







# Test Case Prioritization Using Test Similarities

Alireza Haghghatkhah<sup>(✉)</sup>, Mika Mäntylä<sup>(iD)</sup>, Markku Oivo<sup>(iD)</sup>,  
and Pasi Kuvaja<sup>(iD)</sup>

M3S Research Unit, University of Oulu, P.O. Box 3000, 90014 Oulu, Finland  
{alireza.haghghatkhah,mika.mantyla,markku.oivo,pasi.kuvaja}@oulu.fi

**Abstract.** A classical heuristic in software testing is to reward diversity, which implies that a higher priority must be assigned to test cases that differ the most from those already prioritized. This approach is commonly known as similarity-based test prioritization (SBTP) and can be realized using a variety of techniques. The objective of our study is to investigate whether SBTP is more effective at finding defects than random permutation, as well as determine which SBTP implementations lead to better results. To achieve our objective, we implemented five different techniques from the literature and conducted an experiment using the defects4j dataset, which contains 395 real faults from six real-world open-source Java programs. Findings indicate that running the most dissimilar test cases early in the process is largely more effective than random permutation (Vargha–Delaney A [VDA]: 0.76–0.99 observed using normalized compression distance). No technique was found to be superior with respect to the effectiveness. Locality-sensitive hashing was, to a small extent, less effective than other SBTP techniques (VDA: 0.38 observed in comparison to normalized compression distance), but its speed largely outperformed the other techniques (i.e., it was approximately 5–111 times faster). Our results bring to mind the well-known adage, “don’t put all your eggs in one basket”. To effectively consume a limited testing budget, one should spread it evenly across different parts of the system by running the most dissimilar test cases early in the testing process.

**Keywords:** Test case prioritization · Regression testing  
Test diversity · Test similarity

## 1 Introduction

The software industry is moving toward an agile, continuous delivery paradigm in which software changes are released more frequently and considerably faster than before [29]. This development paradigm has brought many benefits but posed several challenges, particularly regarding software quality [25, 29]. To ensure software correctness, software developers employ regression testing (RT), which involves running a dedicated regression test suite after each revision to verify

that recent changes have not negatively impacted the software's functionality [4]. Industrial software-intensive systems often comprise many test cases, and the execution of these test cases require several hours or even days. For example, the JOnAS Java EE middleware requires running 43,024 test cases to verify all of its 16 configurations [21]. To improve RT processes, the software engineering literature has proposed many solutions [34]. Test case prioritization (TCP) [30] is one of these solutions; it is concerned with the ideal ordering of test cases to maximize desirable properties (i.e., early fault detection). From the fault detection viewpoint, TCP seems to be a safe approach because it does not eliminate test cases and simply permutes them within the test suite.

To increase the likelihood of detecting faults, one potential strategy is spreading the testing budget evenly across different parts of the system [11, 18, 23], and realizing this strategy involves utilizing a diverse set of test cases. To devise a diverse test suite, one needs to measure similarities among the test cases. The notion of similarity measurement is a subject of interest for many applications. The degree to which two objects share characteristics is called similarity, and the degree to which they differ is termed distance. In the software testing literature, a point of particular interest is quantifying similarities among test cases. For example, in coverage-based testing, coverage information has been used as a proxy to measure the similarities among test cases [18]. More recently, several other properties have been described in the literature i.e., the overlap between test paths and their coverage in model-based testing [3, 12], as well as the source code of test cases [23], test input and output [16], topic models extracted from test scripts [32], and even English document of manual test cases [15].

The main intuition is that *test cases that capture the same faults tend to be more similar to each other, and test cases that capture different faults tend to be more dissimilar* [11, 18, 23]. The number of published empirical studies that support this intuition are growing (e.g., [1, 5, 7, 13, 15, 32]). The implication for TCP is that a higher priority must be assigned to test cases that are most dissimilar to those already prioritized. This can be realized by maximizing the distances among test cases ordered in the test suite. Similarity-based test prioritization (SBTP) is a black-box static technique (i.e., it does not require the source code and execution of the system under test) that can potentially be applied, for example, where code instrumentation is too costly or impossible.

