What Kind of Questions Do Developers Ask on Stack Overflow? A Comparison of Automated Approaches to Classify Posts Into Question Categories

Stefanie Beyer[1] Christian Macho[1] Massimiliano Di Penta[2] Martin Pinzger[1]

Abstract: Stack Overflow (SO) is among the most popular question and answers sites used by developers. Labeling posts with tags is one of the features to facilitate searching and browsing SO posts. However, existing tags mainly refer to technological aspects but not to the purpose of a question. In this paper, we argue that tagging posts with their purpose can facilitate developers to find the posts that provide an answer to their question. We first present a harmonization of existing taxonomies of question categories, that represent the purpose of a question, into seven categories. Next, we present two approaches to automate the classification of posts into the seven question categories, one using regular expressions and one using machine learning. Evaluating both approaches on an independent test set, we found that our regular expressions outperform machine learning. Applying the regular expressions on posts related to Android app development, showed that the categories API USAGE, CONCEPTUAL, and DISCREPANCY are most frequently assigned. By integrating our approach into SO, posts could be manually tagged with our categories which would allow developers to search posts by question category.

Keywords: Stack Overflow; Classification; Question Categories; Program Understanding

Many developers use question and answer forums, such as Stack Overflow (SO), to discuss and solve their development issues. To refine the search and describe the questions briefly, each question post on SO is labeled with 1 to 5 tags. These tags often describe technological aspects of the questions but lack to describe the motivation of the author which is necessary to understand the issue [Be17].

Given the number of questions that are newly posted each day, manually assigning such tags to the posts is considered no feasible. In this work, we set out to automate the process of labeling posts with tags that represent the *why* questions are asked (question categories). These tags are important to understand the most difficult aspects of software development and the usage of APIs [AS13].

We manually classified 1000 posts into seven question categories that we obtained by comparing taxonomies found by prior studies [AS13, Be17, BP14, RS15, TBS11]. Additionally, we marked 2.192 phrases that indicate a particular question category. Then, we used the manually created data set to supervise the automated classification of posts into question categories. First, we implemented a classifier based on regular expressions that we derived

[1] University of Klagenfurt, stefanie.beyer@aau.at
[2] University of Sannio, dipenta@unisannio.it

doi:10.18420/SE2021_03

by combining recurrent patterns in the phrases. Second, we trained machine learning (ML) models using Random Forest and Support Vector Machine on the phrases to classify the posts into the seven question categories.

We evaluated the performance of our approaches on an independent test set of 110 SO posts that were neither used to extract patterns for the regular expressions nor used to train and test the models before. The results showed that the regex approach achieves an average precision and recall of 0.90 and 0.90, respectively, which outperforms the ML approaches. Furthermore, the regex approach is much faster and easier to adapt. The application of the regex approach to all studied questions confirmed our findings that API USAGE, DISCREPANCY, and CONCEPTUAL are the most frequently occurring question categories. Furthermore, the results show that the majority of the posts is classified in one to three categories and that the categories are mostly not overlapping.

By integrating the regex classifier into SO, several improvements could be achieved. First, the automated nature of our approach can help to tag historical posts that still lack tags that describe non-technical aspects of the posts. Tagging existing posts automatically will increase the chances that historical posts will also be tagged with newly introduced tags. Second, we enable developers posting questions by applying our approach to their post draft and suggesting related question categories to improve the characterization of the posts through the assigned tags. Third, both of these applications consequently help to find appropriate posts when developers are searching for help on SO. Lastly, the question categories can be used to improve existing approaches, such as Seahawk and Prompter, that suggest suitable code snippets to provide more accurate postings.

Literaturverzeichnis

[AS13] Allamanis, M.; Sutton, C.: Why, when, and what: Analyzing Stack Overflow questions by topic, type, and code. In: Proceedings of the Working Conference on Mining Software Repositories. IEEE, S. 53–56, May 2013.

[Be17] Beyer, S.; Macho, C.; Di Penta, M.; Pinzger, M.: Analyzing the Relationships between Android API Classes and their References on Stack Overflow. Technical report, University of Klagenfurt, University of Sannio, 2017.

[BP14] Beyer, S.; Pinzger, M.: A manual categorization of android app development issues on Stack Overflow. In: Proceedings of the International Conference on Software Maintenance and Evolution. IEEE, S. 531–535, 2014.

[RS15] Rosen, C.; Shihab, E.: What are mobile developers asking about? A large scale study using stack overflow. Empirical Software Engineering, 21:1–32, 2015.

[TBS11] Treude, C.; Barzilay, O.; Storey, M. A.: How Do Programmers Ask and Answer Questions on the Web? (NIER Track). In: Proceedings of the International Conference on Software Engineering. ACM, S. 804–807, 2011.