Prioritizing User-session-based Test Cases for Web Applications Testing

Sreedevi Sampath†, Renée C. Bryce‡,
Gokulanand Viswanath†, Vani Kandimalla‡, A. Güneş Koru†
†Department of Information Systems
University of Maryland, Baltimore County
{sampath, gokul1, gkoru}@umbc.edu
‡School of Computer Science
University of Nevada Las Vegas
{reneebryce, kandimal}@cs.unlv.edu

Abstract

Web applications have rapidly become a critical part of business for many organizations. However, increased usage of web applications has not been reciprocated with corresponding increases in reliability. Unique characteristics, such as quick turnaround time, coupled with growing popularity motivate the need for efficient and effective web application testing strategies. In this paper, we propose several new test suite prioritization strategies for web applications and examine whether these strategies can improve the rate of fault detection for three web applications and their pre-existing test suites. We prioritize test suites by test lengths, frequency of appearance of request sequences, and systematic coverage of parameter-values and their interactions. Experimental results show that the proposed prioritization criteria often improve the rate of fault detection of the test suites when compared to random ordering of test cases. In general, the best prioritization metrics either (1) consider frequency of appearance of sequences of requests or (2) systematically cover combinations of parameter-values as early as possible.

1. Introduction

Web applications are a critical component of many businesses. Failures in this domain result in losses of millions of dollars to organizations [17, 21]. Web applications often have a large and dynamic user base that demands a high level of robustness and reliability. Most web applications must be available 24/7. This requires testers to fix bugs in an application and deploy a new version within a short time frame. In such situations, testers execute regression test suites to ensure that the new version of the application is working as expected. With quick turnaround time being critical for web applications, testers can benefit from test suites that can detect faults early in the test execution cycle.

Unlike other application domains, in web applications, logs of actual usage data are easily available to testers. Converting usage data into test cases is known as user-session-based testing [9, 26]. These user-sessions provide testers with information on exactly how users interact with a web application, making them ideal for test development. A user-session-based test case is a sequence of base requests and parameter-value pairs (also called name-value pairs). Cookies and the originating IP address information are used to convert requests in a web server log into a sequence of user-session-based test cases. Requests with parameter-values specified by a user may access back-end code or retrieve/store data in the database, thus exercising complex interactions between application components. Further, user sessions identify the most frequently accessed parts of an application. This is important in testing because frequently accessed components of a system have significant impact on the user-perceived reliability of an application.

In [26], Sampath et al. reduce user-session-based test suites while maintaining their effectiveness. In this paper, we expand upon Sampath et al.’s [26] previous work. However, instead of further reducing the size of test suites, we prioritize user-session-based test suites with the goal of improving the rate of fault detection. Prioritizing test cases is particularly significant in user-session-based testing because limited resources make it difficult to execute all of the test cases that can accumulate for a frequently-used application in the logged or originally recorded order.

Several strategies exist to prioritize test cases for C programs [6, 7, 11, 24] and Java programs [4, 5, 11]. We expand upon previous work to develop criteria specifically for web application testing. In particular, we consider frequency of user requests and interactions of parameter-values in requests. The main contributions of this paper are: (1) Strategies to prioritize user-session-based web application test suites, (2) Empirical evaluation of the strategies using user-session-based test suites, and (3) Guidance to testers based on the results of our empirical evaluation.

Section 2 presents background in web application testing, user-session-based testing, and test case prioritization.
Section 3 presents the prioritization metrics that we use to order user-session-based tests. Section 4 describes our subject applications and experimental methodology. We present and analyze the results in Section 5. Section 6 concludes.

2. Background

2.1. Web Applications

A web application is a set of web pages that are accessible through a browser over a network. A web page can be either static—in which case the content is the same for all users—or dynamic, such that content depends on user input. Web applications may include integration of numerous technologies; third-party reusable modules; a well-defined, layered architecture; dynamically generated pages with dynamic content; and extensions to an application framework. Large web-based software systems can require millions of lines of code, contain many interactions between objects, and involve significant interaction with users. Changes in user profiles and frequent maintenance changes also complicate automated testing [13].

While several approaches exist for model-based web application test case generation [1, 14, 16, 23], to our knowledge there is little work in the area of test case prioritization for web applications.