The natural question that arises is whether running the most dissimilar test cases early in the testing process improves the test suite's fault-detection capability. SBTP can be implemented in a variety of ways, such as applying different similarity metrics. Thus, a follow-up question that arises is which implementation yields the best results. A similar objective was pursued by Ledru et al. [23] in 2012. The authors conducted a comprehensive experiment on the Siemens test suite and evaluated four classical string metrics using a pairwise algorithm. This study extends prior research by investigating the effectiveness and performance of five different SBTP techniques (4 additional in comparison to Ledru et al.). These techniques rely on different similarity metrics and were selected from the literature based on the results of recent experimental studies [5, 6, 14, 23, 26].

The ultimate objective of our study is to detect regression faults early in the testing process, allowing software developers to perform RT more frequently and continuously. To achieve this objective, we conducted an experiment using the defects4j dataset [20], which contains 395 real faults from 6 real-world open-source Java programs. Findings from our study indicate that:

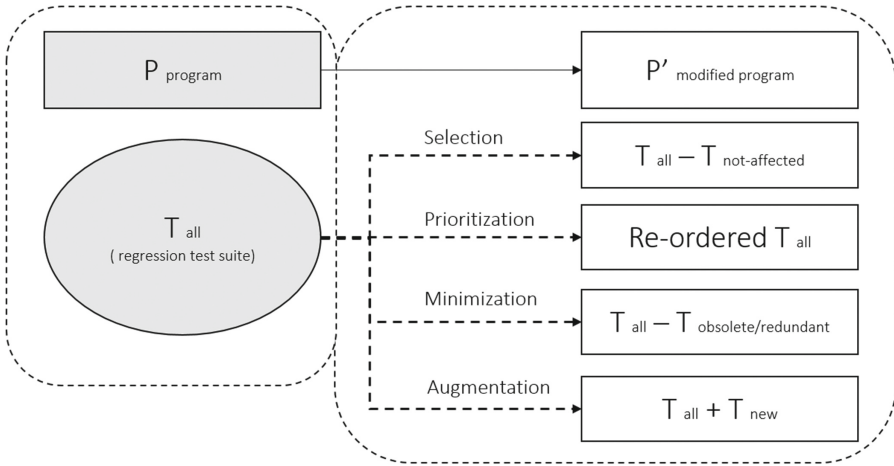
- When their average percentage of faults detected (APFD) are compared, test suites ordered by SBTP are largely more effective than random permutation (VDA: 0.76–0.99 observed using normalized compression distance [NCD] across all subjects), which means running the most dissimilar test cases early in the testing process improves the test suite’s fault-detection capability.
- Of the 5 SBTP implementations investigated, no technique was found to be superior with respect to the effectiveness. Locality-sensitive hashing (LSH) was, to a limited extent, less effective than other SBTP techniques (VDA: 0.38 observed in comparison to NCD), but its speed largely outperformed the other techniques (i.e., it was approximately 5–111 times faster).

Our findings yield important academic and practical implications. From the academic perspective, we provide empirical evidence that supports test diversity and its impact on TCP. From the of practitioners’ perspective, our results bring to mind the well-known adage, “don’t put all your eggs in one basket”. To effectively consume a limited testing budget, one should spread it evenly across different parts of the system by running the most dissimilar test cases early in the process. The remainder of the paper is organized as follows. Section 2 discusses the background and related works. Section 3 describes the research methodology, and Sect. 4 presents answers to the research questions. The findings are discussed in Sect. 5, and conclusions are discussed in Sect. 6.

## 2 Background and Related Work

### 2.1 Background

Figure 1 presents a general model of RT techniques. Let  $P$  be a program,  $P'$  be a modified version of the program, and  $T$  be a test suite developed for  $P$ . In the transition from  $P$  to  $P'$ , a previously verified behavior of  $P$  may have become faulty in  $P'$ . RT seeks to capture regressions in  $P'$  and verify that changes to the system have not negatively impacted any previously verified functionalities. During RT, several techniques may be employed. One of the techniques is test suite minimization; it seeks to identify and permanently eliminate obsolete or redundant test cases from the test suite. Another technique, regression test selection, aims to select only the subset of test cases affected by the recent changes. TCP is concerned with the ideal ordering of test cases to maximize desirable properties (i.e., early fault detection), while test suite augmentation aims to identify newly added source code and generate new test cases accordingly.



**Fig. 1.** General model of RT techniques

## 2.2 Related Work

A similarity metric, which is also known as similarity/distance function, is a metric that measures the similarity or distance (i.e., inverse similarity) between two objects. Similarity metrics have been widely applied in the literature (e.g., classification problems, plagiarism detection, sequence and image analysis).

