

Dynamic Random Testing of Web Services: A Methodology and Evaluation

Chang-ai Sun, *Senior Member, IEEE*, Hepeng Dai, Guan Wang, Dave Towey, *Member, IEEE*, Kai-Yuan Cai, *Member, IEEE*, and Tsong Yueh Chen, *Member, IEEE*,

Abstract— */** Dave [3]: no mention of partition testing in the abstract? **/* In recent years, Service Oriented Architecture (SOA) has been increasingly adopted to develop distributed applications in the context of the Internet. To develop reliable SOA-based applications, an important issue is how to ensure the quality of web services. In this paper, we propose a dynamic random testing (DRT) technique for web services that is an improvement over the widely-practiced random testing (RT). We examine key issues when adapting DRT to the context of SOA, including a framework, guidelines for parameter settings, and a prototype for such an adaptation. Empirical studies are reported where DRT is used to test three real-life web services, and mutation analysis is employed to measure the effectiveness. Our experimental results show that, compared with the two baseline techniques, RT and Random Partition Testing (RPT), DRT demonstrates higher fault-detection effectiveness with a lower test case selection overhead. Furthermore, the theoretical guidelines of parameter setting for DRT are confirmed to be effective. The proposed DRT and the prototype provide an effective and efficient approach for testing web services.

Index Terms—Software Testing, Random Testing, Dynamic Random Testing, Web Service, Service Oriented Architecture.

1 INTRODUCTION

SERVICE oriented architecture (SOA) [2] defines a loosely coupled, standards-based, service-oriented application development paradigm in the context of the Internet. Within SOA, three key roles are defined: service providers (who develop and own services); service requestors (who consume or invoke services); and a service registry (that registers services from providers and returns services to requestors). Applications are built upon services that present functionalities through publishing their interfaces in appropriate repositories, abstracting away from the underlying implementation. Published interfaces may be searched by other services or users, and then invoked. Web services are the realization of SOA based on open standards and infrastructures [3]. Ensuring the reliability of SOA-based applications can become critical when such applications implement important business processes.

Software testing is a practical method for ensuring the quality and reliability of software. However, some SOA features can pose challenges for the testing of web services [4], [5]. For instance, service requestors often do not have access to the source code of web services which are published and owned by another organization, and, consequently, it is not

possible to use white-box testing techniques. Testers may, therefore, naturally turn to black-box testing techniques.

Random Testing (RT) [6] is one of the most widely-practiced black-box testing techniques. Because test cases in RT are randomly selected from the input domain (which refers to the set of all possible inputs of the software under test), it can be easy to implement. Nevertheless, because RT does not make use of any information about the software under test (SUT), or the test history, it may be inefficient in some situations. In recent years, many efforts have been made to improve to RT in different ways [7]–[10]. Adaptive random testing (ART) [8], for example, has been proposed to improve RT by attempting to have a more diverse distribution of test cases in the input domain. */** Dave [4]: I wonder if we could include some other ART references? DART? RART?.... **/*

In contrast to RT, partition testing (PT) attempts to generate test cases in a more “systematic” way, aiming to use fewer test cases to reveal more faults. When conducting PT, the input domain of the SUT is divided into disjoint partitions, with test cases then selected from each and every one. Each partition is expected to have a certain degree of homogeneity—test cases in the same partition should have similar software execution behavior. Ideally, a partition should also be homogeneous in fault detection: If one input can reveal a fault, then all other inputs in the same partition should also be able to reveal a fault. */** Dave [5]: faults or failures? **/*

RT and PT are based on different intuitions, and each have their own advantages and disadvantages. Because it is likely that they can be complementary to each other, detecting different faults, it is intuitively appealing to investigate the their integration. Accordingly, Cai et al. [7] have proposed the random partition testing (RPT) strategy. In RPT, the input domain is first divided into m partitions,

A preliminary version of this paper was presented at the 36th Annual IEEE Computer Software and Applications Conference (COMPSAC 2012) [1].

C.-A. Sun, H. Dai, and G. Wang are with the School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China. E-mail: casun@ustb.edu.cn.

D. Towey is with the School of Computer Science, University of Nottingham Ningbo China, Ningbo 315100, China. E-mail: dave.towey@nottingham.edu.cn

K.-Y. Cai is with the School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China. E-mail: kycai@buaa.edu.cn.

T.Y. Chen is with the Department of Computer Science and Software Engineering, Swinburne University of Technology, Hawthorn VIC 3122, Australia. Email: tychen@swin.edu.au.

s_1, s_2, \dots, s_m , where each s_i is allocated a probability p_i of selection. A partition s_i is randomly selected according to the testing profile $\{\langle s_1, p_1 \rangle, \langle s_2, p_2 \rangle, \dots, \langle s_m, p_m \rangle\}$, where $p_1 + p_2 + \dots + p_m = 1$. A concrete test case is then randomly selected from the chosen s_i .

In traditional RPT testing, the partitions and corresponding test profiles remain constant throughout testing, which may not be the best strategy. Independent researchers from various areas have observed that fault-revealing inputs tend to cluster into “continuous regions” [11], [12]—there is similarity in the execution behavior of neighboring software inputs. Based on software cybernetics, Cai et al. proposed dynamic random testing (DRT) [7], which aims to improve on both RT and RPT. Unlike the original RPT, where the values of p_i are fixed, DRT attempts to dynamically change the values: If a test case from a partition s_i reveals a fault, the corresponding p_i will be increased by a constant ε ; otherwise, it is decreased by ε .

*/** Dave [6]: Please confirm that I have the correct meaning in the following paragraph **/* In practice, web services have usually been tested by the service providers, and simple or easy-to-test faults have been removed, meaning that the remaining faults are normally hard to detect. For ensuring a higher reliability of the web services, a simple RT strategy may not be an appropriate technique, especially when the scale is large, or there are some stubborn faults. Studies have shown that DRT can improve on RT in term of fault detection effectiveness [13], [14], [15].

In this paper, we present a dynamic random testing (DRT) approach for web services, as an enhanced version of RT that is an adaptation of DRT to the context of SOA. We examine key issues of such an adaption, and conduct empirical studies to evaluate the feasibility and effectiveness of the proposed DRT, with the experimental results showing DRT to outperform RT. */** Dave [7]: outperform in what sense? Failure/fault finding? **/* The contributions of this work include:

- We develop an effective and efficient testing technique for web services. This includes a DRT framework that addresses key issues for testing web services, and a prototype that partly automates the framework.
- We evaluate the performance of DRT through a series of empirical studies on three real web services. These studies show that DRT has significantly higher fault-detection effectiveness than RT and RPT, and that its test case selection overhead is lower than RPT. */** Dave [8]: I presume it is not lower than RT **/*
- We provide guidelines for the DRT parameter settings, supported by theoretical analysis, and validated by the empirical studies.

The rest of this paper is organized as follows. Section 2 introduces the underlying concepts for DRT, web services and mutation analysis. Section 3 presents the DRT framework for web services, guidelines for its parameter settings, and a prototype that partially automates DRT. Section 4 describes an empirical study where the proposed DRT is used to test three real-life web services, the results of which are summarized in Section 5. Section 6 discusses related work and Section 7 concludes the paper.

2 BACKGROUND

In this section, we present some of the underlying concepts for DRT, web services, and mutation analysis.

2.1 Dynamic Random Testing (DRT)

DRT combines RT and PT [31], with the goal of benefitting from the advantages of both. Given a test suite TS classified into m partitions (denoted s_1, s_2, \dots, s_m), suppose that a test case from s_i ($i = 1, 2, \dots, m$) is selected and executed. If this test case reveals a fault, $\forall j = 1, 2, \dots, m$ and $j \neq i$, we then set

$$p'_j = \begin{cases} p_j - \frac{\varepsilon}{m-1} & \text{if } p_j \geq \frac{\varepsilon}{m-1} \\ 0 & \text{if } p_j < \frac{\varepsilon}{m-1} \end{cases}, \quad (1)$$

and then

$$p'_i = 1 - \sum_{\substack{j=1 \\ j \neq i}}^m p'_j. \quad (2)$$

Alternatively, if the test case does not reveal a fault, we set

$$p'_i = \begin{cases} p_i - \varepsilon & \text{if } p_i \geq \varepsilon \\ 0 & \text{if } p_i < \varepsilon \end{cases}, \quad (3)$$

and then for $\forall j = 1, 2, \dots, m$ and $j \neq i$, we set

$$p'_j = \begin{cases} p_j + \frac{\varepsilon}{m-1} & \text{if } p_i \geq \varepsilon \\ p_j + \frac{p'_i}{m-1} & \text{if } p_i < \varepsilon \end{cases}. \quad (4)$$

The detailed DRT algorithm is given in Algorithm 1. In DRT, the first test case is taken from a partition that has been randomly selected according to the initial probability profile $\{p_1, p_2, \dots, p_m\}$ (Lines 2 and 3 in Algorithm 1). After each test case execution, the testing profile $\{\langle s_1, p_1 \rangle, \langle s_2, p_2 \rangle, \dots, \langle s_m, p_m \rangle\}$ is updated by changing the values of p_i : */** Dave [9]: there was some inconsistency for the plural of formula: I'm changing all to formulas (not formulae), but please let me know if it should be formulae **/* If a fault is revealed, Formulas 1 and 2 are used; otherwise, Formulas 3 and 4 are used. The updated testing profile is then used to guide the random selection of the next test case (Line 8). This process is repeated until a termination condition is satisfied (Line 1). Examples of possible termination conditions include: “testing resources have been exhausted”; “a certain number of test cases have been executed”; and “a certain number of faults have been detected”.

*/** Dave [10]: Formulas 1 to 3, or 1 to 4? **/* As can be seen from Formulas 1 to 4, updating the testing profile involves m simple calculations, thus requiring a constant time. Furthermore, the selection of partition s_i , and subsequent selection and execution of the test case, all also involve a constant time overhead. The execution time for one iteration of DRT is thus a constant, and therefore the overall time complexity for DRT to select n test cases is $O(n)$.

Algorithm 1 DRT

Input: $\varepsilon, p_1, p_2, \dots, p_m$

- 1: **while** termination condition is not satisfied
- 2: Select a partition s_i according to the testing profile $\{\langle s_1, p_1 \rangle, \langle s_2, p_2 \rangle, \dots, \langle s_m, p_m \rangle\}$.
- 3: Select a test case t from s_i .
- 4: Test the software using t .
- 5: **if** a fault is revealed by t
- 6: Update p_j ($j = 1, 2, \dots, m$ and $j \neq i$) and p_i according to Formulae 1 and 2.
- 7: **else**
- 8: Update p_j ($j = 1, 2, \dots, m$ and $j \neq i$) and p_i according to Formulas 3 and 4.
- 9: **end_if**
- 10: **end_while**

2.2 Web Services

A web service is a platform-independent, loosely coupled, self-contained, programmable, web-enabled application that can be described, published, discovered, coordinated and configured using XML artifacts for the purpose of developing distributed interoperable applications [2]. A web service consists of a description (usually specified in WSDL) and implementation (that can be written in any programming language). Web services present their functionalities through published interfaces, and are usually deployed in a service container. Invocation of a web service requires analysis of the input message in its WSDL, and then appropriate value assignment. */** Dave [11]: assign values to what? **/*

A web service is a basic component of SOA software. Adoption of SOA, in addition to changing the architecture of a system, causes changes to the process of building the system and using it, and this in turn impacts on the testing of the system [5]. */** Dave [12]: did we want to introduce the following list? **/*

- *Lack of access to the service code and structure:* The service users and registry only have access to the service interfaces, which means that white-box testing approaches are not possible.
- *Dynamics and adaptiveness:* With SOA, testers may not be able to determine the component invoked, or the set of possible targets. */** Dave [13]: I'm not sure that I have the intended meaning here. Can you check, or perhaps rephrase? **/*
- *Lack of control:* With traditional software testing, testers can control the components under test, but with SOA, only the services provider has control over the service.
- *Lack of trust:* A service provider may only offer an incomplete or inaccurate description of a service's functional and non-functional behavior, thus further complicating the testing task.