2.2. User-session-based Testing

Web applications often rapidly evolve. With each evolution, testers need to run regression tests. One source of regression tests for web-applications is that of user-sessions captured from previous releases of the software. A user-session-based test case is a sequence of HTTP requests containing base requests and name-value pairs that are recorded when a user accesses the application. In the example test case in Table 1, for the following request: Login.jsp

```
name="john"&pswd="doe"
```

the base request is Login.jsp and the parameter-value pairs are name="john" and pswd="doe". Base requests can be HTTP request accesses to both static and dynamic web page content. When available, cookies were used to generate a user-session-based test case. Otherwise, a user-session-based test case begins when a request from a new IP address arrives at the server and ends when the user leaves the web site or the session times out. A 45 minute gap between two requests from a user is considered equivalent to a session timing out. Different strategies can construct test cases for the collected user sessions [9, 20, 22, 27].

Elbaum et al. [9] show that using usage data as test cases is efficient at detecting faults, but unscalable with larger numbers of user sessions. Sampath et al. [26] reduce the size of user-session-based test suites and empirically evaluate the effectiveness of the reduced suites.

The reduction techniques are based on criterion, such as covering all base requests in the application while maintaining the use case representation. The criteria create a test suite smaller than the original suite, but tests are in no particular order. Indeed, test suite reduction techniques strive to reduce the size of a test suite, while maintaining overall fault finding effectiveness. Whereas, test suite prioritization uses the entire test suite for execution, but the test cases are ordered based on pre-determined criteria that attempt to detect faults as quickly as possible in the test execution cycle.

2.3. Test Case Prioritization

In the life cycle of an application, a new version of the application is created as a result of (a) bug fixes and (b) requirements modifications [19]. A large number of test cases may be available from testing previous application versions which can be reused to test a newer version of the application. However, running such tests may take a significant amount of time. Rothermel et. al. report an example for which it can take weeks to execute all of the test cases from a previous version [24]. Due to time constraints, a tester must often select and order these test cases for execution.

One approach to selecting test cases is to schedule the test cases according to some criterion to satisfy a performance goal, scheduling test cases in this manner is known as test case prioritization. Examples of criteria include code coverage, fault likelihood, and fault exposure potential [6, 7, 24]. Binkley uses the semantic differences between two programs to reduce the number of tests that must be run during regression testing [3]. Jones et al. reduce and prioritize test suites that are MC/DC adequate [12]. Lee et al. reduce test suites by using tests that provide coverage of the requirements [15]. Offutt et al. use strategies that reorder tests to select a smaller number of test cases [18].

Additional criteria exist for GUI-based programs. For instance, Bryce and Memon prioritize pre-existing test suites for GUI-based programs by the lengths of tests (i.e., the number of steps in a test case, where a test case is a sequence of events that a user invokes through the GUI), early coverage of all unique events in a test suite, and early event-interaction coverage between windows (i.e., select tests that contain combinations of events invoked from different windows which have not been covered in previously selected tests) [4]. In half of their experiments, event-interaction-based prioritization results in the fastest fault detection rate.

Our work extends prior prioritization work for our specific application to web-based testing. We leverage actual
user-sessions that serve as our test cases and define and evaluate alternate criteria for test case prioritization.

3. Test Case Prioritization Strategies

The test suite prioritization problem is defined in [24]. Given \((T, \Pi, f)\), where \(T\) is a test suite, \(\Pi\) is the set of all test suites that are prioritized orderings of \(T\) obtained by permuting the tests of \(T\), and \(f\) is a function to evaluate the orderings from \(\Pi\) to the real numbers, the problem is to find a permutation, \(\pi \in \Pi\) such that \(\forall \pi' \in \Pi, f(\pi) \geq f(\pi')\).

Prioritization can be based on any criteria. Examples include code coverage, cost estimates, event coverage, and others [4, 7, 8, 24, 30]. In this work, we exploit the characteristics of user-session-based test cases and examine the following functions (or criteria) for web-based applications (each described in more detail in Sections 3.1-3.4):

- **Test length based on number of base requests (Req-LtoS, Req-StoL):** order by the number of HTTP requests in a test case
- **Frequency-based prioritization (MFAS, AAS):** order such that test cases that cover most frequently accessed pages/sequence of pages are selected for execution before test cases that exercise the less frequently accessed pages/sequences of pages.
- **Unique coverage of parameter-values (1-way):** order tests to cover all unique parameter-values as soon as possible
- **2-way parameter-value interaction coverage (2-way):** order tests to cover all pairwise combinations of parameter-values between pages as soon as possible
- **Test length based on number of parameter-values (PV-LtoS, PV-StoL):** order by number of parameter-values used in a test case
- **Random:** randomly permute the order of tests

3.1. Test Lengths

Prioritization of length by base requests selects a next test with the maximum count of base requests, counting duplicates. Since the number of requests in a test case partially determines how much of the application code is exercised by the test case, ordering test cases based on their length can affect the rate of fault detection of the ordered test suite. We order test cases in descending (Req-LtoS) and ascending (Req-StoL) order of length, where length of a test case is defined as the number of base requests the test case contains. For the example test case \(tc1\) in Table 1, the length of \(tc1\) is the number of base requests in \(tc1\), i.e., four.

