systems;** *Social recommendation;* •**Human-centered computing** → Empirical studies in collaborative and social computing;

Keywords

stackoverflow, reopened questions, closed questions, edits

1. INTRODUCTION

Community-driven Question and Answering (Q&A) platforms provide an excellent means to the internet users with specific information needs, to get responses from the experts. These platforms also serve a vital role in knowledge base creation. Stack Exchange is one such popular network of community driven Q&A platforms which has 148 domain specific platforms. This network has about 300 Million (M) unique global visitors who have looked upon 3.1 M questions which in-turn have 4.5 M submitted answers and 17 M comments[1]. Stack Overflow is one of the most popular Q&A platforms of the Stack Exchange group with nearly 4.5 M registered users, and contains computer programming related questions (9.9 M) and answers (16.5 M) as per August, 2015 data dump [2].

Stack Overflow lays specific guidelines on the scope and quality of the content on the platform and in order to keep itself focused it provides a voting and reputation mechanism to its users . The users can follow these guidelines and vote (*Upvote*, *Downvote*, *Closevote*, *Reopenvote* and *Deletevote*) questions and answers based on the usefulness and appropriateness. The users earn reputation by actively participating on the platform. The primary way to gain reputation is by posting good questions and useful answers. Voting on such posts may help in gaining reputation as well.

The reputation of a user provides a rough measure of how much the community trusts the user. A user with atleast 1,000 reputation points has been participating for a fair amount of time, and is referred as *Established User* by the platform [1]. In this study, we aim to study the problem of edit suggestion (for closed questions) under the categorization of established and non-established users (users with reputation less than 1000).

A question is the seed level element of Stack Overflow, therefore it becomes important to pay special attention to its quality. Stack Overflow through its guidelines strictly discourages unfit questions that do not add value to the platform, such as off-topic, duplicate, and opinion based questions. The users of platform vote to close such questions. There are 4.5% questions that are closed on Stack Overflow that counts to 448,508 questions. Individually, the established and non-established users of the platform have approximately 4% of their questions that have been closed via voting.

A *Closed* question can be *Reopened* if it is properly edited. On the other hand, if there is no scope of improvement of a *Closed* question, it is *Deleted* from the platform. We find that around 65% closed questions for non-established users and 54% closed questions by established users have been edited (atleast once) before getting reopened. This clearly indicates that a large proportion of closed questions are edited before reopening takes place, therefore emphasizing the role of edits in reopening process and the need of an automated system to suggest edits for the closed questions of users (both established and non-established).

CODS '16, March 13-16, 2016, Pune, India
© 2016 ACM. ISBN 978-1-4503-4217-9/16/03…$15.00
DOI: http://dx.doi.org/10.1145/2888451.2888460

[1] http://stackexchange.com/about

Researchers have extensively looked into problems such as identifying unfit questions [4, 6], assessing thier quality [5], finding best answers for them [8], and suggesting expert users to answer them [7] , however, to the best of our knowledge, we are the first to address the problem of improving quality of unfit questions. We study the reopening process by categorizing the users amongst established and non-established users.

In this study, we propose the novel problem of edit suggestion for reopening of closed questions posted by the users on Stack Overflow platform. The idea is to suggest users with edit features that can help them to improve text of their closed questions. These edit features can be best suggested by learning from the edit style of skilled users on the platform. The non-established users, being poor at editing skills, have significantly less percentage of reopened questions (4.4%) as compared to established users (6.4%) of the platform. Since established users possess better edititng skills, we propose to suggest edits to users by learning discriminative edit features from edited questions of established users.

2. QUESTIONS AT STACK OVERFLOW

The users on Stack Overflow can ask questions and get answers to their questions from other users. Each question on Stack Overflow platform majorly contains three sections: *Title*, *Body*, and *Tags*[3]. The title is the first thing about the question that the potential answerers will see. It should sum up the entire question and be able to catch the attention of the related community. The body of the question should expand the summary provided by its title. Stack Overflow provides special formatting for inserting *code* in the body. The code can be used to reproduce the problem and thus help in better understanding of the question. The body can also have few *links* which can be used to create a live example of the problem. The tags are the words or phrases that highlight the main topics of the question.

3. PROPOSED APPROACH

The automated edit suggestion system aims to suggest the appropriate edits which may be beneficial in reopening of question. We, therefore, intend to learn the key edits that are helpful in reopening of the closed questions. The idea is to present a unified model that can be used by the users at large. We formulate the problem of edit suggestion as a binary classification problem. We create a collection of edited closed questions by extracting closed questions whose either body, title, or tags section was edited once. The questions among these which got reopened later are treated as positive set and rest as negative set. To identify the key edits, we define a state-space of features derived from body, title and tags sections of a question, such as number of popular tags, length of title, and number of urls in body. We build classifier on a set of established users, based on these features, to detect for all users (established or non-established) if the edited version of a question will reopen or not. It helps to identify top features among body, title and tag sections of the closed question. These top features represent the areas to focus during editing of the closed question.

4. EXPERIMENT & RESULT

The Stack Exchange provides all the user-contributed content on its various platforms under cc by-sa 3.0 license. We perform experiments with the Stack Overflow August, 2015 data dump [2] provided by Stack Exchange. Out of 4.5M registered users, we find that only 1.8% of the registered users are established users, whereas rest of them (i.e. 98.2%) are non-established users. The closed questions count to 445,000 of which 22,000 got reopened later.

We experiment with 37 features computed based on the edited version of the closed questions. These features are extracted from the three sections of the question: 25 from body, 5 from title, and 7 from tags. The features are either frequently used text features, such as number of characters and number of special characters, or stackoverflow specific features such as number of urls and number of code snippets. We use Adaboost with base classifier as Decsion tree as the classfication technique and obtain the F1 score of 0.6 for established users and 0.5 for non-established users. Based on results of established users, question words, length of question, popular tags are amongst the top features.

5. CONCLUSION & FUTURE WORK

In this study, we categorize the users on Stack Overflow platform as established and non-established users. We learn predictive model for reopening of edited closed questions based on 37 features using Adaboost as the classification technique. The classifier performs reasonably well for established users with an F1-score of 0.6 (better than non-established). We identify salient feature for reopening task based on the model learnt on established users and thus present a unified model for edit suggestions.

We further aim to perform an extensive analysis of various facets related to closed and reopened questions of the Established and Non-Established users on Stack Overflow platform. Using the insights from analysis, we aim to extend the feature set wth more sophisticated features based on semantics of question.

6. REFERENCES

[1] Established users. http://stackoverflow.com/help/privileges/established-user.

[2] Stack overflow data dump. https://archive.org/details/stackexchange.

[3] Stack overflow: How to ask? http://stackoverflow.com/help/how-to-ask.

[4] Correa, D., and Sureka, A. Fit or unfit: analysis and prediction of closed questions' on stack overflow. In *COSN* (2013), pp. 201–212.

[5] Correa, D., and Sureka, A. Chaff from the wheat: characterization and modeling of deleted questions on stack overflow. In *WWW* (2014), pp. 631–642.

[6] Ponzanelli, L., Mocci, A., Bacchelli, A., Lanza, M., and Fullerton, D. Improving low quality stack overflow post detection. In *ICSME* (2014), pp. 541–544.

[7] Riahi, F., Zolaktaf, Z., Shafiei, M., and Milios, E. Finding expert users in community question answering. In *WWW* (2012), pp. 791–798.

[8] Shah, C., and Pomerantz, J. Evaluating and predicting answer quality in community qa. In *SIGIR* (2010), pp. 411–418.