Clearly, SOA testing can be more challenging than traditional testing. Although RT is a widely used software testing method, some of its characteristics may make it inefficient for testing web services.

2.3 Mutation Analysis

Mutation analysis [16] has been widely used to assess the adequacy of test suites and the effectiveness of testing techniques. */** Dave [14]: any other references? **/* Mutation operators are used to seed various faults into the program under test, and thus generate a set of variants, called mutants. If a test case causes a mutant to behave differently to the program under test (for example, by giving different output for the same input), then we say that this test case “kills” the mutant, and thus detects the injected fault. The mutation score (MS) is used to measure how thoroughly a test suite “kills” the mutants. The MS is defined as:

$$MS(p, ts) = \frac{N_k}{N_m - N_e} \quad (5)$$

where p is the program being mutated; ts is the test suite under evaluation; N_k is the number of mutants killed; N_m is the total number of mutants; and N_e is the number of equivalent mutants (mutants whose behavior is always the same as that of p). It has been highlighted that, compared with manually seeded faults, automatically generated mutants can be more similar to real-life faults, and thus the mutant score is a good indicator of the effectiveness of a testing technique [17]. */** Dave [15]: any more recent references? **/* In this paper, we use mutation analysis to evaluate the effectiveness of our proposed DRT for web services.

3 DRT FOR WEB SERVICES

In this section, we describe a framework for applying DRT to web services, discuss guidelines for DRT's parameter settings, and present a prototype that partially automates DRT for web services.

3.1 Framework

Considering the principle of DRT and the features of web services, we propose a DRT for web services framework, as illustrated in Figure 1. In the figure, the DRT components are inside the box, and the web services under test are located outside. Interactions between DRT components and the web services are depicted in the framework. We next discuss the individual framework components.

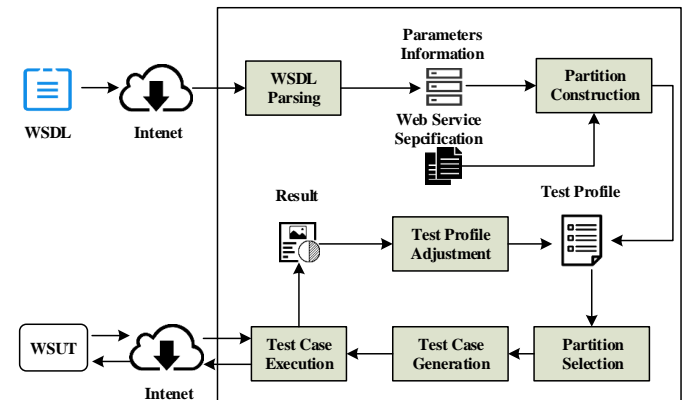


Fig. 1. DRT for web services framework

- 1 *WSDL Parsing*. Web services are composed of services and the relevant WSDL documents. By parsing the WSDL document, we can get the input information for each operation in the services. This includes the number of parameters, their names and types, and any additional requirements that they may have.
- 2 *Partition Construction*. Partition testing (PT) refers to a class of testing techniques that break the input domain into a number of partitions [18]. Because DRT is a black-box testing technique, combining RT and PT, the PT approaches used are at the specification level. Various approaches and principles for achieving convenient and effective partitions have been discussed in the literature [18]–[21]. The input domain of the web service under test (WSUT) can be partitioned based on the WSUT specifications and the parsed parameters. Once partitioned, testers can assign probability distributions to the partitions as an initial testing profile. This initial testing profile can be assigned in different ways, including using a uniform probability distribution, or one that sets probabilities according to the importance of the partition. */** Dave [16]: How is the importance determined? **/*
- 3 *Partition Selection*. DRT randomly selects a partition according to the testing profile.
- 4 *Test Case Generation*. Given the selected partition s_i , a test case is then randomly and independently generated within s_i . Because the WSDL document has been parsed, generation of this test case is non difficult, and can be automated.
- 5 *Test Case Execution*. The relevant DRT component receives the generated test case, converts it into an input message, invokes the web service(s) through the SOAP protocol, and intercepts the test results (from the output message).
- 6 *Test Profile Adjustment*. Upon completion of each test, its pass or fail status is determined by comparing the actual and expected results (with the test passing if both are the same). The pass or fail status is then used to adjust the (partition) probability distribution accordingly. Situations where determination of the test outcome status is not possible (i.e. in the presence of the oracle problem */** Dave [17]: NEED REFERENCES **/*) may potentially be addressed using metamorphic testing [22].

*/** Dave [18]: I'm not sure that I have the correct intended meaning in the next paragraph. Please confirm **/* Generally speaking, DRT test case generation is both in accordance with the probability distribution (for selection of the relevant partition), and with the principles of RT, taking advantage of the ease of RT and the effectiveness of PT. */** Dave [19]: are we sure that PT is effective? **/* Furthermore, many of the DRT for web services framework components can be automated. To make DRT for web services more efficient and practical, we developed a prototype that will be described in Section 3.3.

3.2 Guidelines for Parameter Setting

Our previous work [1] found that the DRT performance can be influenced by the parameter ε (*/** Dave [20]: shall we explain ε here? **/*) and by the number of partitions. We next explore these impacts through a theoretical analysis,

which, to be mathematically tractable, has the following assumptions:

- 1 The failure rate θ_i of each partition s_i ($i = 1, 2, \dots, m$, and $m > 1$) is unknown, but can be estimated.
- 2 Each failure rate θ_i ($i = 1, 2, \dots, m$, and $m > 1$) remains unchanged throughout the testing process (faults are not removed after their detection).
- 3 Test cases are selected with replacement, which means that same test cases may be selected more than once.

A principle of the DRT strategy is to increase the selection probabilities (by amount ε) of partitions with larger failure rates. In addition to the impact of the parameter ε , the number of partitions also influences the speed of updating the testing profile (Formulas 1 and 3). */** Dave [21]: why (here and elsewhere) directly using numbers here and not references/labels for the L^AT_EX for the formulas? **/* Therefore, for a given number of partitions, we are interested in investigating what values of ε yield the best DRT performance.

*/** Dave [22]: note that in the next paragraph I have assumed that more than one partition may have θ_M **/* Letting θ_M denote the maximum failure rate, and s_M denote partitions with that failure rate, then p_i^n denotes the probability of executing the n^{th} test case from partition s_i . As testing proceeds, the probability p_M of partition s_M being selected is expected to increase:

$$p_M^{n+1} > p_M^n \quad (6)$$

Initially, the testing profile is $\{\langle s_1, p_1^0 \rangle, \langle s_2, p_2^0 \rangle, \dots, \langle s_m, p_m^0 \rangle\}$, which, after n test cases have been executed, is then updated to $\{\langle s_1, p_1^n \rangle, \langle s_2, p_2^n \rangle, \dots, \langle s_m, p_m^n \rangle\}$. During the testing process, p_i^n is increased or decreased by the value ε , which is relatively small (set to 0.05 in previous studies [13], [15]). Because the initial p_i^0 is larger than ε , and the adjustment of p_i is relatively small (Formulas 1 to 4), the following two situations are rare, and thus not considered here: $p_i < \varepsilon/(m-1)$ or $p_i < \varepsilon$ ($i = 1, 2, \dots, m$).

As part of exploring the relationship between p_i^{n+1} and p_i^n , we calculate the conditional probability, $p(i|\delta)$, of the following four situations (denoted $\delta_1, \delta_2, \delta_3$, and δ_4 , respectively):

*/** Dave [23]: here and elsewhere, there are full-sops after some formulas (e.g. (7)). Should these be removed? **/*

*/** Dave [24]: We'll probably need to come back later and reword/reformat some of the maths stuff **/*

*/** Dave [25]: Please confirm that the changes I've made below are correct **/*

Situation 1 (δ_1): If $t_n \notin s_i$ and a fault is detected by t_n , then $p(i|\delta_1)$ is calculated according to Formula 1:

$$p(i|\delta_1) = \sum_{i \neq j} \theta_j (p_i^n - \frac{\varepsilon}{m-1}). \quad (7)$$

Situation 2 (δ_2): If $t_n \in s_i$ and a fault is detected by t_n , then $p(i|\delta_2)$ is calculated according to Formula 2:

$$p(i|\delta_2) = \theta_i (p_i^n + \varepsilon). \quad (8)$$

Situation 3 (δ_3): If $t_n \in s_i$ and no fault is detected by t_n , then $p(i|\delta_3)$ is calculated according to Formula 3:

$$p(i|\delta_3) = (1 - \theta_i)(p_i^n - \varepsilon). \quad (9)$$

Situation 4 (δ_4): If $t_n \notin s_i$ and no fault is detected by t_n , then $p(i|\delta_4)$ is calculated according to Formula 4:

$$p(i|\delta_4) = \sum_{j \neq i} (1 - \theta_j)(p_i^n + \frac{\varepsilon}{m-1}). \quad (10)$$

Therefore, p_i^{n+1} for all cases together is:

$$\begin{aligned} p_i^{n+1} &= p_i^n \theta_i (p_i^n + \varepsilon) + p_i^n (1 - \theta_i) (p_i^n - \varepsilon) \\ &\quad + \sum_{j \neq i} p_j^n \theta_j (p_i^n - \frac{\varepsilon}{m-1}) \\ &\quad + \sum_{j \neq i} p_j^n (1 - \theta_j) (p_i^n + \frac{\varepsilon}{m-1}) \\ &= (p_i^n)^2 \theta_i + p_i^n \theta_i \varepsilon + (p_i^n)^2 - p_i^n \varepsilon - (p_i^n)^2 \theta_i + p_i^n \theta_i \varepsilon \\ &\quad + (p_i^n - \frac{\varepsilon}{m-1}) \sum_{j \neq i} p_j^n \theta_j + (p_i^n + \frac{\varepsilon}{m-1}) \sum_{j \neq i} p_j^n \\ &\quad - (p_i^n + \frac{\varepsilon}{m-1}) \sum_{j \neq i} p_j^n \theta_j \\ &= (p_i^n)^2 + 2p_i^n \theta_i \varepsilon - p_i^n \varepsilon + (p_i^n - \frac{\varepsilon}{m-1} - p_i^n \\ &\quad - \frac{\varepsilon}{m-1}) \sum_{j \neq i} p_j^n \theta_j + (p_i^n + \frac{\varepsilon}{m-1})(1 - p_i^n) \\ &= p_i^n + (p_i^n)^2 - (p_i^n)^2 + 2p_i^n \theta_i \varepsilon - p_i^n \varepsilon + \frac{\varepsilon}{m-1} - \\ &\quad \frac{\varepsilon}{m-1} p_i^n - \frac{2\varepsilon}{m-1} \sum_{j \neq i} p_j^n \theta_j \\ &= p_i^n + \frac{\varepsilon}{m-1} (2p_i^n \theta_i m - p_i^n m - 2p_i^n \theta_i + 1) \\ &\quad - \frac{2\varepsilon}{m-1} \sum_{j \neq i} p_j^n \theta_j \\ &= p_i^n + Y_i^n \end{aligned} \quad (11)$$

where

$$\begin{aligned} Y_i^n &= \frac{\varepsilon}{m-1} (2p_i^n \theta_i m - p_i^n m - 2p_i^n \theta_i + 1) \\ &\quad - \frac{2\varepsilon}{m-1} \sum_{j \neq i} p_j^n \theta_j. \end{aligned} \quad (12)$$

From Formula 12, we have:

$$\begin{aligned} Y_M^n - Y_i^n &= \frac{\varepsilon}{m-1} (2p_M^n \theta_M m - p_M^n m - 2p_M^n \theta_M + 1) \\ &\quad - \frac{2\varepsilon}{m-1} \sum_{j \neq M} p_j^n \theta_j - \frac{\varepsilon}{m-1} (2p_i^n \theta_i m - p_i^n m \\ &\quad - 2p_i^n \theta_i + 1) + \frac{2\varepsilon}{m-1} \sum_{j \neq i} p_j^n \theta_j \\ &= \frac{\varepsilon}{m-1} (2m(p_M^n \theta_M - p_i^n \theta_i) - m(p_M^n - p_i^n) - \\ &\quad 2(p_M^n \theta_M - p_i^n \theta_i)) - \sum_{j \neq M} p_j^n \theta_j + \sum_{j \neq i} p_j^n \theta_j \\ &= \frac{2\varepsilon}{m-1} (m(p_M^n \theta_M - p_i^n \theta_i) - \frac{m(p_M^n - p_i^n)}{2} - \\ &\quad (p_M^n \theta_M - p_i^n \theta_i)) + \frac{2\varepsilon}{m-1} (p_M^n \theta_M - p_i^n \theta_i) \\ &= \frac{2\varepsilon}{m-1} (m(p_M^n \theta_M - p_i^n \theta_i) - \frac{m(p_M^n - p_i^n)}{2}). \end{aligned} \quad (13)$$

Before presenting the final guidelines, we need the following lemma.