<table>
<thead>
<tr>
<th>Test case</th>
<th>Parameter-value pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Register.jsp</td>
<td>name=john, pswd=doe, fname=John, lname=Doe</td>
</tr>
<tr>
<td>Login.jsp</td>
<td>name=john, pswd=doe</td>
</tr>
<tr>
<td>Search.jsp</td>
<td>bookid=10</td>
</tr>
<tr>
<td>Logout.jsp</td>
<td>null</td>
</tr>
</tbody>
</table>

Table 1. Example Test Case

3.2. Frequency-based Prioritization

In this prioritization methodology, we prioritize test cases based on the count of the most-frequently accessed sequences of pages that appear in the test case. Since failures in frequently accessed application components have more impact on user-perceived reliability of the application, we hypothesize that by favoring test cases that exercise frequently accessed application components, the rate of fault detection of the ordered test suite can be improved.

We identify the total number of times that each unique sequence of pages is accessed in the entire test suite and construct a frequency table. We consider sequences in terms of base requests (i.e., ignoring the parameter-value pairs) and only sequences that involve interactions between JSP and Java servlet pages, i.e., we do not include sequences that contain static HTML pages. We assume that faults will exist in the application code and thus consider only sequences between pages that access application code. In this paper we consider page sequences of size 2. Using the frequency table to identify the most frequently accessed sequences, we then prioritize the test cases in two ways.

**Most Frequently Accessed Sequence (MFAS).** This approach identifies the most frequently accessed request sequence, \(s_i\), in the test suite and orders test cases in decreasing order of the number of times that \(s_i\) appears in the test case.

**All Accessed Sequences (AAS).** In AAS, the frequency of access of all sequences is used to order the test suite. For each sequence, \(s_i\), in the application, beginning with the most frequently accessed sequence, test cases that have maximum occurrences of these sequences are selected for execution before other test cases in the test suite. Figure 1 describes the AAS test case prioritization algorithm.

Consider the example shown in Table 2. Sequences \(s1, s2, s3,\) and \(s4\) represent all the sequences of pages in the test suite, ordered in decreasing order of frequency of access. Starting with sequence \(s1\), one test case, say \(tc1\), is selected at random. Then, all of the other sequences that \(tc1\) covers are marked satisfied and removed from further consideration, in this example the only other sequence that \(tc1\) cov-
Algorithm: All Accessed Sequences
Input: set of sequences in web application, S, set of test cases T
Output: Prioritized order of test cases, P
for each sequence os ∈ OS do
    T(os) = test cases t ∈ T such that t has maximum occurrences of os, in T
end for

Figure 1. All Accessed Sequences Algorithm

ers is s4. Then, the test cases that cover s2 are considered and one is selected at random, assume tc3. Since tc3 does not cover any other sequence other than s2 maximum number of times, only s2 is marked as satisfied in this step. The algorithm continues selecting test cases until all of the test cases from the table are ordered. Remaining test cases that do not appear in the table are randomly ordered and appended.

Table 2. AAS Example

3.3. Systematic Prioritization by Parameter-Values

Web applications contain pages that contain parameters for which users may specify values. For instance, consider the example test case in Table 1. The Login.jsp page accessed in the test case has two parameters, “name” and “pswd” that can take on values. The user sessions that we prioritize include a discrete number of values that have been specified for these parameters. For instance, test case, tc1 in Table 1 has the parameter “name” set to the value “john”. We refer to this as a parameter-value.

Table 3. Four parameters can take on one of three values each

Unique parameter-value coverage. The 1-way criterion selects a next test that maximizes the number of parameter-values that have not appeared in previously selected tests.

Parameter-value Interaction Coverage. The 2-way criterion selects a next test that maximizes the number of l-way parameter-value interactions between pages that occur in a test. In this paper, we set l=2 for pairwise coverage of parameter-values. Consider the example of 4 parameters that can each take on one of three values from Table 3. Table 4 shows an example of parameter-values that occur in a set of test cases. Table 5 lists the six pairwise parameter-value interactions that occur in Test 1. In our experiments, we count the number of previously uncovered parameter-values in each test and prioritize by selecting the test with the maximum count.

