COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING
DEPARTMENT OF SOFTWARE ENGINEERING

Selected Topic paper 2

Title - Software Reuse:

Group members	**ID No**
1. Gelila Adugna	ETS0312/12
2. Hara Birhanu	ETS0336/12
3. Kokebe Negalgn	ETS0419/12
4. Mastewal Tesfaye	ETS0435/12

SECTION B

Submitted to Mr. Enchalew Y.

Submission Date May 10,05,2024

Introduction

The paper titled "Towards Exploring the Code Reuse from Stack Overflow during Software Development" presents an empirical study on the code reuse activities between Stack Overflow (SO) and GitHub projects, focusing on the reuse of code snippets from SO during software development. The study aims to investigate the extent, trends, and implications of code reuse from SO in the context of software development. It collects data from 793 open-source Java projects and 1,355,617 posts on Stack Overflow and utilizes the CCFinder clone detection algorithm to analyze the code reuse ratio, the relationship between developer experience and code reuse, and the impact of code reuse on bug-related commits. The findings reveal an average code reuse ratio of 6.32% across projects, with a maximum of 8.38%, showing an increasing trend over the years. The study also identifies potential security risks associated with code reuse and the types of Stack Overflow posts that are more likely to be reused by developers, with a focus on popular Java-related technologies such as Android and Spring. The paper concludes by discussing the future work and potential expansion of the study to include more Q&A website platforms and open-source projects with other programming languages.

Problem dealt with the paper.

The problem dealt with in the given paper is the exploration of code reuse from Stack Overflow (SO) during software development. The study aims to investigate the extent, trends, and implications of code reuse from SO in the context of software development. It addresses the issue of how programmers reuse code on SO during the development process, focusing on capturing the development process of programmers in a project and analyzing the code reuse between modified code snippets in commits and code snippets on SO. The paper also aims to determine the code reuse ratio in bug-related modified code snippets and the influence of code reuse on software development. Additionally, the study seeks to identify the types of Stack Overflow posts that are more likely to be reused by developers, with a focus on popular Java-related technologies such as Android and Spring. Overall, the research aims to provide insights into the prevalence, patterns, and implications of code reuse from Stack Overflow in the context of software development.

Methodology and data used by paper.

The paper "Towards Exploring the Code Reuse from Stack Overflow during Software Development" presents a comprehensive empirical study on code reuse activities between Stack Overflow (SO) and GitHub projects. The study focuses on the reuse of code snippets from SO during software development and investigates various aspects of code reuse, including the code reuse ratio, the relationship between developer experience and code reuse, the impact of code reuse on bug-related commits, and the types of Stack Overflow posts that are more likely to be reused by developers.

The methodology employed in the study involves the collection of data from 793 open-source Java projects and 1,355,617 posts on Stack Overflow. The researchers utilize the CCFinder code clone detection tool to extract the cloning relationship between code snippets on SO and modified code snippets involved in commits, and then determine their code reuse relationship according to chronological order. The study also involves the construction of a Stack Overflow code database and an open-source Java project code database, and the use of the Change Distilling algorithm to extract modified code snippets in commits. The data collection process includes obtaining Q&A pair data from the SO website, extracting code snippets from each post, and crawling GitHub popular projects to obtain historical commit files.

The paper discusses the threats to validity, including external validity related to the data collection from SO and the potential impact of expanding the database to other Q&A websites, as well as internal validity related to the definition of code reuse and the limitations of the CCFinder clone detection algorithm.

Findings:

1. The average code reuse ratio across the 793 GitHub projects studied is 6.32%, with a maximum of 8.38%. The code reuse ratio has shown an increasing trend over the years, indicating that code reuse activities are becoming more prevalent in software development.

2. Experienced developers are more likely to reuse knowledge from Stack Overflow, suggesting that developer experience influences code reuse behavior.

3. The code reuse ratio in bug-related modified code snippets is slightly higher (6.35%) than in non-bug-related modified code snippets involved in the commits (6.31%).

4. The code reuse ratio in Java class files that have undergone multiple modifications is more than double the overall code reuse ratio, indicating a higher proportion of code reuse in files that have been extensively modified.

5. The study also identifies the types of Stack Overflow posts that are more likely to be reused by developers, with a focus on popular Java-related technologies such as Android and Spring.

Future Work:

1. The researchers plan to expand the analysis to include more Q&A website platforms and open-source projects with other programming languages, aiming to generalize the findings to other languages and investigate whether the results hold for other programming languages.