Lemma 1. If $p_i^n - p_M^n > 2(p_i^n \theta_i - p_M^n \theta_M)$, then $p_M^{n+1} > p_M^n$.

Proof: The condition $p_i^n - p_M^n > 2(p_i^n \theta_i - p_M^n \theta_M)$ can be equivalently expressed as:

$$\frac{p_M^n - p_i^n}{2} < p_M^n \theta_M - p_i^n \theta_i. \quad (14)$$

From Formula 14, $(p_M^n \theta_M - p_i^n \theta_i) - \frac{p_M^n - p_i^n}{2} > 0$, and because $0 < \varepsilon < 1$, and $m > 1$, therefore:

$$\frac{2m\varepsilon}{m-1} ((p_M^n \theta_M - p_i^n \theta_i) - \frac{p_M^n - p_i^n}{2}) > 0. \quad (15)$$

Furthermore:

$$\frac{2\varepsilon}{m-1} (m(p_M^n \theta_M - p_i^n \theta_i) - \frac{m(p_M^n - p_i^n)}{2}) > 0. \quad (16)$$

According to Formulas 16 and 13, if $p_i^n - p_M^n > 2(p_i^n \theta_i - p_M^n \theta_M)$, then $Y_M^n - Y_i^n > 0$.

Also, because $\sum_{i=1}^m p_i^{n+1} = 1$, and $\sum_{i=1}^m p_i^n = 1$, therefore $Y_M^n > 0$, and thus $\sum_{i=1}^m Y_i^n = 0$.

According to Formula 11, $p_M^{n+1} = p_M^n + Y_M^n$. Because $Y_M^n > 0$, therefore $p_M^{n+1} > p_M^n$. \square

Accordingly, we can now present the following theorem that states a sufficient condition for achieving $p_M^{n+1} > p_M^n$.

Theorem 1. For failure rate $\theta_{min} = \min\{\theta_1, \dots, \theta_m\}$, $\theta_M > \theta_{min}$, if $0 < \theta_{min} < \frac{1}{2}$, the following condition is sufficient to guarantee that $p_M^{n+1} > p_M^n$:

$$\frac{2m\theta_{min}^2}{1 - 2\theta_{min}} < \varepsilon < \frac{(m-1)m\theta_{min}}{2(m+1)}. \quad (17)$$

Proof: In order to guarantee $p_M^{n+1} > p_M^n$, we consider the following three situations (where $i \in \{1, 2, \dots, m\}$ and $i \neq M$).

Situation 1 ($p_i^n = p_M^n$): Because $\theta_i < \theta_M$, therefore $(p_i^n \theta_i - p_M^n \theta_M) < 0$.

Therefore, $(p_i^n - p_M^n) > 2(p_i^n \theta_i - p_M^n \theta_M)$.

According to Lemma 1, we have $p_M^{n+1} > p_M^n$.

Situation 2 ($p_i^n > p_M^n$): According to Formula 17, we have the following:

$$\varepsilon > \frac{2m\theta_{min}^2}{1 - 2\theta_{min}}.$$

Because

$$\frac{2m\theta_{min}^2}{1 - 2\theta_{min}} = \frac{\theta_{min}}{1/2m\theta_{min} - 1/m},$$

we have the following:

$$\varepsilon > \frac{\theta_{min}}{1/2m\theta_{min} - 1/m}.$$

Because $\theta_{min} < 1/2$, therefore $1/2m\theta_{min} - 1/m > 0$ and $\varepsilon(1/2m\theta_{min} - 1/m) > \theta_{min}$, which gives $\varepsilon/2m\theta_{min} > \theta_{min} + \varepsilon/m$.

Because $\varepsilon > 0$, and $m > 1$, therefore

$$\frac{1}{2\theta_{min}} > \frac{(\theta_{min} + \varepsilon/m)}{(\varepsilon/m)}.$$

$(1/2\theta_{min})(p_i^n - p_M^n) > (p_i^n - p_M^n)(\theta_{min} + \varepsilon/m)/(\varepsilon/m)$ as $p_i^n > p_M^n$, and

$$p_i^n - p_M^n > 2\theta_{min}(p_i^n - p_M^n) \frac{\theta_{min} + \varepsilon/m}{\varepsilon/m}.$$

Because $(\theta_{min} + \varepsilon/m)/(\varepsilon/m) > 1$, therefore

$$2\theta_{min}(p_i^n - p_M^n) \frac{\theta_{min} + \varepsilon/m}{\varepsilon/m} > 2\theta_{min}(p_i^n - p_M^n).$$

Because $\theta_{min} < \theta_M$, therefore

$$2\theta_{min}(p_i^n - p_M^n) > 2(p_i^n\theta_{min} - p_M^n\theta_M).$$

Thus,

$$p_i^n - p_M^n > 2(p_i^n\theta_{min} - p_M^n\theta_M).$$

According to Lemma 1, we have $p_M^{n+1} > p_M^n$.

Situation 3 ($p_i^n < p_M^n$): For this proof, we make the assumption that $\frac{1}{2} < \theta_M < 1$.

Because we have

$$\varepsilon < \frac{(m-1)m\theta_{min}}{2(m+1)}$$

and

$$\frac{(m-1)m\theta_{min}}{2(m+1)} = \frac{2m - (m+1)}{2(m+1)}m\theta_{min},$$

thus

$$\varepsilon < \left(\frac{m}{m+1} - \frac{1}{2}\right)m\theta_{min}.$$

Obviously, $\varepsilon/m < (m/(m+1) - 1/2)\theta_{min}$ as $m > 1$.

Furthermore, we have

$$-\frac{\varepsilon}{m} > \left(\frac{1}{2} - \frac{m}{m+1}\right)\theta_{min}$$

and

$$\frac{m\theta_{min}}{m+1} - \frac{\varepsilon}{m} + \frac{2\varepsilon}{m} > \frac{\theta_{min}}{2} + \frac{2\varepsilon}{m}$$

which means that

$$\frac{m\theta_{min}}{m+1} + \frac{\varepsilon}{m} > \frac{1}{2}(\theta_{min} + \frac{4\varepsilon}{m}).$$

It follows that

$$(m\theta_{min}/(m+1) + \varepsilon/m)/(4\varepsilon/m + \theta_{min}) > 1/2$$

for any $m > 1, \varepsilon > 0$, and $0 < \theta_{min} < 1$.

Because $\frac{1}{2} < \theta_M < 1$, therefore $(m\theta_{min}/(m+1) + \varepsilon/m)/(4\varepsilon/m + \theta_{min}) > 1/2\theta_M$.

Thus, we have

$$2(p_M^n - p_i^n)\theta_M \frac{\frac{\varepsilon}{m} + \frac{m\theta_{min}}{m+1}}{\frac{4\varepsilon}{m} + \theta_{min}} > p_M^n - p_i^n$$

as $p_M^n > p_i^n$.

Because $\varepsilon/m < 4\varepsilon/m$, and $m\theta_{min}/(m+1) < \theta_{min}$, therefore

$$\frac{\frac{\varepsilon}{m} + \frac{m\theta_{min}}{m+1}}{\frac{4\varepsilon}{m} + \theta_{min}} < 1$$

and

$$2(p_M^n - p_i^n)\theta_M > 2(p_M^n - p_i^n)\theta_M \frac{\frac{\varepsilon}{m} + \frac{m\theta_{min}}{m+1}}{\frac{4\varepsilon}{m} + \theta_{min}}$$

Hence we have

$$2(p_M^n - p_i^n)\theta_M > p_M^n - p_i^n,$$

which can be equivalently expressed as

$$p_i^n - p_M^n > 2(p_i^n - p_M^n)\theta_M.$$

Because $\theta_{min} < \theta_M$, therefore $2(p_i^n - p_M^n)\theta_M > 2(p_i^n\theta_{min} - p_M^n\theta_M)$, and thus

$$p_i^n - p_M^n > 2(p_i^n\theta_{min} - p_M^n\theta_M).$$

According to Lemma 1, we have $p_M^{n+1} > p_M^n$. \square

In summary, when $\frac{1}{2} < \theta_M < 1$, there is always an interval E : */** Dave [26]: Is "E" the correct notation? **/*

$$\varepsilon \in \left(\frac{2m\theta_{min}^2}{1 - 2\theta_{min}}, \frac{(m-1)m\theta_{min}}{2(m+1)}\right) \quad (18)$$

where $\theta_{min} \leq \theta_i, i \in \{1, 2, \dots, m\}$, and $\theta_i \neq 0$, which can guarantee $p_M^{n+1} > p_M^n$.

From the proof above, it is clear that the value of θ_M affects the upper bound (E_{upper}) of E . When $\theta_{min} < \theta_M < \frac{1}{2}$, the value of E_{upper} should close to the lower bound of E . In practice, we should set

$$\varepsilon \approx \frac{2m\theta_{min}^2}{1 - 2\theta_{min}}. \quad (19)$$

3.3 Prototype

Figure 2 shows a screenshot of a prototype tool that partially automates DRT for web services. To start, testers input the address of the web service being tested (the URL of the WSDL), and press the Parse button to analyze the input and output formats. Next, an operation is selected from the operation list (in the bottom left). The tool provides two options for the partitions and test suites: */** Dave [27]: I assume that automatically generating the partitions will also mean automatically generating the test cases. Please confirm. **/* either to automatically generate the partitions (and test cases); or to upload the predefined partitions and test suites. Before beginning testing (by pressing the Test button), testers must

set the maximum number of tests (Test Repetition Limit). During the testing, if a failure is detected before having executed the maximum number of tests, then the tool suspends testing and asks for the tester's instruction. Testers can choose to remove defects and continue testing, or to stop testing. When all tests have completed, the test report is summarized and output in a file.

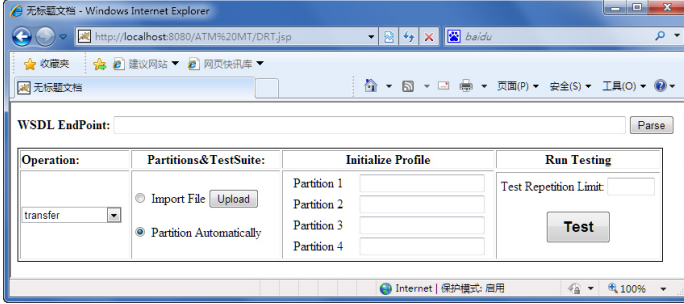


Fig. 2. Prototype interface

4 EMPIRICAL STUDY

We conducted a series of empirical studies to evaluate the performance of DRT.

