Connecting the Average and the Non-Average: A Study of the Rates of Fault Detection in Testing WS-BPEL Services

Changjiang Jia, City University of Hong Kong, Tat Chee Avenue, Hong Kong & National University of Defense Technology, Changsha, China
Lijun Mei, IBM Research—China, Beijing, China
W.K. Chan, City University of Hong Kong, Tat Chee Avenue, Hong Kong
Yuen Tak Yu, City University of Hong Kong, Tat Chee Avenue, Hong Kong
T.H. Tse, The University of Hong Kong, Pokfulam, Hong Kong

ABSTRACT

Many existing studies measure the effectiveness of test case prioritization techniques using the average performance on a set of test suites. However, in each regression test session, a real-world developer may only afford to apply one prioritization technique to one test suite to test a service once, even if this application results in an adverse scenario such that the actual performance in this test session is far below the average result achievable by the same technique over the same test suite for the same application. It indicates that assessing the average performance of such a technique cannot provide adequate confidence for developers to apply the technique. The authors ask a couple of questions: To what extent does the effectiveness of prioritization techniques in average scenarios correlate with that in adverse scenarios? Moreover, to what extent may a design factor of this class of techniques affect the effectiveness of prioritization in different types of scenarios? To the best of their knowledge, the authors report in this paper the first controlled experiment to study these two new research questions through more than 300 million APFD and HMFD data points produced from 19 techniques, eight WS-BPEL benchmarks and 1000 test suites prioritized by each technique 1000 times. A main result reveals a strong and linear correlation between the effectiveness in the average scenarios and that in the adverse scenarios. Another interesting result is that many pairs of levels of the same design factors significantly change their relative strengths of being more effective within the same pairs in handling a wide spectrum of prioritized test suites produced by the same techniques over the same test suite in testing the same benchmarks, and the results obtained from the average scenarios are more similar to those of the more effective end than otherwise. This work provides the first piece of strong evidence for the research community to re-assess how they develop and validate their techniques in the average scenarios and beyond.

Keywords: Adverse Scenario, Average Scenario, Controlled Experiment, Correlation, Empirical Study, Regression Testing, Test Case Prioritization, WS-BPEL, XML-based Artifact

DOI: 10.4018/IJWSR.2015070101
INTRODUCTION

In an ecosystem of services, a WS-BPEL program (WS-BPEL Version 2.0, 2007) offers a service by invoking external services (Mei et al. 2014b) to implement its workflow steps. If the business requirements of the ecosystem for the program are not met, its service consumers may discontinue consuming the service it provides, and switch to competing services dynamically. A conventional wisdom is that if a customer discards a product or service, it is intuitively difficult to attract the same customer to reuse the same product or service. Hence, to stay competitive, developers need to rapidly maintain and deploy the service to meet these latest business requirements. Moreover, each modified service should be rapidly and thoroughly tested to reduce the potential impact of any latent fault on its consumers. In short, from the testing viewpoint, maintaining such a service demands highly efficient test sessions, and every test session should be as efficient as possible.

In a regression test session (Leung and White 1989; Onoma et al. 1998), developers execute a modified service over a suite of regression test cases to assess to what extent this modified service passes the regression test. Suppose that these test cases have been prioritized (Rothermel et al. 2001; Rothermel et al. 2002), meaning that some test cases are scheduled to execute earlier than others, in order to expose all the faults in the service detectable by these test cases as fast as possible. Developers would expect that executing these test cases according to their priority can quickly expose faults in their own situations.

Existing studies on test case prioritization (Do et al. 2004; Elbaum et al. 2002; Mei et al. 2014b), or TCP for short, evaluate various design factors of TCP techniques or these techniques directly based on their effectiveness on average (that is, mean or median effectiveness statistically). Yet, in practice, in a regression test session, a developer only applies at most one test case prioritization technique to one test suite once to test the same version of the same program. The same developers do not have the luxury to apply multiple test suites or the same test suite multiple times to look for the average effectiveness of the technique on the service under test.

Thus, even when the average effectiveness of a TCP technique or a design factor in TCP is excellent, if the technique or the factor performs ineffectively in scenarios that are far below average (hereafter simply referred to as the adverse scenarios), the technique or the factor may not be reliably used in practice. This problem is general across a wide range of software domains, and we are particularly interested in it within the regression testing of services.

In the preliminary version of this paper (Jia et al. 2014), we reported the first controlled experiment investigating whether the effectiveness results (the rate of fault detection measured by APFD (Elbaum et al. 2002)) of both TCP techniques and their design factors can be extrapolated from the average scenarios to the adverse scenarios. The controlled experiment included 10 TCP techniques plus two control techniques, eight benchmarks, and 100 test suites per benchmark with 100 repeated applications of every TCP technique on every such test suite. In total, we computed 0.96 million raw APFD values. It compared the consistency between the whole effectiveness dataset of each technique against that of random ordering (labeled as C1 in the Preliminaries section of this paper) and the dataset consisting of the lowest 25 percentile of the former dataset. The results showed that less than half of all the techniques and factors exhibited such consistency.

This paper significantly extends the preliminary version (Jia et al. 2014). We investigate two completely new research questions (denoted by RQ1 and RQ2), which are refined from the research questions presented in the preliminary version. RQ1 investigates the correlation of the change of effectiveness (relative to random ordering) between the average scenarios and adverse scenarios. RQ2 investigates whether switching among levels of a design factor of TCP techniques may result in statistically
Search of Web Service Based on Weighted Association Rule

Promoting Social and Solidarity Economy Through Big Data