

Research Statement - Meiyappan Nagappan

My research interests are in the field of Software Engineering specifically focusing on,

- Software Fault Prediction and Identification;
- Software System Log File Analysis;
- Operational Profiling of Software Systems for regression testing; and
- End user interaction (i.e. how do users interact with this information)

I am particularly interested in complimenting the analysis of software metrics data collected about the software system during its development with the analysis of log file data collected from the software system once it is deployed, to identify abnormal behavior and provide the relevant information to the developer/system administrator/technical assistance personnel to help locate the source of the problem. My work is strongly interdisciplinary and lies in the intersection of software engineering, statistics and user centered interaction.

In the research community there are a wide variety of tools that use different techniques to analyze log files. These analysis techniques range from building graphs or automations of the steps in a log file to using clustering and statistical techniques. My research is also complimentary to the dynamic analysis of programs for defect localization and fault identification. But unlike dynamic analysis, there is not much detail in log files. Also none of the prior studies in this area look at how humans debug using these log files and what information is needed/missing in the log files that could help analysis. This is a crucial factor as the human (developer/engineer) is the major decision maker in the process. Other major factors include log abstraction, and incompatibility of these tools across different log files.

In a previous project, I had also interviewed the developers of the Virtual Computing Lab System in NCSU to find out what they looked for in a log file on a regular basis. (I am also starting a project with Dr. Lucas Layman who is at the Fraunhofer Center for Experimental Software Engineering at the University of Maryland. This project deals with qualitatively analyzing interviews with developers to find out what they look for when they are trying to debug applications.)

Therefore I intend to focus my research on the following problems.

- 1) Framework: to develop a component-based adaptable end-to-end framework for the analysis of logs,
- 2) Abstraction and Analytics: to develop examples of pro-active, scalable and adaptive abstractions and analytics suitable for very large scale and very complex logs, and
- 3) Evaluation: to research assess, scalability, adaptability, usability, effectiveness and efficiency metrics for evaluation of the log processing and information extraction algorithms.

As part of my internships at ABB-US Corporate Research Center (ABB-USCRC) and MSR Cambridge (MSR-C), I developed techniques and built prototype tools to help solve some of the problems stated above. More specifically, at ABB-USCRC I helped build tools for operational profiling, and at MSR-C I helped build the framework and tools for finding the root developmental causes for injecting bugs in code.

Also related to this as part of my research project in Lawrence Berkeley National Labs, I developed a new technique to get the operational profile of a system from the execution log files [1]. The technique developed was completely automatic compared to the semi-automated and manual techniques available in the literature for operational profiling. The other projects that I am working on are about adaptive logging, and intelligent log abstraction based on the frequency of words in the log file [2].

Research Directions:

My research goals are two pronged. Firstly I would like to work on projects to develop tools and techniques to help predict and prevent failures, find the root cause of these failures, fix these failure inducing bugs, and find the reasons these bugs were injected in the first place, by leveraging the available data on the software development process and the log file data from the deployed systems.

Secondly I would like to study what information or artifact is required for answering the above problems with greater efficiency and accuracy. I want to investigate why (if) we don't collect those artifacts now, how we can collect it, and how collecting this information would improve the solutions to the above problems.

My experience has been in quantitatively analyzing the production log files, crash data, and data about the development of the products. I have also worked on qualitatively looking at what developers do and where they look for additional information when they debug applications.

I would like to work on projects that use both these kinds of analysis to provide the developers and testers with information they can use to achieve their day to day goals. My experience in working with large and complex data sets and developing solutions to better analyze them, and the experience gained in the industrial labs leads me to believe that a combination of quantitative and qualitative techniques would enable me to address these pressing problems, given the explosion of web services, service –oriented architectures and log file data generated.

References

[1] Nagappan, M., Wu, K., Vouk, M.A., "Efficiently Extracting Operational Profiles from Execution Logs using Suffix Arrays." In the proceedings of the 20th International Symposium on Software Reliability Engineering, 16-19 Nov, 2009, Mysuru, India.

[2] Nagappan, M., Vouk, M.A., "Abstracting Log Lines to Log Event Types for Mining Software System Logs". In the proceedings of Mining Software Repositories (Co-Located with ICSE 2010), 2-3 May, 2010, Cape Town, South Africa.