4.1 Research Questions

In our experiments, we focused on addressing the following three research questions:

- RQ1 How effective is DRT at detecting web service faults?
Fault-detection effectiveness is a key criterion for evaluating the performance of a testing technique. In our study, we chose three commonly used, real-life web services as subject programs, and applied mutation analysis to evaluate the effectiveness.
- RQ2 How do the number of partitions and the DRT parameter ε impact on the failure detection efficiency of DRT? */** Dave [28]: efficiency or effectiveness? **/*
In our earlier work [1], we found that the DRT parameter ε had a significant effect on DRT efficiency, and that the optimal value of the parameter could be related to the number of partitions. The relationship between ε and the number of partitions is examined through theoretical analysis, and verified through the empirical studies.
- RQ3 What is the actual test case generation overhead when using the DRT strategy?
In Section 2.1, we have showed that DRT only requires linear time to generate test case. We wish to validate this theoretical finding through empirical examination of the actual test case generation and execution.

4.2 Subject Web Services

We selected three real-life web services as the subject programs for our study: Aviation Consignment

Management Service (ACMS), China Unicom billing service (CUBS), and Parking billing service (PBS). We used mutation analysis to generate a total of 1563 mutants. After removing equivalent mutants, we then also removed mutants that were too easily detected — deleting mutants that could be detected with less than 20 randomly generated test cases. Table 1 summarizes the basic information of the used web services and their mutants. A detailed description of each web service is given in the following.

TABLE 1
Studied web services

Web service	LOC	Number of mutants
ACMS	116	3
CUBS	131	11
PBS	129	4

4.2.1 Aviation Consignment Management Service (ACMC)

ACMS helps airline companies check the allowance (weight) of free baggage, and the cost of additional baggage. Based on the destination, flights are categorised as either domestic or international. For international flights, the baggage allowance is greater if the passenger is a student (30kg), otherwise it is 20kg. Each aircraft offers three cabins classes from which to choose (economy, business, and first), with passengers in different classes having different allowances. The detailed price rules are summarized in Table 2, where $price_0$ means economy class fare.

4.2.2 China Unicom Billing Service (CUBS)

CUBS provides an interface through which customers can know how much they need to pay according to cell-phone plans, calls, and data usage. The details of several cell-phone plans are summarized in Tables 3, 4, and 5. */** Dave [29]: Please check the tables: it seems strange to talk about Plan A/B/C as the heading of each table, and then have column headings in each table for plan₁, etc. **/* */** Dave [30]: There also seem to be some formatting issues for the tables — width, etc. It may be a good idea to also check what kind of formatting style the journal wants (all capitals; title case, etc.) **/*

4.2.3 Parking Billing Service (PBS)

Consider a parking billing service that accepts the parking details for a vehicle, including the vehicle type, day of the week, discount coupon, and hours of parking. This service rounds up the parking duration to the next full hour, and then calculates the parking fee for according to the hourly rates in Table 6. If a discount voucher is presented, a 50% discount off the parking fee is applied. */** Dave [31]: Should the \$ signs be on the left side, not right, of the figures in Table 6. **/*

To facilitate better parking management, at the time of parking, customers may provide an estimation of parking duration, in terms of three different time ranges ((0.0, 2.0], (2.0, 4.0], and (4.0, 24.0]). If the estimation and actual parked hours fall into the same time range, then the customer will receive a 40% discount; but if they are different ranges, then a 20% markup is applied. A customer may choose to either use a discount coupon, or provide

TABLE 2
ACMC baggage allowance and pricing rules

	Domestic Flights			International Flights		
	first class	business class	economy class	first class	business class	economy class
Carry on (kg)	5	5	5	7	7	7
Free checked-in (kg)	40	30	20	40	30	20/30
Additional baggage pricing (kg)	$price_0 * 1.5\%$			$price_0 * 1.5\%$		

TABLE 3
Plan A

Plan details		cell-phone plan										
		$plan_1$	$plan_2$	$plan_3$	$plan_4$	$plan_5$	$plan_6$	$plan_7$	$plan_8$	$plan_9$	$plan_{10}$	$plan_{11}$
Basic	free calls (min)	50	50	240	320	420	510	700	900	1250	1950	3000
	free data (MB)	150	300	300	400	500	650	750	950	1300	2000	3000
	free incoming calls	domestic (including video calls)										
Extra	incoming calls (CNY/min)	0.25	0.20	0.15								
	data (CNY/KB)	0.0003										
	video calls (CNY/min)	0.60										

TABLE 4
Plan B

Plan details		cell-phone plan					
		$plan_1$	$plan_2$	$plan_3$	$plan_4$	$plan_5$	$plan_6$
Basic	free calls (min)	120	200	450	680	920	1180
	free data (MB)	40	60	80	100	120	150
	free incoming calls	domestic (including video calls)					
Extra	incoming calls (CNY/min)	0.25	0.20	0.15			
	data (CNY/KB)	0.0003					
	video calls (CNY/min)	0.60					

TABLE 5
Plan C

Plan details		cell-phone plan (CNY)		
		$plan_1$	$plan_2$	$plan_3$
Basic	free calls (min)	260	380	550
	free data (MB)	40	60	80
	free data (MB)	domestic (including video calls)		
Extra	incoming calls (CNY/min)	0.25	0.20	0.15
	data (CNY/KB)	0.0003		
	video calls (CNY/min)	0.60		

an estimation of parking duration, but may not do both. (Obviously, a customer may also choose to neither provide an estimation, nor use a discount coupon.) No vehicles are allowed to remain parked for two consecutive days on a continuous basis.

4.3 Variables

4.3.1 Independent Variables

The independent variable in this study is the testing technique, DRT. RPT and RT were used as baseline techniques for comparison.

4.3.2 Dependent Variables

The dependent variable for RQ1 is the metric for evaluating the fault-detection effectiveness. Several effectiveness metrics exist, including: the P-measure (the probability of at least one fault being detected by a test suite); the E-measure

(the expected number of faults detected by a test suite); the F-measure (the expected number of test cases executions required to detect the first fault); and the T-measure (the expected number of test cases required to detect all faults). */** Dave [32]: Perhaps some references for the different measures? **/* Among these, the F-measure and T-measure are the most appropriate for evaluating the fault-detection effectiveness of the DRT testing technique. */** Dave [33]: why? **/* We use F and T to represent the F-measure and the T-measure of a testing method. As shown in Algorithm 1, the testing process may not terminate after the detection of the first fault. Furthermore, because the fault detection information can lead to different probability profile adjustment mechanisms, it is also important to see what would happen after the first fault is revealed. For our study, therefore, we introduce the F2-measure, which is the number of additional test cases required to reveal the second fault after detection of the first fault. We use $F2$ to represent the F2-measure of a testing method, and $SD_{measure}$ to represent the standard deviation of metrics (where $measure$ can be F , $F2$, or T).

An obvious metric for RQ3 is the time required to detect faults. Corresponding to the T-measure, in this study we used $T-time$, the time required to detect all faults. $F-time$ and $F2-time$ denote the time required to detect the first fault, and the additional time needed to detect the second fault (after detecting the first), respectively.

For each of these metrics, smaller values indicate a better performance.

4.4 Experimental Settings

4.4.1 Partitioning

In our study, we made use of a decision table [23], [24] to set the partitions. A decision table is based on a simple principle of sets of actions for sets of constraints, and can be used to present a large amount of complex information in a simple, straightforward manner. Constraints are the pre-conditions of a contract, and need to be resolvable to true or false. */** Dave [34]: do we need to explain contracts? **/* Actions specify valid system responses with respect to constraints, and are called *rules*. They represent the post-conditions of a contract.

TABLE 6
Hourly parking rates

Actual parking hours	Hourly parking rates					
	Weekday			Saturday and Sunday		
	Motorcycle	Car: 2-door coupe	Car: others	Motorcycle	Car: 2-door coupe	Car: others
(0.0, 2.0]	4.00\$	4.50\$	5.00\$	5.00\$	6.00\$	7.00\$
(2.0, 4.0]	5.00\$	5.50\$	6.00\$	6.50\$	7.50\$	8.50\$
(4.0, 24.0]	6.00\$	6.50\$	7.00\$	8.00\$	9.00\$	10.00\$

Thus, every rule of the decision table defines a *contract*. In our study, every *contract* corresponds to a partition.

To investigate the performance of the three testing techniques under different scenarios, we developed the two partition schemes shown in Table 7:

- **Scheme 1:** Every *contract* in the decision table corresponds to a partition. */** Dave [35]: this repeats the last line of the previous paragraph ... **/*
- **Scheme 2:** After constructing the decision table, we group those constraints with the same responses, then every *contract* in the decision table corresponds to a partition. */** Dave [36]: I'm not sure that I understand the intended meaning here. Can you explain or rephrase? **/*

*/** Dave [37]: scheme or schema? Please decide. Do note that some figures have "schema" written in them **/*

TABLE 7
Two partition schemes

Web service	Scheme 1	Scheme 2
ACMS	24	7
CUBS	20	3
PBS	18	3

4.4.2 Initial Test Profile

*/** Dave [38]: Please check that I have the correct meaning in the following **/* Because test cases may be generated randomly during the test process, a feasible method is to use a uniform probability distribution as the initial testing profile. On the other hand, testers may also use past experience to guide a different probability distribution as the initial profile.

4.4.3 Constants

In the experiments, we were interested in exploring the relationship between the number of partitions and the DRT strategy parameter ε , and therefore selected a set of parameter values: $\varepsilon \in \{1.0E-05, 5.0E-05, 1.0E-04, 5.0E-04, 1.0E-03, 5.0E-03, 1.0E-02, 5.0E-02, 1.0E-01, 2E-01, 3E-01, 4E-01, 5E-01\}$. It should be noted that $\varepsilon = 5E-01$ is already a large value. Consider the following scenario. For PBS, when the test is carried out under partition scheme 2, if $\varepsilon = 7.5E-01$ and a uniform probability distribution is used as the testing profile (that is, $p_i = 1/3$), then suppose that the first test case belonging to c_1 is executed and does not reveal any faults, then, according to Formula 3, the value of p_1 would become 0. It is important, therefore, that the initial value of ε should not be set too large.

4.5 Experimental Environment

Our experiments were conducted on a virtual machine running the Ubuntu 11.06 64-bit operating system, with two CPUs, and a memory of 2GB. The test scripts were written in Java. To ensure statistically reliable values of the metrics (F-measure, F2-measure, T-measure, F-time, F2-time, T-time), each testing session was repeated 30 times, with 30 different seeds, and the average value calculated. */** Dave [39]: Is 30 enough times? Do we have evidence? Why "30"? **/*

4.6 Threats To Validity

4.6.1 Internal Validity

A threat to internal validity is related to the implementations of the testing techniques, which involved a moderate amount of programming work. However, our code was cross-checked by different individuals, and we are confident that all techniques were correctly implemented.

4.6.2 External Validity

*/** Dave [40]: the threats to external validity seem quite strong. Have we any other arguments to defend against this/these threat(s)? **/* An obvious possible threat to external validity is that we only considered three web services in the study. However, these are real-life services. Furthermore, 17 distinct faults were used to evaluate the performance. */** Dave [41]: Can we elaborate on "17 distinct faults"? **/* Although we have tried to improve the generalisability of the findings by applying different partitioning granularities, and 13 kinds of parameters, we cannot be certain as to whether or not similar results would be observed in other types of web services.

4.6.3 Construct Validity

The metrics used in our study are simple in concept and straightforward to apply, and hence there should be little threat to the construct validity.