2. There is a plan to leverage clone detection tools that can detect type-III and type-IV clone types for more precise detection, as the current clone detection algorithm used in the study can only detect type I and type II clone types.

3. The study also aims to investigate the influence of code reuse on software development and the potential security risks associated with code reuse, particularly in cases where defective code snippets from Stack Overflow may propagate to multiple projects.

Strength and weakness of paper

Strengths:

Large dataset: The study analyzed a large dataset of 793 popular Java projects on GitHub and 1,355,617 Java-related posts from Stack Overflow, which provides a good representation of the code reuse behavior of developers.

Methodology: The authors employed the Change Distilling algorithm to extract modified code snippets from the commits and the CCFinder algorithm to identify code clones between the code snippets on SO and the modified code snippets in the commits, which is a robust approach to identify code reuse.

Insights: The study provides valuable insights into the code reuse behavior of developers, including the types of projects that are more likely to reuse code from SO and the characteristics of the reused code.

Relevance: The study is relevant to the field of software development, as it highlights the importance of Drawbacks:

weakness

Restricted scope: The study solely looks at Stack Overflow and Java projects, which might not be indicative of other platforms and computer languages.

Data integrity: The study depends on the accuracy of the information gathered from Stack Overflow and GitHub, which can be impacted by problems like biased data gathering procedures, missing or corrupted data, or both.

Absence of control variables: Other factors, such as project size and complexity, developer experience level, and kind of development activities, that might influence developers' code reuse behavior are not considered in this study.

Limited generalizability: The results of this study might not apply to other demographics or circumstances, such as developers working on various kinds of projects or understanding the role of SO in the development process.

Based on the paper, a summary of the Motivation, Literature, and Problem sections:

Motivation:

- The paper motivates the study by highlighting the importance of code reuse in software development, as it can reduce development time, improve code quality, and increase productivity.
- The authors also mention that code reuse is a common practice in software development, but it is often done manually, which can be time-consuming and error prone.
- The paper aims to investigate the code reuse phenomenon in software development, specifically focusing on the reuse of code from Stack Overflow (SO) during the development process.

Literature

- The paper reviews the existing literature on code reuse, highlighting the importance of understanding the code reuse behavior of developers.
- The authors discuss the various approaches to code reuse, including manual copying, code libraries, and frameworks.
- The paper also reviews the existing literature on the role of online communities, such as Stack Overflow, in the development process.

Problem:

- The paper identifies the problem of understanding the code reuse behavior of developers, particularly in the context of online communities like Stack Overflow.
- The authors note that while there is a large body of research on code reuse, there is a lack of empirical studies on the reuse of code from online communities.
- The paper aims to address this gap in the literature by investigating the code reuse behavior of developers in the context of Stack Overflow. Based on the paper, here is a summary of the Methodology section, including the data collection and analysis:

Data Collection

- The study collected data from two sources: GitHub and Stack Overflow.
- From GitHub, the authors collected data on 793 popular Java projects, including the commit history and code changes.
- From Stack Overflow, the authors collected data on 1,355,617 Java-related posts, including the question-and-answer text, user information, and tags.

Data Preprocessing:

- The authors used the Change Distilling algorithm to extract modified code snippets from the commit history of the GitHub projects.
- The authors used the CCFinder algorithm to identify code clones between the code snippets on Stack Overflow and the modified code snippets in the commits.

Data Analysis

- The authors analyzed the data using a combination of statistical and machine learning techniques.
- authors used a logistic regression model to identify the factors that influence the likelihood of code reuse from Stack Overflow.
- The authors used a random forest model to predict the probability of code reuse based on the features extracted from the data.

Feature Extraction:

The authors extracted the following features from the data:

- **Project characteristics**: number of commits, number of authors, and project size.
- **Code snippet characteristics:** length, complexity, and frequency of modification.
- **Stack Overflow post characteristics:** question type, answer type, and user reputation.
- **User characteristics:** user type, experience level, and location.

Evaluation Metrics:

- ✓ The authors evaluated the accuracy of the code reuse prediction model using the following metrics:
    - **Precision**: proportion of correctly predicted code reuse instances.
    - **Recall**: proportion of correctly predicted code reuse instances among all actual code reuse instances.

Results:

- The authors presented the results of the study, including the accuracy of the code reuse prediction model, the factors that influence the likelihood of code reuse, and the characteristics of the reused code.