Table 5. 2-way Parameter-Value Interactions

Length by parameter-value counts. During a user-session, a user may specify any number of parameter-values. We prioritize tests by the number of parameter-values in a test case (duplicates included). This includes selecting those tests with the largest number of parameter-values in a test first, called PV-LtoS. We also prioritize in the opposite manner by selecting those tests with the smallest number of parameter-values first, called PV-StoL.

Table 4. A set of test cases

4. Experimental Evaluation

In our experiments, we study the effectiveness of the prioritization strategies by evaluating their fault detection rate.
Independent and Dependent Variables. Independent variables in our study are the user-session-based test suites, the seeded faults and the test case prioritization techniques. Dependent variables are rate of fault detection, average percent of faults detected (APFD) [24], and test execution time.

Subject Applications and Test Suites. We used three web-based applications and their pre-existing test suites, where test suites are the previously recorded user-sessions in experiments by Sampath et al. [26] and Sprenkle et al. [28]. The subject programs have different characteristics: an open-source, e-commerce bookstore (Book) [10], a Course Project Manager (CPM), and the web application used for the Mid-Atlantic Symposium on Programming Languages and Systems (Masplas). Table 6 shows the subject programs and test suite characteristics.

Book. Book allows users to register, login, browse for books, search for books by keyword, rate books, add books to a shopping cart, modify personal information, and logout. Since our interest was in testing consumer functionality, we did not include the administration code in our experiments [26]. Book uses JSP for its front-end and a MySQL database back-end. Sampath et al. [26] collected 125 test cases by sending emails to local newsgroups and by posting advertisements in the University of Delaware’s classifieds web page asking for volunteer users.

CPM. In CPM, course instructors login and create grader accounts for teaching assistants. Instructors and teaching assistants create group accounts for students, assign grades, and create schedules for demonstration time slots. Users interact with an HTML application interface generated by Java servlets and JSPs. CPM manages state in a file-based datastore. Sampath et al. [26] and Sprenkle et al. [28] collected 890 test cases from instructors, teaching assistants, and students using CPM during the 2004-05 and 2005-06 academic years at the University of Delaware.

Masplas. Masplas is a web application developed at the University of Delaware for a regional workshop. Users can register for the workshop, upload abstracts and papers, and view the schedule, proceedings, and other related information. Masplas is written using Java, JSP, and MySQL. Sampath et al. [26, 25] and Sprenkle et al. [29] collected 169 test cases that we use in our experiments.

Evaluation Metrics. For evaluating the prioritization techniques, we assume that prior knowledge of the faults detected by the regression test suites is available to the testers. We evaluate the test suites with respect to their rate of fault detection, the average percent of faults detected (APFD) [24], and the test suite execution time. The rate of fault detection is defined as the total number of faults detected for a given subset of the prioritized test case order. We present the average percent of faults detected (APFD) using the notation in [24]. For a test suite, \( T \) with \( n \) test cases, if \( F \) is a set of \( m \) faults detected by \( T \), then let \( TF_i \) be the position of the first test case \( t \) in \( T' \), where \( T' \) is an ordering of \( T \), that detects fault \( i \). Then, the APFD metric for \( T' \) is given as

\[
APFD = 1 - \frac{TF_1 + TF_2 + \cdots + TF_n}{mn} + \frac{1}{2n} \quad (1)
\]

Informally, APFD measures the area under the curve that plots test suite fraction and the number of faults detected by the prioritized test case order.

In our empirical study we are interested in finding the most faults in the earliest tests (i.e., in the first 10% of the tests executed) and locating 100% of the faults earliest.

We also measure the time taken by each prioritized suite to achieve 100% fault detection. From Sprenkle et al. [28] and through recent experiments, we measured the execution time for replaying the entire suite in the logged or originally recorded order. We then compute the average time per request by dividing the total suite execution time by the total number of requests in the test suite. We use the average time per request to compute the execution time per test case for each of the prioritized order of test cases. Note that these execution times are the time taken for a replay of test cases in the logged order and not the time taken for fault detection replay (difference between the two types of replay is described in Sprenkle et al. [28]).