In software engineering, particularly in software testing, similarity metrics have been applied. For instance, Shahbazi and Miller [31] conducted a large empirical study on black-box automated test-case generation using several string metrics. Their results indicate that superior test cases can be generated by controlling the diversity and length distribution of the string test cases. Hemmati et al. [13] proposed a similarity-based test case selection technique that selects the most diverse subset of test cases among those generated by applying a coverage criterion on a test model. Feldt et al. [5] proposed the test set diameter (TSDm) technique, which was developed based on NCD for multisets. Their results indicate that test selection using TSDm leads to higher structural and fault coverage than random selection. NCD multisets, which provides similarity measurement at the level of entire sets of elements rather than between pairs, have also been applied in the TCP literature recently [16].

To implement SBTP, the distances among test cases must be measured using a specific metric, and this information must then be leveraged to perform TCP. Ledru et al. [23] conducted a comprehensive experiment on the Siemens test suite and evaluated four classical string metrics for TCP purposes (i.e., Cartesian, Levenshtein, Hamming, and Manhattan distance). Their findings indicated that TCP using string metrics is more effective than random prioritization, and on average, Manhattan distance yields better results than the other investigated metrics. To calculate the distance between a test case  $t$  and set of test cases  $T'$ ,

Ledru et al. proposed the following function, which uses distance measure  $d$ :

$$distance(t, T', d) = \min\{d(t, t_i) | t_i \in T', t_i \neq t\}$$

Ledru et al. used the min operation because an empirical study by Jiang et al. [18] showed that maximize-minimum is more efficient than maximize-average and maximize-maximum. Ledru et al. also proposed an algorithm (Algorithm 1) that iteratively picks the most dissimilar test case (i.e., having the greatest distance from a set of already prioritized test cases).

---

**Algorithm 1.** Similarity-based TCP Using a Pairwise Algorithm

---

**Data:** Test Suite  $TS$

**Result:** Prioritized Suite  $PS$

- 1 Find  $t \in TS$  with the maximum  $distance(t, TS)$ ;
  - 2 Append  $t$  to  $PS$  and remove from  $TS$ ;
  - 3 **while**  $TS$  is not empty **do**
  - 4     Find  $t \in TS$  with the maximum  $distance(t, PS)$ ;
  - 5     Append  $t$  to  $PS$  and remove from  $TS$ ;
  - 6 **end**
- 

Using SBTP with a pairwise algorithm comes with the cost of pairwise comparison, and its performance becomes inefficient as the test suite becomes larger. The underlying issue in SBTP can be defined as a similarity search problem, which involves searching within a large set of objects for a subset of objects that closely resemble a given query object. One popular approach to solving similarity search problems is LSH, which was originally introduced by Indyk and Motwani [17] in 1998. LSH hashes input items so that similar items map to the same buckets with high probability [24]. LSH is widely used in the literature (see the many references in Google Scholar to [17]) but is only occasionally applied to software engineering problems (e.g., clone detection [19] and test generation [31]). More recently, Miranda et al. [26] proposed an approach based on LSH, which provides scalable SBTP in both white-box and black-box fashion.

The purpose of our study is to investigate whether SBTP is more effective at finding defects than random permutation and which SBTP implementations yield the best results. A similar objective was pursued by Ledru et al. [23] in 2012. In comparison to their work, we have investigated five different techniques with respect to their effectiveness and performance. These techniques rely on different similarity metrics and were selected from the literature based on the results of recent experimental studies [5, 6, 14, 23, 26]. The rationale behind their selection and details about their implementation is described in Sect. 3.3. The Siemens test suite, which was used by Ledru et al., is a classical dataset and widely used in the software testing literature. However, its representative character has been debated for several reasons (e.g., in [27], which was also acknowledged by Ledru et al. in [23]). In this work, we report an experiment conducted on the defects4j dataset [20], which contains 395 real faults from 6 open-source Java programs.

### 3 Research Method

In this section, the study’s objective and research questions, study subject, study design, and evaluation methods are discussed.