4.6.4 Conclusion Validity

We have run a sufficient number of trials to ensure the statistical reliability of our experimental results. */** Dave [42]: Are we sure? How do we know? **/* Furthermore, as will be discussed in Section 5, statistical tests were conducted to confirm the significance of our results.

5 EXPERIMENTAL RESULTS

5.1 RQ1: Fault Detection Effectiveness

The F-, F2- and T-measure results are summarized in Tables 8 to 10, and their distributions for each program are

TABLE 8
ACMS results

Strategy		Partition Scheme 1						Partition Scheme 2					
		F	SD_F	F2	SD_{F2}	T	SD_T	F	SD_F	F2	SD_{F2}	T	SD_T
RT		13.30	10.34	14.90	24.64	41.80	27.13	13.30	10.34	14.90	24.64	41.80	27.13
RPT		12.04	11.13	12.26	19.08	35.31	25.95	9.03	8.24	15.13	16.08	34.29	29.53
DRT	1.0E-5	11.42	10.16	11.93	18.62	34.93	22.46	8.73	9.88	13.28	18.83	36.19	28.96
	5.0E-5	12.42	10.63	12.05	20.27	36.90	25.39	8.65	9.30	13.53	19.07	35.76	29.86
	1.0E-4	11.34	10.65	11.66	21.27	35.19	26.01	7.36	8.49	13.09	17.89	33.00	27.94
	5.0E-4	12.16	12.13	11.50	19.36	34.19	26.18	7.80	8.37	13.45	17.36	33.59	27.67
	1.0E-3	11.46	11.10	11.39	19.01	34.86	23.17	7.68	8.11	15.07	19.85	33.65	29.95
	5.0E-3	11.02	9.67	12.24	18.83	32.92	20.75	7.47	8.65	15.40	18.81	35.14	29.01
	1.0E-2	10.48	9.60	9.46	14.17	29.59	19.10	7.66	9.18	14.93	18.69	34.30	30.15
	5.0E-2	8.75	6.59	7.06	10.35	23.05	11.67	7.26	7.74	14.70	18.15	34.81	28.14
	1.0E-1	8.59	6.66	6.37	9.41	21.36	10.81	6.67	7.34	16.26	17.54	34.27	27.48
	2.0E-1	8.50	6.21	5.80	9.07	22.10	10.56	5.71	6.04	16.63	10.33	33.58	30.78
	3.0E-1	9.23	6.84	7.12	10.34	22.57	11.13	5.43	6.28	17.60	10.39	33.86	30.33
	4.0E-1	9.22	7.03	6.72	9.36	22.57	10.44	5.14	5.19	17.56	19.11	33.83	28.94
	5.0E-1	8.61	6.41	7.70	10.45	22.64	10.95	5.86	6.50	16.18	17.06	33.31	27.40

TABLE 9
CUBS results

Strategy		Partition Scheme 1						Partition Scheme 2					
		F	SD_F	F2	SD_{F2}	T	SD_T	F	SD_F	F2	SD_{F2}	T	SD_T
RT		21.93	20.37	38.17	38.17	4203.07	3219.40	21.93	20.37	38.17	38.17	4203.07	3219.40
RPT		21.96	20.36	28.74	27.56	2590.38	1768.49	23.66	21.92	27.31	26.99	4195.72	2777.89
DRT	1.0E-5	21.21	23.24	25.62	23.29	2720.70	2051.85	22.90	22.51	25.29	28.21	4106.81	2589.29
	5.0E-5	19.50	19.94	25.77	24.78	2503.45	1873.76	23.88	23.25	26.68	26.84	4130.03	2588.36
	1.0E-4	20.71	22.51	25.59	26.00	2516.91	1843.11	23.55	22.76	26.71	30.26	4196.01	2247.57
	5.0E-4	21.79	21.10	26.82	28.21	2519.39	1942.65	25.27	28.74	26.11	24.36	4190.61	2753.74
	1.0E-3	20.95	20.79	31.41	31.54	2532.84	1752.56	23.84	25.44	27.34	27.64	4291.41	2884.39
	5.0E-3	22.32	21.90	25.93	26.48	2535.97	1572.42	24.11	23.27	26.80	25.20	4218.74	2887.01
	1.0E-2	22.47	21.55	26.01	23.85	2550.88	1873.01	23.46	25.01	26.74	26.43	4117.11	2798.92
	5.0E-2	21.61	20.04	27.66	29.12	2559.56	1777.16	24.01	24.32	26.52	27.04	4105.51	2570.57
	1.0E-1	21.72	21.71	28.31	28.91	2533.08	1774.39	23.30	24.45	26.07	27.91	4271.32	3011.37
	2.0E-1	21.71	21.83	28.68	32.75	2552.29	1879.60	23.55	25.20	28.25	30.31	4170.32	2796.61
	3.0E-1	22.82	21.65	26.68	31.28	2623.00	1770.36	23.40	25.52	27.35	27.84	4138.49	2594.70
	4.0E-1	23.34	24.02	27.32	27.42	2664.34	1886.21	23.07	24.08	29.18	30.39	4192.68	2706.73
	5.0E-1	22.18	22.32	27.14	28.54	2599.40	1640.05	23.30	23.63	26.98	27.71	4195.45	2535.00

TABLE 10
PBS results

Strategy		Partition Scheme 1						Partition Scheme 2					
		F	SD_F	F2	SD_{F2}	T	SD_T	F	SD_F	F2	SD_{F2}	T	SD_T
RT		20.17	16.32	16.50	13.20	252.80	191.54	20.17	16.32	16.50	13.20	252.80	191.54
RPT		17.71	16.78	16.72	25.12	178.91	147.96	16.49	15.76	15.45	21.76	182.66	130.43
DRT	1.0E-5	18.07	19.08	14.27	21.70	176.18	146.49	15.24	15.06	15.31	23.97	163.51	135.48
	5.0E-5	18.23	18.34	15.39	24.35	176.93	143.93	15.13	15.35	14.30	23.57	160.82	116.59
	1.0E-4	16.39	17.42	13.96	21.04	171.61	142.79	15.14	15.05	13.96	22.31	159.93	121.90
	5.0E-4	16.75	17.56	12.60	21.98	165.94	140.33	14.95	12.66	14.98	24.52	166.18	125.20
	1.0E-3	17.96	18.93	15.40	22.22	164.56	138.74	15.72	16.03	16.30	24.41	166.27	128.63
	5.0E-3	16.93	17.23	14.34	22.34	160.70	117.02	15.54	12.79	15.45	23.70	170.14	129.61
	1.0E-2	17.12	16.60	15.10	22.68	166.15	136.59	15.19	15.02	15.13	25.78	166.60	124.97
	5.0E-2	17.02	18.70	16.46	24.75	168.20	129.72	17.10	17.23	15.07	24.09	170.09	130.88
	1.0E-1	17.09	16.78	14.12	22.00	172.94	153.50	16.02	17.16	15.98	24.97	174.16	134.34
	2.0E-1	17.45	18.36	14.14	22.77	174.71	139.02	15.27	15.54	15.52	23.86	167.07	132.81
	3.0E-1	17.23	18.45	16.95	27.43	178.09	161.44	15.43	15.54	15.15	22.04	175.02	136.66
	4.0E-1	17.05	17.72	16.21	20.08	169.21	145.43	15.28	15.93	15.28	24.48	164.67	124.24
	5.0E-1	17.17	18.26	16.23	25.83	172.29	134.94	16.10	15.78	15.26	23.80	167.21	124.05

displayed using boxplots in Figures 3 to 5. In each boxplot, the upper and lower bounds of the box represent the third and first quartiles of the metric, respectively; the middle line represents the median value; the upper and lower whiskers mark, respectively, the largest and smallest data within the range of $\pm 1.5 \times IQR$ (where IQR is the interquartile range); outliers beyond the IQR are denoted with hollow circles; and each solid circle represents the mean value of the metric.

It can be observed from the tables and figures that, in general, DRT is the best performer, followed by RPT. We also conducted statistical testing to verify the significance of this observation, using the Holm-Bonferroni method [25] (with p-value equal to 0.05) to determine which pairs of testing techniques had significant differences. The statistical data are shown in Tables 11 to 13, where each cell gives the number of scenarios where the technique above (in the table) performed better than one to the left. Where the difference is significant, the number is displayed in bold. */* Dave [43]: perhaps we could consider a different way to indicate significance? brackets when not significant, or underlying when significant, for example. Bold is not so easy to see ... */* For example, the bold 69 */* Dave [44]: this originally said 76. Please confirm that it should be 69 */* in the top right cell of Table 12 indicates that, of 78 scenarios (13 parameters \times tow partition schemes \times three web services), DRT had lower F2-measure scores than RT for 69, with the fault-detection capabilities of these two techniques being significantly different. */* Dave [45]: is "scenarios" correct? */*

TABLE 11

Number of scenarios where the technique on the top row has a lower F-measure score than that on the left column

	RT	RPT	DRT
RT	—	4	60
RPT	2	—	61
DRT	18	17	—

TABLE 12

Number of scenarios where the technique on the top row has a lower F2-measure score than that on the left column

	RT	RPT	DRT
RT	—	4	69
RPT	2	—	64
DRT	9	14	—

TABLE 13

Number of scenarios where the technique on the top row has a lower T-measure score than that on the left column

	RT	RPT	DRT
RT	—	6	75
RPT	0	—	64
DRT	3	14	—

Tables 11 to 13 clearly show that the difference between each pair of testing techniques is always significantly different.

5.2 RQ2: Relationship between Partition Number and ε

In 3.2, we analyzed the relationship between the number of partitions and the DRT strategy parameter ε . In this sec-

TABLE 14
Theoretical optimal values of DRT parameter

Web service	partition scheme	θ_{min}	h
ACMS	1	5.452E-2	1.601E-1
	2	2.797E-3	1.102E-4
CUBS	1	1.193E-3	5.702E-5
	2	1.397E-3	1.734E-5
PBS	1	1.760E-3	1.118E-4
	2	1.492E-3	1.340E-5

tion, we show that our theoretical analysis provides useful guidance to testers to set the value of ε .

We used three web services to validate our theoretical analysis. Before starting the test, the failure rate θ_i of partition s_i was obtained by executing (k) test cases from s_i until revealing a fault, then $\theta_i = k/k_i$, where k_i is the total number of test cases in s_i . According to Formula 19, the theoretically optimal values of ε in each scenario for each web service is shown in Table 14, where h denotes the theoretical value of ε . */* Dave [46]: why "h"? */* We ran a series of experiments with the parameters set according to those in Table 14: The F-, F2-, and T-measure results for each program are shown in Figure 6, where $h1$ and $h2$ denote the theoretically optimal value of parameter ε in different partition schemes. */* Dave [47]: why "h1" and "h2"? */* */* Dave [48]: it may be good to give some more explanation of the axes and values in Figure 6 */* Apart from the DRT strategy parameter ε , all other experimental settings remained the same as in Section 5.1.

From Figure 6, we have the following observations:

- In most scenarios, the DRT strategy with theoretically optimum parameter value performs best. Furthermore, the DRT strategy performs better when the parameter values are near the theoretically optimum value than when not.
- From Figure 6 (a), it can be observed that the DRT strategy with larger parameter values performs better than with the theoretically optimum value, in terms of the F-measure. The main reason for this is that, for this scenario, the maximum failure rate ($\theta_M = 4.781E - 3$) is large and the number of partitions is small: When the parameter value is large, the probability of selecting partitions with lower failure rates is quickly reduced, and the probability of selecting partitions with larger failure rates is quickly increased, according to Formulas 3 and 4. */* Dave [49]: I'm not sure that I've captured the intended meaning in the paragraph/point above. Please check and confirm. */*

5.3 RQ3: Selection Overhead

Tables 15 to 17 summarize the F-, F2-, and T-time results, respectively, and their distributions for each web service is shown in Figures 7 to 9. */* Dave [50]: These figures don't say Scheme/a 1 or 2... */* It can be observed from the figures that, in general, DRT had the best performance, and RPT just marginally outperforms RT.