Experimental Methodology. From previous work by Sampath et al. [26] and Sprenkle et al. [28, 29] we had information on how many faults are detected by each test case, i.e., a fault matrix, mapping each test case to the faults detected by test case. The fault matrices used in this paper are generated by using the struct oracle for CPM and Masplas and the diff oracle for Book [28, 29]. As described in [26, 28], faults were seeded manually by graduate and undergraduate students. In addition, some naturally occurring faults discovered during deployment were also seeded in the applications [26]. In general, five types of faults were seeded into the applications—data store (faults that exercise application code interacting with the data store), logic (application code logic errors in the data and control flow), form

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Book</th>
<th>CPM</th>
<th>Masplas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classes</td>
<td>11</td>
<td>75</td>
<td>9</td>
</tr>
<tr>
<td>Methods</td>
<td>319</td>
<td>173</td>
<td>22</td>
</tr>
<tr>
<td>Comments</td>
<td>1720</td>
<td>1260</td>
<td>108</td>
</tr>
<tr>
<td>Non-commented Lines of Code</td>
<td>7615</td>
<td>9401</td>
<td>999</td>
</tr>
<tr>
<td>Seeded faults</td>
<td>40</td>
<td>135</td>
<td>29</td>
</tr>
<tr>
<td>Total number of user sessions</td>
<td>125</td>
<td>890</td>
<td>169</td>
</tr>
<tr>
<td>Total number of requests</td>
<td>3640</td>
<td>12552</td>
<td>1107</td>
</tr>
<tr>
<td>Number of unique requests</td>
<td>10</td>
<td>69</td>
<td>24</td>
</tr>
<tr>
<td>Largest user session in number of requests</td>
<td>160</td>
<td>585</td>
<td>69</td>
</tr>
<tr>
<td>Average user session in number of requests</td>
<td>29</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>% of 2-way parameter-value interactions covered in pre-existing test suite</td>
<td>92.5%</td>
<td>97.8%</td>
<td>96.2%</td>
</tr>
</tbody>
</table>

Table 6. Subject Applications and Test Suite Characteristics.
(modifications to parameter-value pairs and form actions), appearance (faults which change the way in which the user views the page), and link (faults that change the hyperlinks location) [26]. Fault categories are not mutually exclusive.

In [28], Sprenkle et al. propose the with state replay mechanism used for collecting the fault matrices that we use in this paper. In with state test suite execution on a faulty application version, the application state is restored before each test case is replayed on the fault-seeded version so that replay closely matches the clean (non-fault seeded) execution. Thus, faults that affect the underlying application state do not propagate in the application state for latter test cases to expose.

We implemented each of the prioritization techniques as described in Section 3 in C++ and Perl. In all of the implementations, in case of tie between two or more tests that meet the prioritization criterion, a random tie-breaking strategy is implemented. To account for the nondeterminism introduced by random tie breaking, we execute each prioritization technique five times and report the average rate of fault detection, APFD, and execution time.

Threats to Validity. Threats to validity in this experiment include our ability to generalize the results to all web-based applications. While we examine three actual web-based applications, the trends observed may not be representative of all web-based applications. Further, many of the faults were manually seeded into the application. Due to challenges faced in seeding faults in object-oriented code, in addition to the challenges of detecting hand-seeded faults described in [2], faults may not be evenly distributed in all the classes. Though we tried to model the seeded faults similar to naturally occurring faults—we even included a few naturally occurring faults in CPM—some of the seeded faults may not be representative of naturally occurring faults. The oracle comparators used in the study can have false positives or false negatives [29]. The execution times reported in this paper are an approximation based on the total time taken to execute the entire test suite. Actual individual test case execution times could vary depending on the test case’s characteristics. Also, the test suites for the three applications were not executed on the same machine—execution time comparisons across applications would not be fair.

5. Results and Analysis

5.1. CPM

Figure 2, Table 7 and Table 8 show the results for CPM. We show the rate of fault detection for different prioritization techniques in logarithmic scale. To better show the trends between the prioritization techniques, for each application we divide the prioritization technique five times and report the average rate of fault detection, APFD, and execution time.

Finding the most faults in the earliest tests. From Table 7 and Figure 2(a), if the APFD in the first 10% of the tests
suite, a test prioritization technique. The remaining prioritization techniques fall in between these best and worst cases. From Table 7, we see that prioritization by AAS has the highest APFD in the first 20% of the test suite’s execution. Figure 4 shows that PV-Stol and Req-Stol are slow starters during the first 10% of tests run, i.e., the first test case in each technique detects only 6 faults, whereas the first test case in the other techniques detects between 15 and 24 faults.

5.2. Masplas

Finding the most faults in the earliest tests. Figure 3 and Table 9 show that prioritization by Req-LtoS is the most effective if APFD during the first 30% of the test suite is of primary concern. The results for AAS, 1-way and 2-way are slightly less competitive than Req-LtoS.

