MANCHESTER
1824
The University of Manchester

Mining Software Repositories: The Research of Ourselves

Zhongyan Chen
Department of Computer Science

30th November 2023

Today is about…

Introducing a field of software engineering research

- Brief introduction
- Example Papers (with quiz)
- Challenges in the field

Question:

- How many of you have used Git and GitLab?
- Did you find Git/GitLab useful?

Nonetheless…

- GitLab tracks the process of building software systems, no matter good or bad

- It is a great resource for developers to collab and improve themselves

Do you know…

…that we are collectively building a rich source of data also for CS researchers?

In fact…

- There is a field called mining software repositories
- It’s big enough to have its own conferences!

Mining Software Repositories

It analyzes the rich data available in software repositories to uncover interesting and actionable information about software systems and projects [1]

Mining Software Repositories

It analyzes the rich data available in software repositories to uncover interesting and actionable information about software systems and projects [1]

What are Software Repositories?

- source control systems
- question-and-answer sites
- CI build servers
- code review repositories
- archived communications between project personnel
- defect tracking systems
- run-time telemetry

What do researchers want to learn from them?

- Better understand the system. e.g. evolution
- Human aspects: how do we write software, how can we improve in future developments
- Empirically validate novel ideas and techniques

Benefits

- Empirical evidence is obtained from developers actions
- Only need to **extract** data to our interests (no interviews, surveys)
- Good scalability

Now…

- Let’s take a look a some examples!
- They are published papers

Quiz Time!

- Each question asks the findings of a paper
- **Disclaimer:** the findings are confined within the scope of their respective paper
- Have fun!

1. Can Java writers write better Python programme, according to paper 1?

A. Yes

B. No

C. Not Necessarily ✓

Do Java Programmers Write Better Python? Studying Off-Language Code Quality on GitHub

- “Our data supports the assumption that being knowledgeable in Java or C++ can actually make someone a better Python programmer regarding commonly accepted and object-oriented best practices, but not necessarily with respect to Python-specific conventions.” [2]

2. What may encourage more answers to questions on Stack Overflow, according to paper 2?

A. Short Issue Descriptions

B. Reproducible Issues ✓

C. Relevant Tags

D. Issues with Reputation Rewards

Can Issues Reported at Stack Overflow Questions be Reproduced? An Exploratory Study

In 400 questions:

- **68%** of the code segments require minor and major modifications
- **22%** code segments completely fail to reproduce the issues. [3]

3. Comments written on which weekday are tend to be more negative, reported by paper 3?

A. Friday

B. Sunday

C. Wednesday

D. Monday ✓

Sentiment Analysis of Commit Comments in GitHub: An Empirical Study

If PHP code is producing errors[-2] with register globals on you are terrible[-4] terrible[-4] [-1 consecutive negative words] programmer.[sentence: 1,-5] If you are using magic quotes you are simply stupid[-3].[sentence: 1,-3]

Total score: -5

Sentiment Analysis of Commit Comments in GitHub: An Empirical Study

Weekday	Mean
Monday	-0.043
Tuesday	0.005
Wednesday	0.008
Thursday	0.001
Friday	-0.016
Saturday	-0.027
Sunday	0.022

Challenges in this field

- Unstructured data
- Quality of repositories

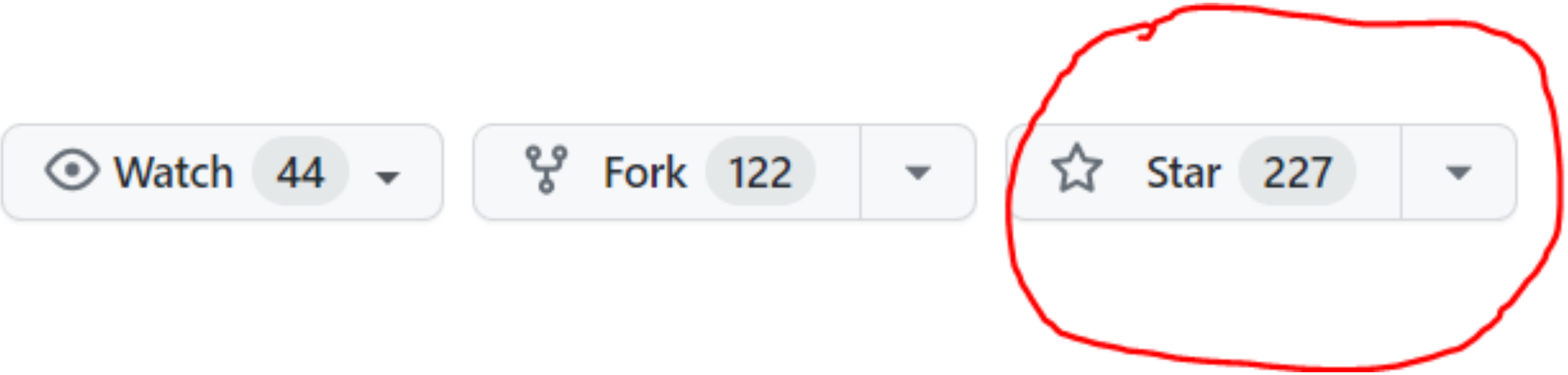
Unstructured data

- Information could be in anywhere, any form

> “In particular, we parse, compile, execute and even carefully examine the code segments from these questions, spent a total of 200 man hours, and then attempt to reproduce their programming issues.” [3]

Quality of repositories

- No common or standardized format for repository records and operations

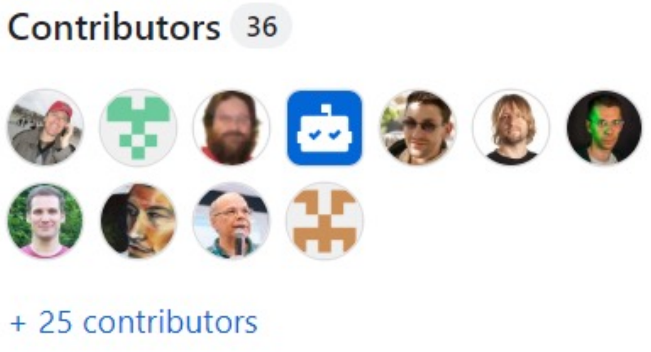


MANCHESTER
1824
The University of Manchester

Ethics

- Why playing with publicly available data has ethical concerns?

Link to a repository →



Contributors 36

+ 25 contributors

→ Email address

⚠

Summary

- An brief introduction to MSR
- Some example papers
- Challenges of MSR

Reference

[1] Hassan A E. The road ahead for mining software repositories[C]//2008 frontiers of software maintenance. IEEE, 2008: 48-57.

[2] Horschig S, Mattis T, Hirschfeld R. Do Java programmers write better Python? studying off-language code quality on GitHub[C]//Companion Proceedings of the 2nd International Conference on the Art, Science, and Engineering of Programming. 2018: 127-134.

[3] Mondal S, Rahman M M, Roy C K. Can issues reported at stack overflow questions be reproduced? an exploratory study[C]//2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). IEEE, 2019: 479-489.

[4] Guzman E, Azócar D, Li Y. Sentiment analysis of commit comments in GitHub: an empirical study[C]//Proceedings of the 11th working conference on mining software repositories. 2014: 352-355.

Want to find out more?

MSR Conference

Paper 1

Paper 2

Paper 3