CS846

Topics in Empirical Software Evolution

Diversity in Software Engineering Research by Nagappan, Zimmermann, and Bird

Presented by Ahmed El Shatshat

Problem Being Solved

- One of the primary goals of Software Engineering Research is to achieve generality
- In software specific research, it's important to collect a diverse and representative set of projects to analyze (e.g. only looking at JSON projects will skew data)
- However, in practice it is difficult to know whether or not you've gathered a sample with high coverage of subgroups
- Thus, there is need for a way to measure the sample you've collected to ensure it has a broad enough coverage of the relevant software space

New Idea

- An introduction of a measure called *sample coverage*
  - Defined as a percentage of projects in a population similar to a given sample
- Not only does this give an accurate measure of the quality of a sample, it also highlights projects that could be added to a sample
- The introduction of a vocabulary of (universe, space and configuration) to calculate similarity metrics

New Idea

Universe:

- Large set of projects
- Also known as population
- Possible universes: all open-source projects, all closed-source projects, all web applications

Space:

- Each software project covers a certain amount of dimensions
- e.g total lines of code, number of developers, main programming language
- The dimension of focus for the research is the **space**

Configuration:

- Similarity between projects is dependent on what dimensions they cover
- A list of similarity functions is called a **configuration**
- A project must be similar in all dimensions to be similar in the universe

New Idea

- Coverage scores do not imply that the research is or is not important; they simply enhance our ability to reason on results
- Generality is difficult to achieve in SE research; regardless, understanding the context of a piece of research greatly aids in gaining deeper insight into research results, even if they differ

Book Chapter: Don't embarrass yourself: Beware of bias in your data

- A biased sample can affect not only your results, but other research that takes from your results
- Not only is identifying bias is important, but also assessing if the bias will actually impact your results
- Worst case scenario, report on your bias so others are aware

Positives

Thorough Evaluation of an Integral Part of Research:

- A bad sampling can completely devalue the worth of a paper, especially if it is not reported on
- The authors provides both good arguments for proper sampling, as well as a thorough methodology for assessing your project's coverage

Well Formatted:

- The authors do a great job of clearly illustrating how their sampling method works, provide a clear example, and follow it with an insightful discussion of the results
- Graphs are easy to understand and provide further insight into results

Universally Applicable:

- Almost all, if not all, research projects have to at least reference other similar projects
- As a result, the research presented provides value to all research; an extremely generalizable result

Negatives

Rambles a Bit Near the End:

- I feel as though the paper could have been cut off much earlier
- Related Work section feels a little backloaded

Feels a Touch Too Conversational At Times

- At times the tone leans a little too conversational for a research paper
- e.g. "consider a researcher who wants to investigate a hypothesis about say distributed development…"

Future Work

- It would be interesting to integrate a tool with research archival platforms like ACM Digital Library to perform these checks
  - Would help with finding similar papers across dimensions
  - Currently it's still quite a pain to find relevant research papers on these platforms
- I wonder if this methodology could extend to other research domains
  - Would need communication with domain experts to see if the Universe/Space/Configuration model would be applicable
  - Would also need to understand what keywords would be relevant if such a model is applicable

Rating

5/5

This paper provides both great insight, and a thorough methodology for the problem it seeks to address.

Discussion Points

- Does a bad sampling ever dissuade you from referencing a research paper?
  - Would that decision be swayed by the paper reporting its sampling as a threat to validity?
- Do you believe that better sampling coverage and reporting sampling scores will lead to more robust results in the SE domain?
- What other considerations do you think could be included in determining coverage?