Locating 100% of the faults earliest. From Table 9, we note that after the first 30% of the test suite is run, prioritization by 2-way has the best APFD for the remaining 70% of the test suite. From Table 8, 2-way finds all of the faults 33.3% of the test suite in 36 seconds, followed closely by 1-way. From Table 9, in the last 70% of the test suite, AAS, Req-LtoS, PV-LtoS are comparable in their APFD. PV-Stol’s APFD suggests that it is the least effective prioritization technique. The remaining prioritization techniques fall in between these best and worst cases.

5.3. Book

Locating 100% of the faults earliest. Table 10 shows that prioritization by 1-way and MFAS have a high APFD.

<table>
<thead>
<tr>
<th>Test No.</th>
<th>1-way</th>
<th>2-way</th>
<th>PV-LtoS</th>
<th>PV-Stol</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94.26</td>
<td>95.64</td>
<td>97.04</td>
<td>94.26</td>
</tr>
<tr>
<td>10</td>
<td>94.26</td>
<td>95.64</td>
<td>97.04</td>
<td>94.26</td>
</tr>
<tr>
<td>100</td>
<td>94.26</td>
<td>95.64</td>
<td>97.04</td>
<td>94.26</td>
</tr>
</tbody>
</table>

Figure 3. Masplas: Rate of fault detection.

Table 8. Percentage of the test suite run (Execution Time in seconds) for 100% Fault Detection

<table>
<thead>
<tr>
<th>Percentage of test suite run</th>
<th>MFAS</th>
<th>AAS</th>
<th>Req-LtoS</th>
<th>Req-Stol</th>
<th>Random</th>
<th>1-way</th>
<th>2-way</th>
<th>PV-LtoS</th>
<th>PV-Stol</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>65.43</td>
<td>85.28</td>
<td>78.17</td>
<td>75.14</td>
<td>48.63</td>
<td>83.79</td>
<td>83.72</td>
<td>83.53</td>
<td>16.38</td>
</tr>
<tr>
<td>20</td>
<td>74.16</td>
<td>88.52</td>
<td>80.34</td>
<td>77.76</td>
<td>57.55</td>
<td>87.78</td>
<td>90.8</td>
<td>88.77</td>
<td>25.6</td>
</tr>
<tr>
<td>30</td>
<td>77.75</td>
<td>89.4</td>
<td>81.77</td>
<td>80.27</td>
<td>64.31</td>
<td>91.54</td>
<td>91.72</td>
<td>88.77</td>
<td>26.44</td>
</tr>
<tr>
<td>40</td>
<td>79.61</td>
<td>89.36</td>
<td>84.58</td>
<td>81.39</td>
<td>69.19</td>
<td>94.79</td>
<td>95.64</td>
<td>92.71</td>
<td>28.16</td>
</tr>
<tr>
<td>50</td>
<td>80.92</td>
<td>91.04</td>
<td>85.58</td>
<td>82.95</td>
<td>73.03</td>
<td>94.79</td>
<td>95.64</td>
<td>92.71</td>
<td>30.33</td>
</tr>
<tr>
<td>60</td>
<td>81.71</td>
<td>91.58</td>
<td>87.14</td>
<td>84.44</td>
<td>75.37</td>
<td>94.79</td>
<td>95.64</td>
<td>94.26</td>
<td>34.64</td>
</tr>
<tr>
<td>70</td>
<td>82.73</td>
<td>92.1</td>
<td>87.74</td>
<td>85.15</td>
<td>77.37</td>
<td>94.79</td>
<td>95.64</td>
<td>94.26</td>
<td>39.15</td>
</tr>
<tr>
<td>80</td>
<td>84.28</td>
<td>92.35</td>
<td>88.27</td>
<td>86.21</td>
<td>78.24</td>
<td>94.79</td>
<td>95.64</td>
<td>94.26</td>
<td>39.58</td>
</tr>
<tr>
<td>90</td>
<td>84.57</td>
<td>92.37</td>
<td>88.3</td>
<td>86.31</td>
<td>78.45</td>
<td>94.99</td>
<td>95.64</td>
<td>94.26</td>
<td>42.18</td>
</tr>
<tr>
<td>100</td>
<td>84.64</td>
<td>92.45</td>
<td>88.36</td>
<td>86.35</td>
<td>78.89</td>
<td>94.99</td>
<td>95.64</td>
<td>94.26</td>
<td>43.09</td>
</tr>
</tbody>
</table>