#### 3.1 Objective and Research Questions

The main objective of our study is to catch regression faults early in the testing process, allowing software developers to perform RT more frequently and continuously. The research questions and their rationales are as follows:

**RQ1: Is prioritization by similarity-based TCP more effective at finding defects than random permutation?** This research question is designed to investigate whether running the most dissimilar test cases early in the testing process improves the test suite’s fault-detection capability in comparison to random ordering.

**RQ2: Which similarity-based TCP technique is the most effective and has the best performance?** This research question is designed to compare the effectiveness and performance of investigated SBTP implementations. The rationale behind the investigated techniques’ selection and details about their implementation are described in Sect. 3.3.

#### 3.2 Subjects Under Study

To answer our research questions, we conducted an experiment using the defects4j dataset [20], which contains 395 real faults from 6 real-world open-source Java programs. The subject’s characteristics are presented in Table 1. Each analyzed subject’s name is presented in the first column, while the second column shows the number of versions analyzed for each program. The third and fourth columns present the median number of test classes and test cases, and the range is in parentheses. The last two columns show the source’s size (kilo line-of-code) and test code for the most recent version, as reported by SLOCCount<sup>1</sup>.

#### 3.3 Study Design

To answer RQ1, we compared the effectiveness of SBTP with random permutation. SBTP does not use a system under test; thus, it can hardly be more effective than TCP techniques, which use code coverage criteria [23]. Thus, like Ledru et al., we used random permutation as the baseline of our experiment. For the sake of a sanity check, we also included a TCP approach in which we minimize the diversity (i.e., maximize similarity among test cases). The rationale behind our sanity check is if diversity is valuable in TCP, then minimizing

---

<sup>1</sup> SLOCCount is a suite of programs used to count lines of code: <https://www.dwheeler.com/sloccount/>.

**Table 1.** Subject characteristics

Project	Versions	Test classes	Test cases	S-LOC	T-LOC
JFreeChart (Chart)	26	323 (301–356)	1789 (1591–2193)	123.527	37.396
Closure Compiler (Closure)	133	216 (118–235)	7389 (2595–8443)	251.855	85.138
Apache Lang (Lang)	65	89 (81–111)	1760 (1540–2291)	45.609	28.199
Apache Math (Math)	106	253 (91–385)	2319 (817–4378)	22.738	12.238
Mockito	38	237 (128–268)	1233 (704–1388)	38.914	10.638
Joda-Time (Time)	27	122 (120–123)	3906 (3749–4041)	176.965	41.536

diversity should, in turn, negatively affect the test suite’s fault-detection capability [16]. Effectiveness was measured using APFD, which is a commonly used metric in the TCP literature and elaborated on in Sect. 3.4.

To answer RQ2, we presented the aggregated the investigated techniques’ performance and effectiveness within and across studied subjects. Using the aggregated values, we can determine which technique achieved the best effectiveness and performance on average. The five SBTP techniques presented in Table 2 were selected from the literature and investigated in this experiment. To calculate the distances, we automatically downloaded the source code for all studied versions and used the source code behind the test classes at their exact version.

**Table 2.** TCP techniques investigated

Name (Acronym)	Objective	Reference
Random Permutation (RND)	Baseline	-
Manhattan Distance (MNH)	Maximize diversity	Ledru et al. [23]
Jaccard Distance (JAC)	Maximize diversity	Hemmati and Briand [14]
Normalized Compression Distance (NCD)	Maximize diversity	Feldt et al. [6]
Sanity Check (SC) using NCD	Maximize similarity	-
NCD Multisets (NCD-MS)	Maximize diversity	Feldt et al. [5]
Locality Sensitive Hasing (LSH)	Maximize diversity	Miranda et al. [26]

We implemented the Manhattan, Jaccard, NCD, and NCD Multisets using the pairwise algorithm proposed by Ledru et al. [23]. The Manhattan distance between two objects is the sum of the differences of their corresponding components. To calculate the Manhattan distance, the source code is converted to a vector of numbers. In practice, each character should be replaced with their ASCII code (or any other numerical coding). The Jaccard similarity between two sets  $x$  and  $y$  is defined as  $JS(x, y) = |x \cap y| / |x \cup y|$ , and their distance is  $JD(x, y) = 1 - JS(x, y)$ . To calculate the Jaccard distance, the source code is converted to a set of  $k$ -shingles (e.g., any substring of length  $k$  found within the text). In our study, we used  $k = 5$ , which is commonly used in the analysis of relatively short documents [24].

