Mining of Rationale from Issue Trackers

Master's Thesis
Chair of Applied Software Engineering
Informatics, TU München

Supervisor: Prof. Bernd Brügge, Ph.D.
Advisor: Rana Alkadhi M.Sc.
Author: Ankur Sinha

1 Motivation

Rationale knowledge is particularly important during software maintenance and evolution because it is valuable to understand the rationale behind the previous decisions. Rationale knowledge, in the context of software engineering, corresponds to the reasoning and logic behind the decision taken during all the phases of software development. Such decisive conversations are spread out across various mediums. Rationale may be discussed face-to-face or over other modes of communication such as chats, emails, or issue trackers to mention a few. It is thus important to extract and mine these messages that were exchanged over different modes of communication to manage the rationale. Open sourced projects comes with licenses that provides the rights to study, develop and also contribute to the same. Therefore, this research focuses on mining rationale from open sources projects such as Apache Lucene, Mozilla Thunderbird and Ubuntu. These rationale captured can help developers analyze decisions, improve understandability, and document the pertinent information better.

2 Problem Statement

The aim of this thesis is to analyze the different elements of rationale present in the messages of issue trackers. The focus would be on the following entities:

- **Issue:** the problem to be solved, or a feature to be implemented
- **Alternatives:** the feasible solutions that could address the issue
- **Arguments:** the argument put across in favor of or against a particular alternative
- **Decision:** the final conclusions made to resolve the issue

3 Approach

In this thesis, we apply an empirical approach comprising of the following steps:

1. Formulate the research questions based on the investigation to be done

2. Collect the data from the issue tracker of the following open sourced projects: Apache Lucene, Mozilla Thunderbird and Ubuntu

3. Apply manual content analysis:
   This phase consists of three parts:

   (a) Coding Guide Development: A coding guide will be developed in this phase where the rationale elements to be studied will be defined and examples will be given.

   (b) Manual Annotation: Based on the coding guide, manual content analysis would be done to annotate and classify them. In the second iteration of this particular phase, the disagreements and conflicts between annotators will be handled.

   (c) Ground Truth: Section 3.1 and Section 3.2 together will help in the formation of ground truth which will be used as the benchmark and training set for measuring the accuracy of the model which will perform the automatic classification of rationale.

4. This phase would involve training the classifier based on the ground truth formation made in the previous step. This phase consists of two parts:

   (a) Binary Classification: Classification of data into two categories. In the context of this research, this would be data containing rationale elements and those that do not contain rationale information.

   (b) Multi-Class Classification: Granular classification of data into multiple classes. In the context of this thesis, this would be classifying data into issues, proposals, arguments and decisions.

4 Results:

We expect to have trained a classifier with commendable accuracy and precision to carry out both kinds of classification: binary classification and multi-class classification.

1. The binary classifier will be able to distinguish between the existence and absence of rationale information in a message.

2. The multi-class classifier will be able to make granular classifications by associating a message to the right kind of rationale element such as issue, proposal, argument and decision.

5 Conclusion

By the end of this thesis, we would have built two kinds of models to carry out automatic extraction of rationale elements from messages and classify them into their appropriate classes based on the ground truth obtained by manual analysis.