Finding your digital sibling: which other GitHub projects are similar to yours?

Contribution of source code identifiers to GitHub project similarity

INTRODUCTION

- Open-source projects.
- Over 420 million total projects, of which over 284 million are public [1].
- Collaborations, version control, task management, feature requests, and wikis.

RESEARCH QUESTIONS

Determining the contribution of source code identifiers in finding similar projects
to be used as role models, examples or inspiration?

1. How well do source code identifiers indicate similarities between GitHub projects?
2. Do source code identifiers help find a similar project with the same topic?
3. What are naming conventions in source code and how do naming conventions of source code identifiers contribute to the similarities between GitHub projects?

RELATED WORK

- MUDABlue: categorization by source code identifiers [2].
- CLAN: similarity by semantic anchors [3].
- RepoPal: similarity by additional data sources [4].
- CrossSim: similarity based on graphs [5].

CONTACT INFO

J.G.M. Crienen
j.g.m.crienen@student.tudelft.nl

TU Delft

Supervisor: S. Huang
Responsible professor: Dr. S. Proksch

RESEARCH

METHODOLOGY

The research is divided into three parts:

1. **Data collection**
   Defining datasets and extracting identifiers.
1. **Data processing**
   Performing text analysis on the identifiers.
1. **Evaluation**
   Compare results to manually labeled data.

DATA COLLECTION

- Collected through a GitHub search tool [6].
- Consists of 570 projects on five different topics.
- Source code is parsed into Concrete Syntax Trees and identifiers are extracted.
- Identifiers are split based on their casing.
- Twenty random projects are used as queries to analyze our data.

DATA PROCESSING

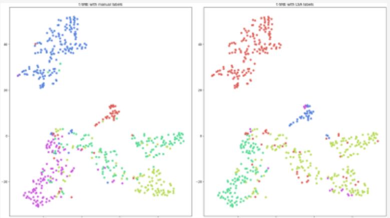
- Identifiers are collected and TF-IDF is used to generate a bag of words.
- Latent Semantic Analysis is used on top of TF-IDF to find hidden relationships.

EVALUATION

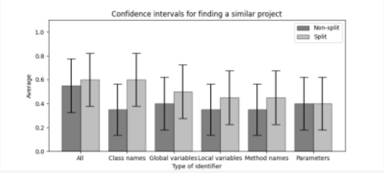
- Interpret plots for topic analysis.
- Analyze success rates in finding similar repositories and topics.
- Compare results to manually labeled data.
- Find whether splitting identifiers results in better performance.

RESULTS

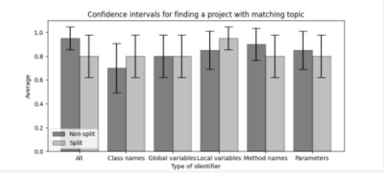
Initial results show that source code identifiers can indicate similar projects and their predicted topics.



We see that combined and split source code identifiers, on average, contribute about 60% with a 95% confidence interval of [0.38, 0.92] to these similarities. Splitting identifiers results in a 10% increase in precision.



Furthermore, we find that source code identifiers are a strong metric for finding projects with the same topic, seeing a 95% success rate with a 95% confidence interval of [0.85, 1.05].



DISCUSSION

- Most projects with the same goal or topic share identifier names.
- Splitting identifiers results in more words to match on and a higher chance for a better match.
- Non-split identifiers are better at finding topic matches because projects often have similar method names or class names.

FUTURE WORK

- What is the influence of type identifiers on the similarity in GitHub projects?
- How well do different clustering techniques model GitHub projects when only looking at source code identifiers?
- What is the effect of a larger test set and query set on our analysis?

CONCLUSIONS

Our analysis shows that source code identifiers are a reasonable metric for finding similar projects, with combined and split identifiers performing the best and therefore do contribute to similarity.

REFERENCES

[1] K. Daigle and GitHub. Octoverse: The state of open source and rise of AI in 2023. The GitHub Blog. 8th Nov. 2023.
[2] S. Kawaguchi et al. "MUDABlue: an automatic categorization system for open source repositories", 2004.
[3] C. McMillan, M. Grechanik and D. Poshyvanyk. "Detecting similar software applications", 2012.
[4] Y. Zhang et al. "Detecting similar repositories on GitHub", 2017.
[5] P. T. Nguyen et al. "CrossSim: Exploiting Mutual Relationships to Detect Similar OSS Projects", 2018
[6] O. Dabic, E. Aghajani and G. Bavota. "Sampling Projects in GitHub for MSR Studies", 2021.