**基于反馈机制的Web服务测试技术与工具研究**

**技术报告**

代贺鹏

1. 问题阐述
2. 背景介绍
3. DRT在Web服务中的应用

4 经验学习

我们实施了一系列的经验性的学习以便评价DRT策略在Web服务中的表现。

这个部分描述了基本的实验设置。

4.1 研究的问题

RQ1 DRT在Web服务中的测试效率？

故障检测效率是一种十分重要的度量标准，可以用来评价测试技术在测试过程中的表现。通过本文的学习我们研究DRT在Web服务测试中是否具有较高的故障检测能力。

RQ2 变异体分布对DRT测试结果的影响

在实验中我们利用Mujava[]工具得到Web服务的大量变异体。一方面，我们用随机测试的方法挑选出至少需要20个测试用例才能“杀死”的变异体；另一方面，在以往的研究中[]发现：故障趋向于集簇在连续的区域。为了充分模拟现实的故障情况，我们运用[]中的技术实现故障的集簇性并且设置5个集簇的梯度：50/50、60/40、70/0、80/20、90/10，其中80/20故障集簇程度就是著名的Pareto原则：80%故障集簇在20%的代码行中。

RQ3 分区数目对DRT测试效率的影响

分区方式不同对实验结果的影响也是不确定的，本文研究同一种分区方式下不同分区数目对DRT策略测试结果的影响。

RQ4 的取值问题

DRT策略中的取值对实验结果有重大的影响，[会议论文]中将参数设置4个值进行研究，由于设置的数目以及覆盖的范围较小，本文对参数的覆盖范围进行扩展，进一步研究参数对DRT策略的影响。

RQ5 相对于RT以及RPT策略在测试时间上的比较

在第2章节我们介绍了DRT策略产生测试用例仅需要线性的时间。我们希望通过经验学习的方式进一步评价DRT策略相对于RT以及RPT策略在时间开销上的表现情况。

4.2 目标程序

在实验中，我们挑选三个真实Web服务：航空托运费用查询服务、联通计费服务以及停车计费服务作为实验对象。每一个实验对象的具体情况如表4.1所示。

表4.1 实验对象

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Web服务 | 代码行数 | 低失效率变异体数目 | M50/50 | M60/40 | M70/30 | M80/20 | M90/10 |
| 航空托运服务 | 116 | 12 | 12 | 12 | 12 | 12 | 12 |
| 联通计费服务 | 130 | 62 | 40 | 35 | 20 | 14 | 10 |
| 停车计费服务 | 129 |  | 106 | 69 | 26 | 25 | 9 |

4.3 变量

4.3.1 自变量

本文的自变量是我们测试Web服务用到的技术：DRT策略、RPT策略以及RT策略。其中RT以及RPT策略是作为基础技术跟DRT策略进行比较。

4.3.2 因变量

因变量用来度量DRT策略在Web services中测试效率。有很多度量标准判

断测试技术是否有效：P-measure(在当前测试用例集中至少检测出一个故障的概率)、E-measure(在当前测试用例集中期望检测的故障数目)、F-measure(当前测试用例集中检测一个故障需要的测试用例数目)、T-measure(当前测试用例集中检测出所有故障需要的测试用例数目)和NF-measure(当前测试用例集中，揭示第一个故障之后再揭示一个故障需要的测试用例数目)。在这些度量标准中我们认为F-measure、NF-measure、T-measure是最合适评估DRT测试有效性的度量标准。

4.4 实验设置

4.4.1 分区设置

三个实验的分区情况如表4.2所示。

表4.2实验对象的分区数目

|  |  |  |
| --- | --- | --- |
| Web服务 | 分区方式1 | 分区方式2 |
| 航空托运 | 24 | 7 |
| 联通计费 | 20 | 3 |
| 停车计费 | 18 | 3 |

航空托运服务有两种分区方式：一，根据规格说明书将所有的输入进行组合形成原始决策表，决策表中的每一条规则对应一个分区；二，在原始决策表的基础上将具有相同动作的规则进行合并，合并后的决策表每一个规则对应一个分区。

联通计费有两种分区方式：一，根据规格说明书将所有的输入进行组合形成决策表，决策表中的每一条规则对应一个分区；二，仅考虑一个参数：套餐类型A、B、C，按照套餐类型行进分区。

停车计费有两种分区方式：一，根据规格说明书将所有的输入进行组合形成决策表，决策表中的每一条规则对应一个分区；二，仅考虑一个参数：交通工具的类型：摩托车、2门的汽车和4门的汽车。

4.4.2 测试剖面

由于测试用例是在测试的过程中随机产生的，因此用均等的概率分布作为初始测试剖面是一个保守、可行的方案。另外测试人员也可以根据以往的测试经验将某一个或某些分区被选择的概率增大，并且减少某些分区被选择的概率也是可行的初始剖面的设置方案。

4.4.3 参数设置

参数的取值为:

已经是一个很大的数，此时很容易产生这样的情况：当某一个分区内的测试用例没有揭示软件中的故障时，并且此时该分区被选择的概率比较低，由于DRT策略的算法原因，此时很有可能导致该分区被选择的概率变为0。此时明显是不合理的。例如在停车计费服务中当在分区方式2下进行测试时，此时初始测试剖面为均等概率分布，即。假如第一次测试在第一个分区选中一个测试用例并执行后没有揭示软件中的故障，由于，所以。即便是失效率很高的分区也不可能每一个测试用例都能揭示软件中的故障，但是当取值很大时，往往会因为一次的测试结果将某一个分区被选择的概率直接降为0，这显然是不合理的。因此取很大的值(始终成立)没有意义。

4.5 实验环境

本文的实验运行在虚拟机上的Ubuntu16.04操作系统上。在该系统中有2个CPUs和2G的内存。Java语言编写的测试脚本。为了得到可靠的F-measure、NF-measure、T-measure均值，我们在实验中每一组实验取30个随机数种子，每一个随机数种子重复30遍，最后再取平均。

4.6 威胁有效性的因素

4.6.1 内部有效性

影响内部有效性的一个明显的因素是——脚本。脚本涉及到大量的程序开发工作。本实验的脚本被不同的开发人员交叉检查，因此我们很有信心我们的脚本正确的实现了不同的技术。

4.6.2 外部有效性

外部有效性的威胁因素可能是我们仅仅考虑了3个Web服务，然而我们挑选的3个Web服务均是真实存在的。并且一共200个具有低失效率的变异体用来评价DRT策略的效率。尽管对DRT策略的参数做了较为详细的划分以及划分了不同数目的分区，我们依然不能确保在其它程序上出现相似的情况。

4.6.3 构造有效性

我们实验中运用的度量标准F-measure、NF-measure、T-measure在概念上是很简单的并且很容易实现，因此对构造有效性的威胁很小。

4.6.4 结论有效性

我们执行了充足的次数(900次)确保实验结果的可靠性。

5 实验结果

F-measure、NF-measure、T-measure的实验结果展示在表5.1、5.2、5.3中。每一个程序不同实验条件下得结果分布展示在盒图5.1中。在5.1中，盒图得箱体上边界和下边界分别表示某一度量指标下的3/4中位数和1/4中位数，中间的线表示中位数，中间的点表示均值。上“胡须”的上定点以及下“胡须”的下顶点表示在范围内的最大值和最小值，其中表示上四分位数与下四分位数的差。

6、相关工作

7、总结