NCD and NCD Multisets both rely on a compressor function  $C$ , which calculates the approximate Kolmogorov complexity and returns the length of the input string after its compression, using a chosen compression program. In this study, we used LZ4, which is a high-speed lossless data compression algorithm<sup>2</sup>. The difference between NCD and NCD Multisets is that the latter performs similarity measurement at the level of entire sets of elements rather than between pairs. For the NCD Multisets, we adapted the pairwise algorithm so that at each iteration, we pick a test  $t \in TS$  that has maximum Kolmogorov complexity when compressed with the entire set of the already prioritized suite  $PS$ . This means that the candidate test has less mutual information with  $PS$  and is more different than any other  $t \in TS$ .

Furthermore, we implemented LSH using the MinHash technique to rapidly estimate Jaccard similarity. In our implementation, we followed the instructions provided by [24], which are also described here. To estimate the Jaccard similarity, we converted the source code to a set of 5-shingles. However, their size is often large, and it is impractical to use them directly. Using MinHashing technique, we replaced these sets with a much smaller representation (e.g., a signature) while preserving the Jaccard similarity between them. Given a hash function  $h$  and an input set  $S$ , we hashed all elements in the set using the hash function and picked the minimum resulting value as MinHash of  $S$ . This process was repeated  $P$  times (i.e., the number of permutations) using different hash functions to calculate the signature of a set (e.g., a sequence of MinHashes). Thereafter, the Jaccard similarity of two sets can be estimated using the fraction of common MinHashes in their signature. Using MinHashing, we were able to compress large sets into a small signature; similarity searches among large numbers of pairs is inefficient.

LSH works with a signature matrix (e.g., MinHash signatures as column) and divides it into  $b$  bands consisting of  $r$  rows each. For each band, LSH takes vectors of numbers (e.g., the portion of one column within that band) and hashes them to the buckets using a hash function. The more similar two columns are, the more likely they collide into some bands. When two items fall into the same bucket, it means a portion of their signature agrees, and they will be added to the candidate set. The candidate set returned by an LSH query only contains a subset of items that are more likely similar (e.g., having Jaccard similarity over a certain threshold). An approximation of this threshold is defined as  $ST = (1/b)^{(1/r)}$ .

Typically, LSH is configured with a high  $ST$  so that the candidate set only contains closely similar items. However, in our context, we are interested in items with a maximum distance from the LSH query. Thus, like Miranda et al. [26], we configured LSH so that we achieved an approximately 0.1 similarity threshold<sup>3</sup>, and the candidate set  $CS$  would contain almost all test cases, and the distant set  $DS$  would include a small number of remaining items with high Jaccard distance.

<sup>2</sup> The LZ4 compression algorithm and details regarding its implementation are available at <http://lz4.github.io/lz4/>.

<sup>3</sup> Permutations: 10; bands: 10; rows: 1.



To employ LSH for TCP purpose, we implemented an algorithm (Algorithm 2) proposed by Miranda et al. [26].

---

**Algorithm 2.** Similarity-based TCP Using Locality-Sensitive Hashing

---

**Data:** Test Suite  $TS$   
**Result:** Prioritized Suite  $PS$

- 1  $signatures \leftarrow \text{MinHashSignature}(TS)$ ;
- 2  $\text{LSH.Index}(signatures)$ ;
- 3  $query \leftarrow \text{MinHashSignature}(\emptyset)$ ;
- 4 **while**  $signatures$  is not empty **do**
- 5  $CS \leftarrow \text{LSH.Search}(query)$ ;
- 6  $DS \leftarrow signatures - CS - PS$ ;
- 7 Find  $i \in DS$  with the maximum JD (estimate) to  $PS$ ;
- 8 Append  $i$  to  $PS$  and remove from  $signatures$ ;
- 9  $query \leftarrow$  Update cumulative MinHash signature of  $PS$ ;
- 10 **end**

---

### 3.4 Evaluation

To compare the investigated TCP techniques, effectiveness and performance are both important. Performance was measured using average method execution time (AMET) in seconds. AMET includes both the preparation time (i.e., calculating the distance matrix or LSH initialization) and the prioritization algorithm itself. To assess effectiveness, we used an APFD metric that was originally introduced by Roethermel et al. [30] and is widely used in the literature [22]. Let  $T$  be an ordered test suite containing  $n$  test cases and  $F$  be a set of  $m$  faults detected by  $T$ ; then  $TF_i$  indicates the number of test cases executed in  $T$  before capturing fault  $i$ . APFD indicates the average percentage of faults detected and is defined as follows:

$$APFD = 100 * \left( 1 - \frac{TF_1 + TF_2 + \dots + TF_M}{nm} + \frac{1}{2n} \right)$$