As was done for the F-, F2-, and T-measure data, we used the Holm-Bonferroni method to check the difference

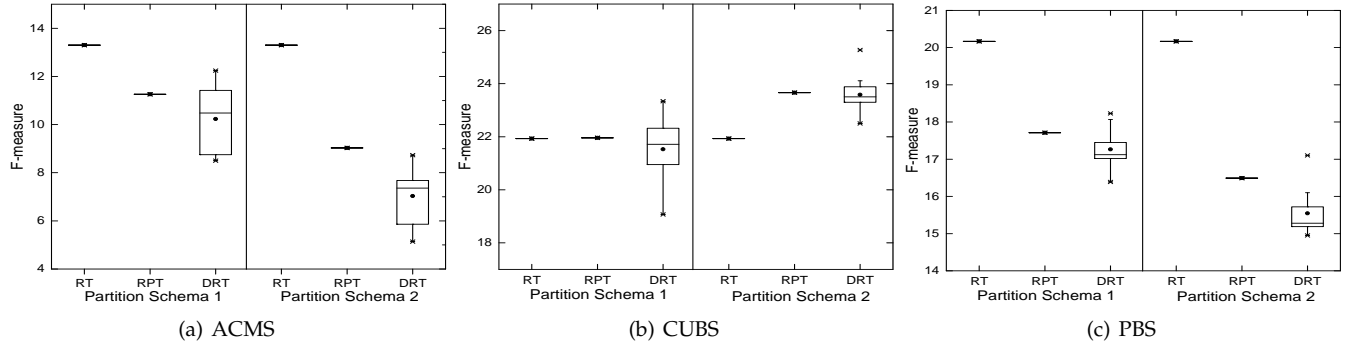


Fig. 3. F-measure boxplots for each program

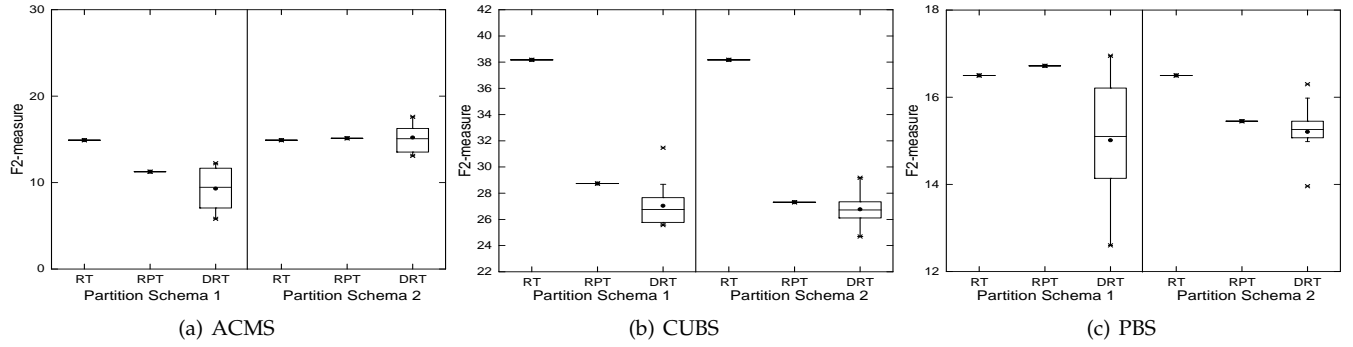


Fig. 4. F2-measure boxplots for each program

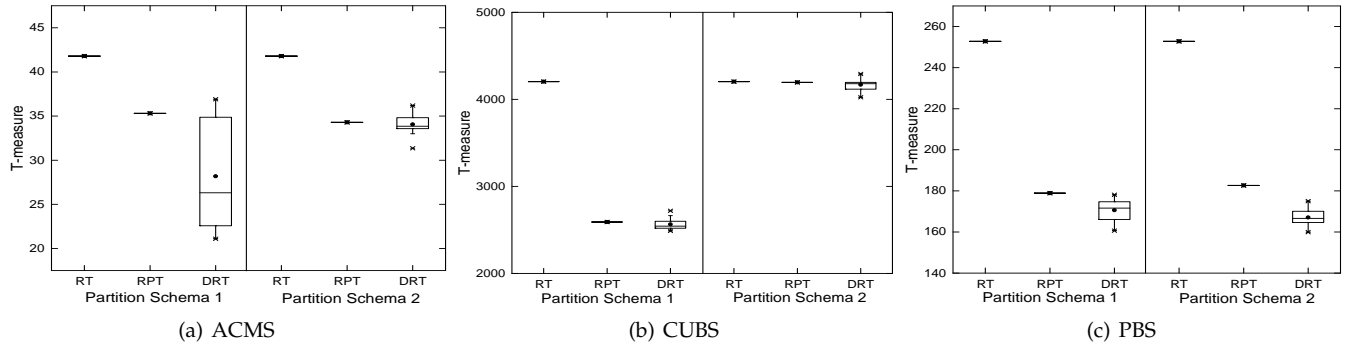


Fig. 5. T-measure boxplots for each program

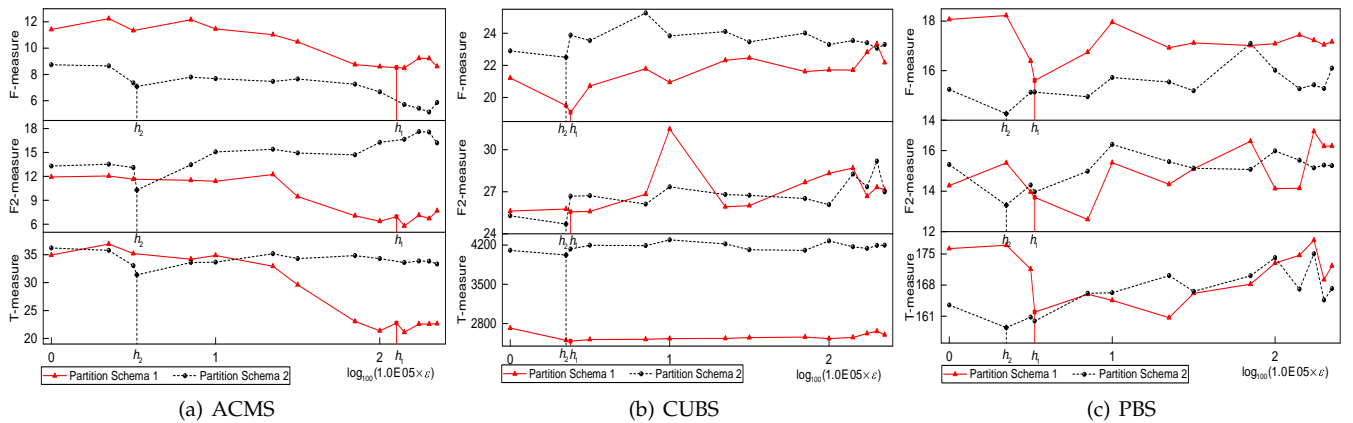


Fig. 6. Line charts of F-measure, F2-measure, and T-measure values for each program (for both the theoretically optimum parameter value, and other values)

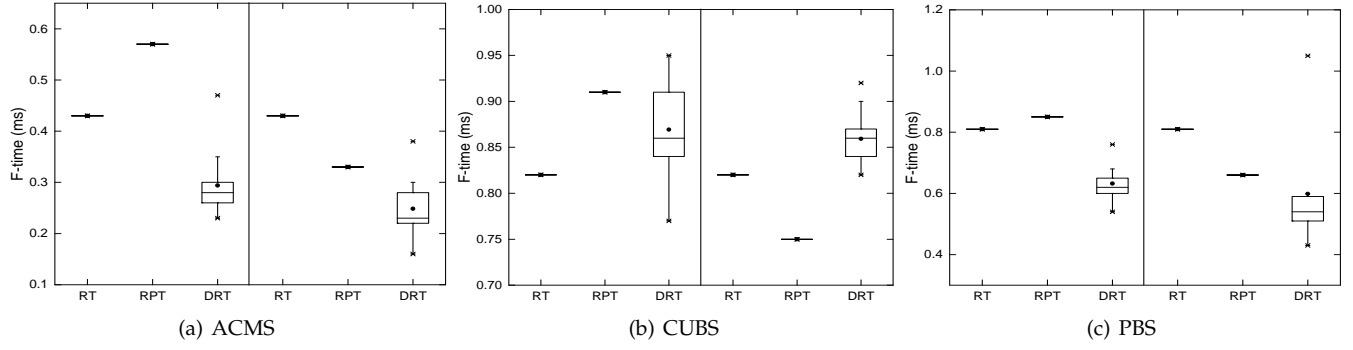


Fig. 7. F-time boxplots for each program

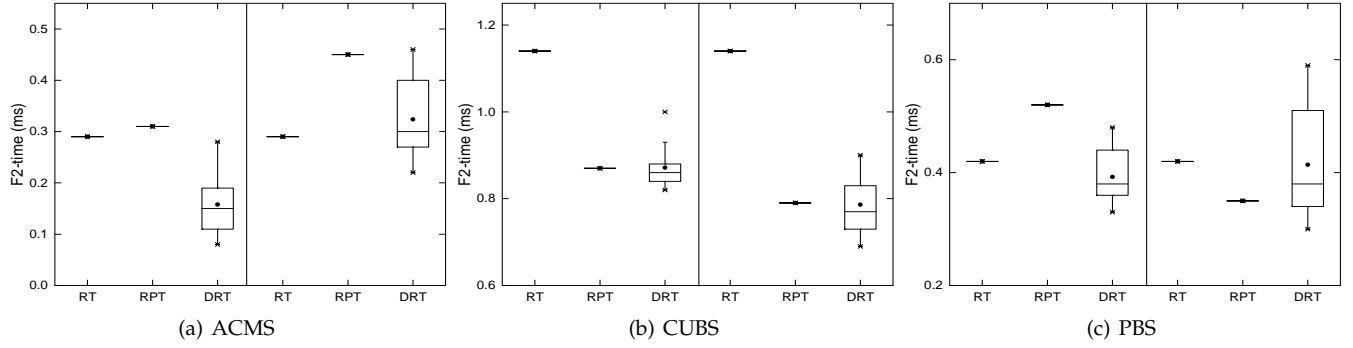


Fig. 8. F2-time boxplots for each program

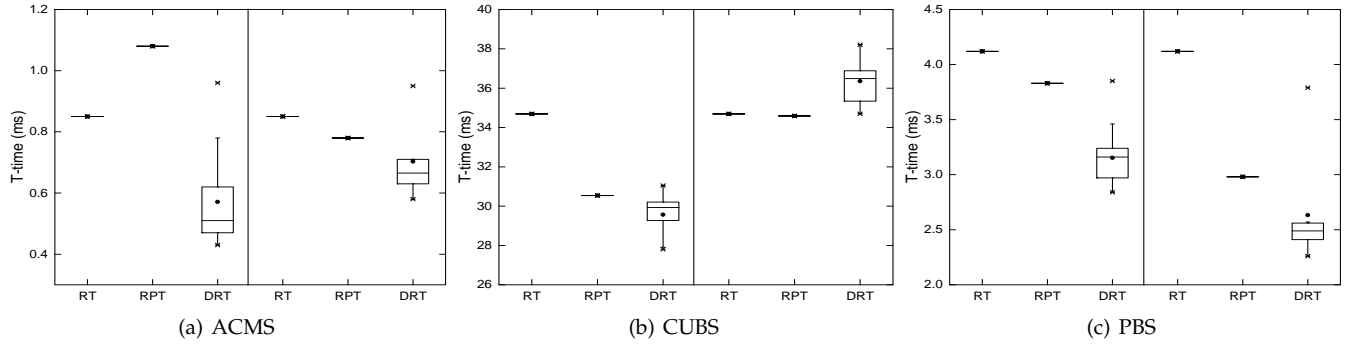


Fig. 9. T-time boxplots for each program

TABLE 15
F-time, F2-time and T-time for Web service ACMS (in ms)