Table 7. CPM: APFD Metric (in percentages). 10% of test suite = 89 test cases
From Table 8, **PV-StoL** detects 100% of the faults earliest, with 54% of the test suite in 300 seconds but has the lowest overall APFD. **2-way**, **PV-LtoS**, **AAS** are the next best prioritization metrics with respect to their APFD. **Req-StoL** performs poorly, in general, indicating that the number of base requests in the test case affect the fault detection rate. **Fault Detection Density**: From Table 10, we note that Random creates a reasonably effective test order with APFD comparable to the other techniques. To study Random’s behavior in Book, we define a metric called the fault detection density, which is a measure of the average number of faults detected by each test case. Given a set of test cases, $t_i \in T$ and a set of faults $F$ detected by test cases in $T$, let $t_{f_i}$ be the number of faults detected by $t_i$, then the fault detection density,

$$FDD = \frac{t_{f_1} + t_{f_2} + \ldots + t_{f_n}}{|T| \times |F|}$$  \hspace{1cm} (2)

FDD is the ratio of the sum of the total number of faults detected by each test case and the total number of test cases, normalized to the total number of faults detected. A fault detection density of 1 for a test suite indicates that each test case in the suite detects every fault. We found that Book’s test cases have a FDD of 0.59 (compared to 0.056 for CPM and 0.19 for Masplas). With a small test suite size (125 test cases) and a high FDD, Random has a greater chance of selecting a test case that detects several of the web application’s faults and thus creates an effective test suite order.

### 5.4. Summary of Results

Our experimental results show that none of our prioritization criteria is clearly the “best criteria” for all three of our web-based applications. However, for two of our three applications, prioritization techniques that consider parameter-value counts or interactions find 100% of faults before the other techniques. In particular **2-way** prioritization finds all of the faults with 38% of the test suite for CPM in 618 seconds, and 33% of the test suite for Masplas in 56 seconds. In both these applications, **2-way** has the highest APFD values overall (after 100% of the test suite is executed). In Book, however, **MFAS** has the highest overall APFD, followed by **1-way**.

In CPM, AAS is the best choice if effectiveness of the first 10% of tests is considered. If the APFD during any of the latter part of the test suite is considered, prioritization by **2-way** is most effective. In Masplas, **Req-LtoS** is the best technique when the APFD during the first 30% of the test suite is of primary concern. In the latter 70% of the test suite, giving preference to covering every 2-way parameter-value interaction creates the most effective test suite order. In Book all the metrics other than **Req-StoL**, **PV-StoL**, **Random** are good options if the goal is to achieve good fault detection early (first 10%) in the test cycle. However, **PV-StoL** is the best prioritization technique, if the goal is to achieve 100% fault detection with the smallest test number of test cases but has the overall lowest APFD. The frequency-based and the parameter-value interaction coverage techniques are better at detecting more faults early in the test execution cycle. Though it appears that Random creates an effective test suite ordering for Book, it is important to note that with larger number of test cases and low fault detection densities, Random’s effectiveness will decrease.

In terms of execution time, in CPM we observe that **2-way** detects 100% of the faults 30% faster than the worst technique, **PV-StoL**, and in Masplas **2-way** detects 100% of the faults 40% faster than the worst technique **PV-StoL**. **PV-StoL** in Book has the fastest rate of fault detection and detects 100% of the faults 74.5% faster than the worst technique, **Req-LtoS**, but has the lowest overall APFD.

**Guidance to Testers.** From our results, we find that depending on the tester’s goal and the characteristics of the web applications, different prioritization strategies may be useful to a tester. If a tester’s primary goal is to find 100% of the faults as soon as possible, prioritization by **2-way** and **PV-LtoS** are good choices. This indicates that systematically covering parameter-values is important when pri-