To properly compare the investigated TCP techniques, we performed statistical analyses. A Mann–Whitney U test [2], which is a non-parametric significance test, was applied to determine whether the difference between two techniques is statistically significant, using  $p < 0.05$  as the significance threshold. The null hypothesis of this test indicates that there is no significant difference between the effectiveness of the techniques under evaluation. This test was selected because the studied data may not follow a normal distribution. The Mann–Whitney U test indicates whether there is any difference between techniques but does not show the degree of difference between them. Thus, we used a VDA measure [2], which is a non-parametric effect size. A VDA measure is a number between 0 and 1. When  $VDA(x, y) = 0.5$ , it indicates the two techniques are equal. When  $VDA(x, y) > 0.5$ , it means  $x$  outperformed  $y$  and vice versa. To compare

the investigated techniques across the subject programs, we presented the mean for the analyzed variables, and a 95% non-parametric confidence interval (CI) based on 1000 bias-corrected and accelerated bootstrap replicates. Furthermore, when comparing TCP techniques, we also provided violin plots to visualize the distribution of APFDs.

## 4 Findings

This section is structured to address the research questions and includes the aggregated results of all execution rounds (see the number of versions presented for each subject in Table 1). The experiments were conducted on a computer with an Intel 2.7 GHz Xeon E5-2680 CPU and 16 GB installed RAM. To accelerate the performance of investigated TCP techniques, we parallelized all techniques.

### 4.1 RQ1: Is Prioritization by Similarity-Based TCP More Effective at Finding Defects Than Random Permutation?

Table 3 presents the effect sizes for differences between analyzed SBTP techniques and random permutation. The analyzed SBTP techniques' effectiveness varies among subjects. However, one can observe that SBTP is largely more effective in finding defects than random permutation (VDA 0.76–0.99 observed using NCD across all subjects). These differences are also statistically significant in nearly all cases, which indicates running the most dissimilar test cases early in the testing process (maximizing the diversity) increases the test suite's fault-detection capability. This was also verified by our sanity check (SC) where the inverse approach was employed. The sanity check indicated maximizing similarities among tests would decrease the test suite's fault-detection capability, and as expected, it was less effective than random ordering (VDA: 0.03–0.34). Figure 2 shows the violin plots for the investigated TCP techniques within the studied subjects.

**Table 3.** VDA effect size - TCP technique vs. RND permutation

Project	MNH	JAC	NCD	NCD-MS	LSH	SC
Chart	<b>0.87</b>	<b>0.81</b>	<b>0.81</b>	<b>0.91</b>	<b>0.76</b>	<b>0.21</b>
Closure	<b>0.89</b>	<b>0.8</b>	<b>0.87</b>	<b>0.87</b>	<b>0.71</b>	<b>0.12</b>
Lang	<b>0.79</b>	<b>0.75</b>	<b>0.77</b>	<b>0.76</b>	<b>0.69</b>	<b>0.25</b>
Math	<b>0.84</b>	<b>0.82</b>	<b>0.84</b>	<b>0.84</b>	<b>0.76</b>	<b>0.16</b>
Mockito	0.58	<b>0.74</b>	<b>0.76</b>	0.6	<b>0.67</b>	<b>0.34</b>
Time	<b>0.96</b>	<b>0.99</b>	<b>0.99</b>	<b>0.96</b>	<b>0.9</b>	<b>0.03</b>
VDA range	0.58–0.96	0.74–0.99	0.76–0.99	0.60–0.96	0.67–0.90	0.03–0.34

