Senior Design

Sociotechnical Solution to Large-Scale Formal Specification Mining

Group May1620

Alex Dana, Deeksha Juneja, Cody Hanika
Advisor: Professor Rajan

Table of Contents

Figures:

Background

Formal behavioral software specifications allow a programmer to specify the behavior of a piece of code to aid in verifying the correctness of the said code. These specifications are not required by the Java programming language and are currently used by very few programmers when writing code. A large factor in the small adoption of formal specifications when writing Java code is the lack of easily available formal specifications for major public libraries. Prior work towards this problem has taken a purely technical approach which has failed to be accurate at creating specifications for widely used code. This project aims to solve this problem. Having learned from the earlier mistakes, our team will be taking a sociotechnical approach that leverages the success of community-run websites, such as Stack Overflow, to solve this problem.

Project Statement

In order to make formal specifications widely available for Java programmers, a community driven web platform will be developed that will allow developers to collaborate to create formal specifications for widely used Java libraries. This project will leverage Stack Overflow, at least initially, as the collaboration platform for creating these specifications. The questions will be created with approximations for code specifications and answers will be created by the community to verify or fix these specifications.

A program is being created to parse, finalize, and persist these specifications created by community answers on Stack Overflow. These specifications will then be made available through Boa.

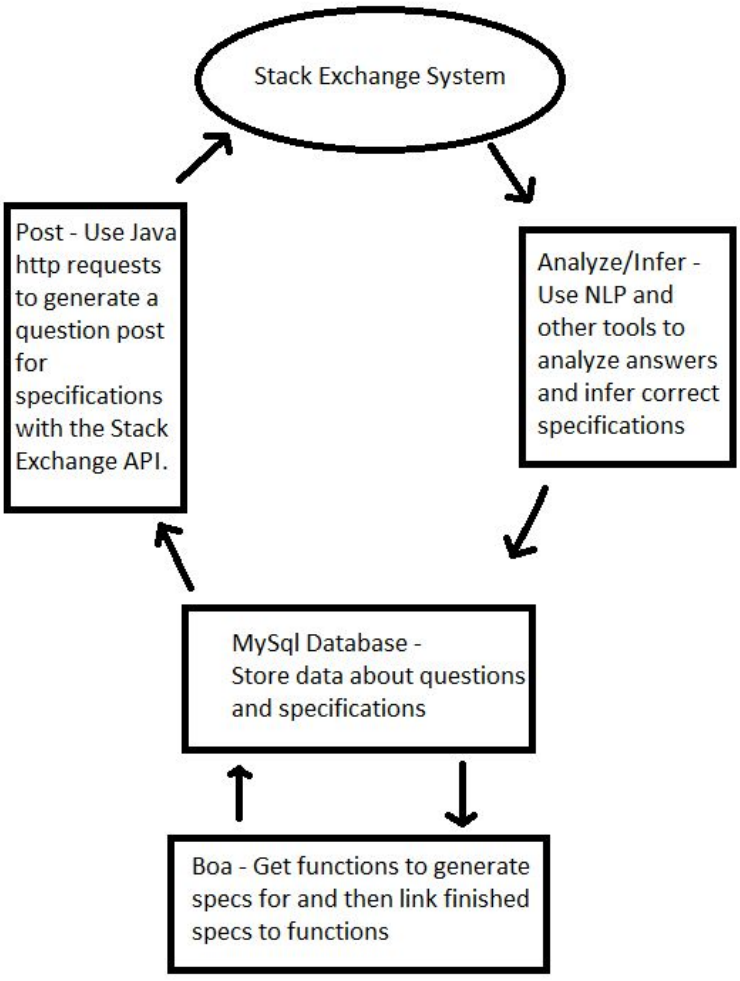
Requirements

- A program that utilizes an existing community Q & A platform to automatically post questions about formal specifications for widely used code (gathered from Boa) for the programming community to answer.
- A program to monitor answers from the Q & A platform that will make decisions on what is most likely to be the correct answer. This program must also be able to persist these specifications.
- Ability to link these specifications to Boa functions