Strategy	Partition Scheme 1			Partition Scheme 2		
	F-time	F2-time	T-time	F-time	F2-time	T-time
RT	0.43	0.29	0.85	0.43	0.29	0.85
RPT	0.57	0.31	1.08	0.33	0.45	0.78
DRT	1.0E-5	0.47	0.28	0.96	0.38	0.41
	5.0E-5	0.35	0.21	0.78	0.30	0.27
	1.0E-4	0.33	0.19	0.66	0.24	0.28
	5.0E-4	0.30	0.17	0.60	0.27	0.25
	1.0E-3	0.29	0.19	0.61	0.22	0.24
	5.0E-3	0.28	0.23	0.62	0.21	0.31
	1.0E-2	0.24	0.17	0.53	0.23	0.29
	5.0E-2	0.28	0.13	0.49	0.23	0.35
	1.0E-1	0.26	0.08	0.43	0.23	0.29
	2.0E-1	0.23	0.12	0.43	0.30	0.46
	3.0E-1	0.24	0.11	0.45	0.18	0.40
	4.0E-1	0.27	0.10	0.47	0.31	0.58
	5.0E-1	0.28	0.11	0.47	0.28	0.45

TABLE 16
F-time, F2-time and T-time for Web service CUBS (in ms)

Strategy	Partition Scheme 1			Partition Scheme 2		
	F-time	F2-time	T-time	F-time	F2-time	T-time
RT	0.82	1.14	34.69	0.82	1.14	34.69
RPT	0.91	0.87	30.54	0.75	0.79	34.59
DRT	1.0E-5	0.91	0.87	30.14	0.87	0.83
	5.0E-5	0.95	0.86	30.21	0.82	0.75
	1.0E-4	0.77	0.89	29.27	0.86	0.69
	5.0E-4	0.79	1.00	31.05	0.87	0.90
	1.0E-3	0.86	0.88	29.93	0.92	0.77
	5.0E-3	0.86	0.83	30.28	0.83	0.83
	1.0E-2	0.95	0.88	29.33	0.84	0.72
	5.0E-2	0.84	0.82	30.56	0.88	0.72
	1.0E-1	0.88	0.93	29.45	0.82	0.76
	2.0E-1	0.83	0.83	30.16	0.86	0.73
	3.0E-1	0.95	0.86	27.81	0.84	0.84
	4.0E-1	0.84	0.84	28.23	0.86	0.82
	5.0E-1	0.87	0.84	27.91	0.90	0.86

TABLE 17
F-time, F2-time and T-time for Web service PBS (in ms)

Strategy		Partition Scheme 1			Partition Scheme 2		
		F-time	F2-time	T-time	F-time	F2-time	T-time
RT		0.81	0.42	4.12	0.81	0.42	4.12
RPT		0.85	0.52	3.83	0.66	0.35	2.98
DRT	1.0E-5	0.76	0.48	3.85	1.05	0.51	3.44
	5.0E-5	0.73	0.35	3.02	0.91	0.52	3.79
	1.0E-4	0.54	0.36	2.97	0.49	0.34	2.26
	5.0E-4	0.68	0.34	3.20	0.50	0.30	2.31
	1.0E-3	0.62	0.38	2.87	0.52	0.38	2.42
	5.0E-3	0.60	0.38	2.99	0.43	0.51	2.40
	1.0E-2	0.58	0.42	2.84	0.51	0.34	2.41
	5.0E-2	0.64	0.44	3.27	0.59	0.36	2.49
	1.0E-1	0.65	0.38	3.46	0.60	0.38	2.56
	2.0E-1	0.60	0.36	3.22	0.54	0.59	2.47
	3.0E-1	0.59	0.44	3.24	0.57	0.46	2.57
	4.0E-1	0.61	0.33	2.90	0.51	0.33	2.54
5.0E-1	0.62	0.44	3.16	0.56	0.36	2.56	

between each pair of testing strategies in terms of F-time, F2-time, and T-time, as shown in Tables 18 to 20.

In Table 20, six entries ("73" & "3" for DRT vs. RT, "58" vs. "20") are in bold font, meaning that, in terms of T-time, DRT was significantly better than RT, and DRT only marginally outperformed RPT. Similar observations can be made regarding the F-time and F2-time results. In other words, the additional computation incurred in DRT by updating the test profile is compensated for in terms of test execution savings.

TABLE 18
Number of scenarios where the technique on the top row has a lower F-time than that on the left column

	RT	RPT	DRT
RT	—	5	62
RPT	1	—	47
DRT	16	31	—

TABLE 19
Number of scenarios where the technique on the top row has a lower F2-time than that on the left column

	RT	RPT	DRT
RT	—	3	52
RPT	3	—	48
DRT	26	30	—

TABLE 20
Number of scenarios where the technique on the top row has a lower T-time than that on the left column

	RT	RPT	DRT
RT	—	5	73
RPT	1	—	58
DRT	3	20	—

It can also be observed from Tables 18 to 20 that DRT only slightly outperformed RPT, but that DRT was significantly better than RT, especially in term of *T-time*.

In summary, the DRT strategy is considered the best testing technique across all six metrics, and RPT marginally outperformed RT.

6 RELATED WORK

In this section, we describe related work from two perspectives: related to testing techniques for web services; and related to improving RT and PT.

6.1 Testing Techniques for Web Services

*/** Dave [51]: I made a lot of changes to this section, please check and confirm **/*

In recent years, a lot of effort has been made to test web services [5], [26]. */** Dave [52]: 2 references, from 8 and 12 years ago may not be sufficient to support our first sentence here ... **/* Test case generation, involving both the generation and selection test cases, is core to testing web services, and model-based [27] and specification-based techniques are two common methods. */** Dave [53]: no reference for "specification-based techniques"? **/* Before making services available on the Internet, testers can use model-based techniques to verify whether or not the behavior of the WSUT meets their requirements. In these techniques, test data can be generated from a data model that specifies the inputs to the software—this data model can be built before, or in parallel to, the software development process. Verification methods using models such as theorem-proving, model-checking and Petri-Nets also exist. Sinha et al. [28], for example, used theorem-proving to generate test cases, making use of existing test generation methods based on extended finite state machine (EFSM) specifications. Majdi et al. [29] generated test data using model-checking. */** Dave [54]: have we anything else to say about Majdi et al. [29]? **/* Dong et al. [30] combined high level Petri-Nets (HPN) with constraint-based test data generation obtaining sufficient test data to detect faults. */** Dave [55]: again, this seems a bit lacking in description ... can we say anything more than "obtaining sufficient test data to detect faults"? **/* When testing web services, because it is often only the service's specification that users can receive, specification-based testing is a natural choice. Typically, the web service specification is contained in the web service description language (WSDL) document, which provides information about the available operations and parameters. Many methods proposed for WSDL-based test data generation are based on the XML schema data type. Hanna and Munro proposed a framework that can be used to test the robustness of a web service [31]. Their framework analyzes the WSDL documents of web services to identify what faults could impact the robustness, facilitating the design of test cases to detect those faults.

The approaches listed above all aim to generate test cases without consideration of the impact of test case execution order on test efficiency. In contrast, Askarunisa et al. [32] performed test case prioritization based on various coverage criteria for composite web services. A major drawback of their approach, however, is the overhead incurred sorting the test cases before testing. Bertolino et al. [33] proposed using the category-partition method [34] with XML schemas to perform XML-based partition testing. Because PT aims to find subsets of all possible test cases to adequately test a system, it can help reduce the required number of test cases. Our proposed approach involved using software cybernetics with PT: In DRT, selection of a partition is done according to the testing profile, which is updated throughout the test

process. An advantage of DRT is that partitions with larger failure rates have higher probabilities of selection. Zhu and Zhang [35] proposed a collaborative testing framework, where test tasks are completed using collaborating test services—a test service is a service assigned to perform a specific testing task. Our framework (Section 3.1) aims to find more faults in the WSUT, with the result of the current test case execution providing feedback to the control system so that the next test case selected has a greater chance to reveal faults.

Most existing web service testing techniques assume that the computed output for each test case is verifiable, something that is not always true in practice. The oracle problem */**Dave [56]: NEED REFERENCES **/* refers to those situations where the test case output is not verifiable, and has meant that many testing techniques may not be applicable in some situations. To address the outstanding oracle problem for testing web services, a metamorphic testing */**Dave [57]: NEED REFERENCES **/* technique has been proposed that not only alleviates the oracle problem, but also presents a feasible and efficient option for testing web services [22]. */** Dave [58]: do we want to elaborate a little more on this approach? **/*

6.2 Improving RT and PT

Based on the observation that failure causing inputs tend to cluster into contiguous regions within the input domain [11], [12], significant work has been done aiming to improve RT. */**Dave [59]: NEED REFERENCES **/* Adaptive random testing [8] is a family of advanced techniques based on random testing that aim to improve the failure detection effectiveness by evenly spreading test cases throughout the input domain. One of the most well-known ART approaches is fixed-size candidate set ART (FSCS-ART), */**Dave [60]: NEED REFERENCES **/*, but many other ART algorithms have been developed and their effectiveness validated through simulation and experiments [8], [36], [37]. */** Dave [61]: shall we list more? RRT? Linear-order ART? etc? **/*

Adaptive testing (AT) [19] is based on the same observation */** Dave [62]: which observation? same as for ART? **/*, and has been shown to outperform both RT and RPT. */** Dave [63]: outperform in what sense? Failure detection? F-measure? E-/P-measure? **/* However, AT may require a very long execution time in practice. To alleviate this, Cai et al. [7] proposed DRT, which uses testing information to dynamically adjust the testing profile. There are several things that can impact on DRT's test efficiency. Yang et al. [14] proposed A-DRT, which adjusts parameters during the testing process. Li et al. [15] developed O-DRT, which changes the testing profile to a theoretically optimal one when a pre-defined criterion is satisfied. */** Dave [64]: can we give an example of the criterion? **/* Lv et al. [13] studied the relationship of parameters in DRT, but did not give the relationship between parameters and the number of partitions. */** Dave [65]: I'm not sure that I understand the intended meaning of the previous sentence. Can you rephrase it? **/*

7 CONCLUSION

In this paper, to address the challenges of testing SOA-based applications, we have presented a dynamic random testing (DRT) technique for web services. Our technique uses random testing to generate test cases, and selects test cases for execution from different partitions in accordance with a testing profile that is dynamically updated in response to the test data collected. In this way, the proposed test technique includes benefits from both random testing and partition testing.

We proposed a framework that examines key issues when applying DRT to testing web services, and developed a prototype to make the method feasible and effective. To help guide testers to correctly set the DRT parameters, we used a theoretical analysis to identify the relationships between numbers of partitions, DRT parameters, and the partitions failure rates. */** Dave [66]: Please check that I have the correct intended meaning in the previous sentence **/* Three real web services were used as experimental objects to validate the feasibility and effectiveness of our approach. The results of the empirical study show that, in general, DRT obtains better performance than both RT and RPT */** Dave [67]: according to what criteria/measures? **/*, and always delivers an outstanding */** Dave [68]: "outstanding"? **/* and stable performance when the DRT parameters lie in our theoretically derived interval. In other words, our theoretical analysis can provide genuinely useful guidance when DRT is used.