---

**Table 9. Masplas: APFD (in percentages). 10% of test suite \( \approx 18 \text{ test cases} \)**

<table>
<thead>
<tr>
<th>Percentage of test suite run</th>
<th>MFAS</th>
<th>AAS</th>
<th>Req-LtoS</th>
<th>Req-StoL</th>
<th>Random</th>
<th>1-way</th>
<th>2-way</th>
<th>PV-LtoS</th>
<th>PV-StoL</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>78.54</td>
<td>92.17</td>
<td><strong>95.12</strong></td>
<td>81.35</td>
<td>76.33</td>
<td>89.6</td>
<td>90.98</td>
<td>86.05</td>
<td>4.44</td>
</tr>
<tr>
<td>20</td>
<td>84.28</td>
<td>92.88</td>
<td><strong>95.12</strong></td>
<td>91.06</td>
<td>80.51</td>
<td>93.04</td>
<td>90.98</td>
<td>89.74</td>
<td>4.44</td>
</tr>
<tr>
<td>30</td>
<td>88.86</td>
<td>94.19</td>
<td><strong>95.12</strong></td>
<td>91.06</td>
<td>85.37</td>
<td>93.04</td>
<td>94.28</td>
<td>97.06</td>
<td>50.16</td>
</tr>
<tr>
<td>40</td>
<td>90.42</td>
<td>95.86</td>
<td>95.68</td>
<td>91.59</td>
<td>87.59</td>
<td>95.36</td>
<td><strong>97.06</strong></td>
<td>94.84</td>
<td>53.91</td>
</tr>
<tr>
<td>50</td>
<td>91.34</td>
<td>95.86</td>
<td>95.68</td>
<td>91.59</td>
<td>89.91</td>
<td>95.36</td>
<td><strong>97.06</strong></td>
<td>94.84</td>
<td>57.00</td>
</tr>
<tr>
<td>60</td>
<td>91.7</td>
<td>95.86</td>
<td>95.68</td>
<td>91.59</td>
<td>90.69</td>
<td>95.36</td>
<td><strong>97.06</strong></td>
<td>94.84</td>
<td>58.1</td>
</tr>
<tr>
<td>70</td>
<td>91.99</td>
<td>95.86</td>
<td>95.97</td>
<td>91.89</td>
<td>90.69</td>
<td>95.36</td>
<td><strong>97.06</strong></td>
<td>94.84</td>
<td>58.5</td>
</tr>
<tr>
<td>80</td>
<td>92.16</td>
<td>96.21</td>
<td>96.14</td>
<td>92.08</td>
<td>90.91</td>
<td>95.36</td>
<td><strong>97.06</strong></td>
<td>94.84</td>
<td>58.5</td>
</tr>
<tr>
<td>90</td>
<td>92.16</td>
<td>96.21</td>
<td>96.22</td>
<td>92.17</td>
<td>90.91</td>
<td>95.36</td>
<td><strong>97.06</strong></td>
<td>94.84</td>
<td>58.5</td>
</tr>
<tr>
<td>100</td>
<td>92.16</td>
<td>96.21</td>
<td>96.22</td>
<td>92.2</td>
<td>90.91</td>
<td>95.36</td>
<td><strong>97.06</strong></td>
<td>94.84</td>
<td>58.5</td>
</tr>
</tbody>
</table>

---

**148**
Based on our results from the FDD metric for Book, we believe that if a tester knows that their test suite has a high fault detection density, they may benefit by ordering test cases randomly. Given no particular tester preference/application characteristic, if the FDD values of pre-existing test suites are low, our results suggest that longer test cases achieve better fault detection than shorter test cases. In this case, instead of random ordering, the tester would benefit by prioritizing by number of base requests.

From the execution time results, we note that choosing the right prioritization could help the tester find and fix faults in the application quickly, which could translate into thousands of dollars in cost savings. Though the actual execution times for our applications are small, since they represent only the time to replay the test suite and do not include the time taken to manually or automatically identify and locate faults, our results indicate that choosing a prioritization technique can have a big time, and thus cost, impact.

6. Conclusions and Future Work

The web-application domain has an advantage that actual user-sessions can be recorded and used for regression testing. While these tests are indicative of user’s interactions with the system, selecting and prioritizing user-sessions has not been thoroughly studied. In this paper, we examine prioritization of such user-sessions for three web applications. We apply several new prioritization criteria to these test suites to identify whether they can be used to increase the rate of fault detection. Results from the experiments show that prioritization by frequency metrics and systematic coverage of parameter-value interactions may in-
crease the rate of fault detection for web applications.

In the future, we will consider the costs associated with the prioritization strategies and extend these studies to consider hybrid approaches to prioritization that combine more than one strategy for prioritizing test cases. Finally, we will extend our frequency-based metrics to consider other metrics, such as critical components of the application as determined by an organization using the web application, or module size. Further, for some web applications, a sequence of size 2 may not be large enough to capture faults in the application [25]. We plan to extend the frequency-based metrics to include sequences of size greater than two. We also plan to study whether the prioritization strategies can be applied to other types of web application test cases that are in the form of HTTP requests.

Acknowledgements. We thank Sara Sprenkle at Washington and Lee University for running experiments for CPM and Masplas execution time results. We thank Richard Kuhn, Raghu Kacker, Jeff Lei, Atif Memon and Sara Sprenkle for providing feedback on the paper.

References