Proposed Initial Solution

Our project must be able to generate specifications and then submit these specifications to the Stack Overflow site in the form of a question. The answers to this question will then be analyzed to refine our initial “guess” at the specification. We will then store this new specification in a database and link it in Boa to its function to be used and possible refined later. The process will follow the figure below:

Figure 1: Diagram of our proposed project structure



To test this we will check to see if we can have questions automatically generated and posted to the Stack Overflow site using the Stack Exchange API. If this succeeds we will check to see if our newly refined specification, acquired from the analysis of comments and answers, is valid. Finally we will make sure that this new specification does indeed persist to the database and Boa.

Professor Rajan had a fairly clear vision of where he wanted to take this project but there were still alternative solutions proposed. The main alternative was to use a different site than Stack

Overflow for the questioning process. However this did not make sense for us since the Stack Overflow site was setup so well for what we wanted. The main weakness to the site is the poor documentation on their API and the aversion of auto generated questions - meaning that we may need to have our program “pretend” to be a person posting the questions. Overall though the benefits of the Stack Overflow site outweigh the negatives so we decided to take this approach. The schema for our MySQL database also has many possible solutions. We have decided to try and conform it to the Boa structure (as much as possible). Alternatively we could have set it up in such a way that made it more simple but this would cause issues and more work when sending information back to Boa so we settled for the former option.

Assessment of the Initial Solution

The proposed solution leaves 3 major challenges to be faced. The first is community participation. The proposed solution relies on the fact that the community will participate in answering specification questions posted to Stack Overflow. To spark interest, we may need a way to promote the positives of formal specifications in our questions posted to Stack Overflow. Once people find value in specifications, there is a positive outlook on people answering questions on them, as Stack Overflow has been so successful in drawing the community in to answer questions. We are in the process of posting test questions to Stack Overflow in order to best determine the format we should use to achieve the best responses from the community.
The second challenge is getting the correct specifications for a piece of code. As specifications are not widely used, there are not many experts in formal specifications. This could make getting initial, correct specifications difficult to obtain. However, Stack Overflow does have many users and the discussion format of the website would allow people to work together and review others’ work.
The final challenge is inferring the specifications from community answers. Discussion on the specifications needs to be allowed and not restricted, but the actual answer specification needs to be able to be inferred from the discussion. We are currently looking into the Stanford Natural Language Processing as a tool to extract valuable pieces of text from the answers that we receive.

A major strength of the proposed solution is utilizing the existing infrastructure that Stack Overflow provides. This will allow our team to not have to deal with creating a platform for collaboration on these specifications. This gives us much more time to focus on the other challenges that we will face, as listed above.

System Description

We are using a Mavenized Java project with JDBC for a MySQL database seen in figure 1. The system is currently being planned for execution on the command line. User interaction should be minimal as our application is an automated process so this will be acceptable. Later on we

will be combining our application with May1639 in order to give them access to our specifications for their own use. Testing of our application is being done with JUnit testing of our basic functions.

Work Breakdown

Project Schedule:

September	October	November	December
Initial meetings and project planning. Come up with an initial system design. Brainstorm solutions and possible issues.	Begin learning the technologies needed to complete the project. Begin research on feasibility of solutions. Create initial project plan.	Begin initial creation of specification generation and question posting. Refine question wording to get best response from subjects.	Begin initial attempts of analysis on responses. Refine specification generation and wording.

Conclusion

This project will approach the previously unsolved problem of collecting formal software specifications at a large scale. To do this, it will collect data about large library APIs from the Boa system, then post questions asking for the specification for a specific API to StackOverflow to receive answers about these questions. The system will analyze these answers and then add the specifications derived from these answers to the Boa knowledge base.

References

[1] Hridesh Rajan, Tien N. Nguyen, Gary T. Leavens, and Robert Dyer, "Inferring Behavioral Specifications from Large-scale Repositories by Leveraging Collective Intelligence ," *ICSE'15: The 37th International Conference on Software Engineering: NIER Track*, Florence, Italy, May 2015.