In our future work, we plan to conduct experiments on more web services to further validate its effectiveness, and identify the limitations of our method.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (under Grant No. 61872039), the Beijing Natural Science Foundation (Grant No. 4162040), the Aeronautical Science Foundation of China (Grant No. 2016ZD74004), and the Fundamental Research Funds for the Central Universities (Grant No. FRF-GF-17-B29).

REFERENCES

- [1] C.-A. Sun, G. Wang, K.-Y. Cai, and T. Y. Chen, "Towards dynamic random testing for web services," in *Proceedings of the 36th IEEE International Computer Software and Applications Conference (COMPSAC'12)*, 2012, pp. 164–169.
- [2] M. P. Papazoglou, P. Traverso, S. Dustdar, and F. Leymann, "Service-oriented computing: a research roadmap," *International Journal of Cooperative Information Systems*, vol. 17, no. 02, pp. 223–255, 2008.
- [3] C.-A. Sun, E. el Khoury, and M. Aiello, "Transaction management in service-oriented systems: Requirements and a proposal," *IEEE Transactions on Services Computing*, vol. 4, no. 2, pp. 167–180, 2011.
- [4] C. Bartolini, A. Bertolino, S. Elbaum, and E. Marchetti, "Whitening soa testing," in *Proceedings of the 7th joint meeting of the European Software Engineering Conference and the ACM SIGSOFT International Symposium on Foundations of Software Engineering*, 2009, pp. 161–170.
- [5] G. Canfora and M. D. Penta, "Service-oriented architectures testing: A survey," *Software Engineering*, vol. 5413, pp. 78–105, 2009.
- [6] R. Hamlet, "Random testing," in *Encyclopedia of Software Engineering*. John Wiley and Sons, Inc., 2002.
- [7] K.-Y. Cai, H. Hu, C. Jiang, and F. Ye, "Random testing with dynamically updated test profile," in *Proceedings of the 20th International Symposium On Software Reliability Engineering (ISSRE'09)*, 2009, pp. 1–2.

- [8] T. Y. Chen, F.-C. Kuo, R. G. Merkel, and T. Tse, "Adaptive random testing: The art of test case diversity," *Journal of Systems and Software*, vol. 83, no. 1, pp. 60–66, 2010.
- [9] F.-C. Kuo, K. Y. Sim, C.-A. Sun, S. F. Tang, and Z. Zhou, "Enhanced random testing for programs with high dimensional input domains," in *Proceedings of the 19th International Conference on Software Engineering and Knowledge Engineering (SEKE'07)*, 2007, pp. 135–140.
- [10] K.-Y. Cai, B. Gu, H. Hu, and Y.-C. Li, "Adaptive software testing with fixed-memory feedback," *Journal of Systems and Software*, vol. 80, no. 8, pp. 1328–1348, 2007.
- [11] P. E. Ammann and J. C. Knight, "Data diversity: An approach to software fault tolerance," *IEEE Transactions on Computers*, vol. 37, no. 4, pp. 418–425, 1988.
- [12] G. B. Finelli, "NASA software failure characterization experiments," *Reliability Engineering and System Safety*, vol. 32, no. 1, pp. 155–169, 1991.
- [13] J. Lv, H. Hu, and K.-Y. Cai, "A sufficient condition for parameters estimation in dynamic random testing," in *Proceedings of the 3rd IEEE International Workshop on Software Test Automation*, Co-located with the 35th IEEE Annual International Computer Software and Applications Conference, 2011, pp. 19–24.
- [14] Z. Yang, B. Yin, J. Lv, K.-Y. Cai, S. S. Yau, and J. Yu, "Dynamic random testing with parameter adjustment," in *Proceedings of the 6th IEEE International Workshop on Software Test Automation*, Co-located with the 38th IEEE Annual International Computer Software and Applications Conference, 2014, pp. 37–42.
- [15] Y. Li, B. B. Yin, J. Lv, and K.-Y. Cai, "Approach for test profile optimization in dynamic random testing," in *Proceedings of the 39th IEEE Annual International Computer Software and Applications Conference (COMPSAC'15)*, 2015, pp. 466–471.
- [16] R. A. DeMillo, R. J. Lipton, and F. G. Sayward, "Hints on test data selection: Help for the practicing programmer," *Computer*, vol. 11, no. 4, pp. 34–41, 1978.
- [17] J. H. Andrews, L. C. Briand, and Y. Labiche, "Is mutation an appropriate tool for testing experiments?" in *Proceedings of the 27th International Conference on Software Engineering (ICSE'05)*, 2005, pp. 402–411.
- [18] E. J. Weyuker and B. Jeng, "Analyzing partition testing strategies," *IEEE Transactions on Software Engineering*, vol. 17, no. 7, pp. 703–711, 1991.
- [19] K.-Y. Cai, T. Jing, and C.-G. Bai, "Partition testing with dynamic partitioning," in *Proceedings of the 29th Annual International Computer Software and Applications Conference (COMPSAC'05)*, 2005, pp. 113–116.
- [20] T. Y. Chen and Y.-T. Yu, "On the relationship between partition and random testing," *IEEE Transactions on Software Engineering*, vol. 20, no. 12, pp. 977–980, 1994.
- [21] —, "On the expected number of failures detected by subdomain testing and random testing," *IEEE Transactions on Software Engineering*, vol. 22, no. 2, pp. 109–119, 1996.
- [22] C.-A. Sun, G. Wang, B. Mu, H. Liu, Z. Wang, and T. Y. Chen, "Metamorphic testing for web services: Framework and a case study," in *Proceedings of the 2011 IEEE International Conference on Web Services (ICWS'11)*, 2011, pp. 283–290.
- [23] D. Gettys, "If you write documentation, then try a decision table," *IEEE Transactions on Professional Communication*, no. 4, pp. 61–64, 1986.
- [24] Y. Tao and H. Chongzhao, "Entropy based attribute reduction approach for incomplete decision table," in *Proceedings of the 20th International Conference on Information Fusion (Fusion'17)*, 2017, pp. 1–8.
- [25] S. Holm, "A simple sequentially rejective multiple test procedure," *Scandinavian Journal of Statistics*, vol. 6, pp. 65–70, 1979.
- [26] M. Bozkurt, M. Harman, Y. Hassoun *et al.*, "Testing web services: A survey," *Department of Computer Science, Kings College London, Tech. Rep. TR-10-01*, 2010.
- [27] S. R. Dalal, A. Jain, N. Karunanithi, J. Leaton, C. M. Lott, G. C. Patton, and B. M. Horowitz, "Model-based testing in practice," in *Proceedings of the 21st International Conference on Software Engineering (ICSE'99)*, 1999, pp. 285–294.
- [28] A. Sinha and A. Paradkar, "Model-based functional conformance testing of web services operating on persistent data," in *Proceedings of the 2006 Workshop on Testing, Analysis, and Verification of Web Services and Applications*, Co-located with the ACM/SIGSOFT International Symposium on Software Testing and Analysis (ISSTA'06), 2006, pp. 17–22.
- [29] M. Ghannoudi and W. Chainbi, "Formal verification for web service composition: A model-checking approach," in *Proceedings of the International Symposium on Networks, Computers and Communications (ISNCC'15)*, 2015, pp. 1–6.
- [30] W. L. Dong and Y. Hang, "Web service testing method based on fault-coverage," in *Proceedings of the International Workshop on Models for Enterprise Computing (IWMEC'06)*, Co-located with the 10th IEEE International Enterprise Distributed Object Conference (EDOC'06), 2006, pp. 43–43.
- [31] S. Hanna and M. Munro, "An approach for wsdl-based automated robustness testing of web services," in *Information Systems Development*, 2009, pp. 1093–1104.
- [32] A. Askarunisa, K. Arockia Jackulin Punitha, and A. M. Abirami, "Black box test case prioritization techniques for semantic based composite web services using owl-s," in *Proceedings of the 2011 International Conference on Recent Trends in Information Technology (ICRTIT'11)*, 2011, pp. 1215–1220.
- [33] A. Bertolino, J. Gao, E. Marchetti, and A. Polini, "Automatic test data generation for xml schema-based partition testing," in *Proceedings of the 2nd international workshop on automation of software test (AST'07)*, Co-located with the 29th International Conference on Software Engineering (ICSE'07). IEEE Computer Society, 2007, p. 4.
- [34] T. J. Ostrand and M. J. Balcer, "The category-partition method for specifying and generating functional tests," *Communications of the ACM*, vol. 31, no. 6, pp. 676–686, 1988.
- [35] H. Zhu and Y. Zhang, "Collaborative testing of web services," *IEEE Transactions on Services Computing*, vol. 5, no. 1, pp. 116–130, 2012.
- [36] T. Y. Chen, F.-C. Kuo, and C.-A. Sun, "Impact of the compactness of failure regions on the performance of adaptive random testing," *Chinese Journal of Software*, vol. 17, no. 12, pp. 2438–2449, 2006.
- [37] A. C. Barus, T. Y. Chen, F.-C. Kuo, H. Liu, R. Merkel, and G. Rothermel, "A cost-effective random testing method for programs with non-numerical inputs," *IEEE Transactions on Computers*, vol. 65, no. 12, pp. 3509–3523, 2016.



Chang-ai Sun is a Professor in the School of Computer and Communication Engineering, University of Science and Technology Beijing. */** Dave [69]: Do you want to say anything about what happened between assistant professor at Jiaotong and full professor at USTB? **/* Before that, he was an Assistant Professor at Beijing Jiaotong University, China, a postdoctoral fellow at the Swinburne University of Technology, Australia, and a postdoctoral fellow at the University of Groningen, The Netherlands. He received the bachelor degree in Computer Science from the University of Science and Technology Beijing, China, and the PhD degree in Computer Science from the Beihang University, China. */** Dave [70]: Should we say "Beijing University of Aeronautics and Astronautics, Beijing, China." instead of "Beihang"? **/* His research interests include software testing, program analysis, and Service-Oriented Computing.



Hepeng Dai is a PhD student in the School of Computer and Communication Engineering, University of Science and Technology Beijing, China. He received the master degree in Software Engineering from University of Science and Technology Beijing, China and the bachelor degree in Information and Computing Sciences from China University of Mining and Technology, China. His current research interests include software testing and debugging.



Guan Wang is a masters */** Dave [71]: MSc ? **/* student at the School of Computer and Communication Engineering, University of Science and Technology Beijing. He received a bachelor degree in Computer Science from University of Science and Technology Beijing. His current research interests include software testing and Service-Oriented Computing.



Dave Towey is an associate professor in the School of Computer Science, University of Nottingham Ningbo China. He received his BA and MA degrees from The University of Dublin, Trinity College, PgCertTESOL from The Open University of Hong Kong, MEd from The University of Bristol, and PhD from The University of Hong Kong. His current research interests include technology-enhanced teaching and learning, and software testing, especially metamorphic testing and adaptive random testing. He is a member of both the IEEE and the ACM.



Tsong Yueh Chen is a Professor of Software Engineering at the Department of Computer Science and Software Engineering in Swinburne University of Technology. He received his PhD in Computer Science from The University of Melbourne, the MSc and DIC from Imperial College of Science and Technology, and BSc and MPhil from The University of Hong Kong. His current research interests include software testing and debugging, software maintenance, and software design.



Kai-Yuan Cai received the BS, MS, and PhD degrees from Beijing University of Aeronautics and Astronautics, Beijing, China, in 1984, 1987, and 1991, respectively. He has been a full professor at Beihang University since 1995. */** Dave [72]: Should we say "Beijing University of Aeronautics and Astronautics, Beijing, China." instead of "Beihang"? **/* He is a Cheung Kong Scholar (chair professor), jointly appointed by the Ministry of Education of China and the Li Ka Shing Foundation of Hong Kong in 1999. His main research interests include software testing, software reliability, reliable flight control, and software cybernatics. */** Dave [73]: I've modified this bio slightly, please confirm **/*