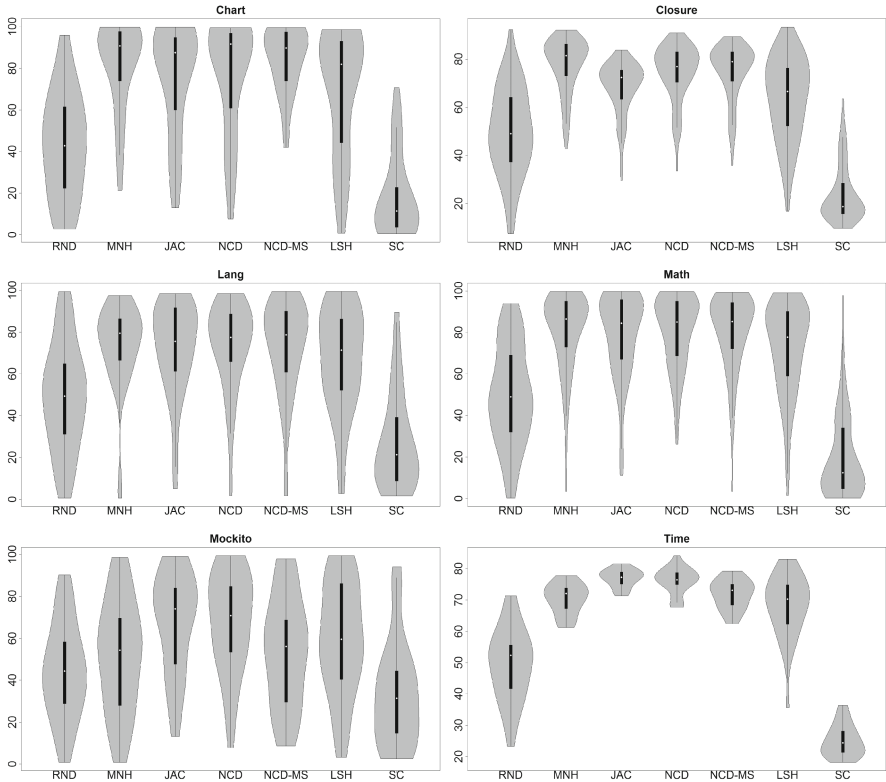


Fig. 2. Effectiveness (APFD) comparison - violin plots

### 4.2 RQ2: Which Similarity-Based TCP Technique Is the Most Effective and Has the Best Performance?

For a TCP approach to be applicable in a real-world environment, effectiveness (measured by APFD) and performance (measured by AMET) are both critical. Table 4 compares the effectiveness of the investigated techniques within and across the studied subjects. However, Table 5 compares the investigated techniques’ performance within and across the studied projects. One can observe that all SBTP techniques except LSH achieved very close mean APFD scores across all subjects (72.69–75.44). LSH achieved the lowest effectiveness (66.79), but had the best performance and scored a very low AMET across all subjects (1.24s). Overall, on average and across all subjects, no technique was found to be superior with respect to the effectiveness. LSH was, to a small extent, less effective than other SBTP techniques (VDA: 0.38 observed in comparison to NCD), but its speed largely outperformed the other techniques (i.e., it was approximately 5–111 times faster).

**Table 4.** Effectiveness (APFD) comparison

Project	MNH	JAC	NCD	NCD-MS	LSH
Chart	81.89	75.72	77.3	84.32	69.94
Closure	77.47	68.24	74.41	74.84	63.65
Lang	74.26	72.14	73.16	72.69	66.85
Math	80.39	78.81	80.37	80.32	72.21
Mockito	51.69	67.09	67.58	53.82	59.52
Time	70.39	76.95	75.99	71.8	68.06
Mean APFD (95% CI)	75.05 (72.90–76.87)	72.69 (70.84–74.47)	75.44 (73.83–77.28)	74.35 (72.51–76.26)	66.79 (64.64–68.81)

**Table 5.** Performance (AMET) comparison

Project	MNH	JAC	NCD	NCD-MS	LSH
Chart	138.22	25.02	15.65	102.99	2.84
Closure	230.15	15.05	5.85	89.99	1.32
Lang	33.18	2.76	0.58	7.22	0.29
Math	142.64	17.22	10.01	97.69	1.62
Mockito	41.54	7.14	5.63	14.32	1.04
Time	56.78	5.88	1.27	18.24	0.39
Mean AMET (95% CI)	138.21 (130.50–146)	12.88 (12.14–13.57)	6.41 (5.87–6.95)	67.11 (61.85–73.29)	1.24 (1.15–1.33)

## 5 Discussion

### 5.1 Overview of Findings, Their Implications, and Future Works

The ultimate objective of our study was to detect regression faults early in the testing process, allowing software developers to perform regression testing more frequently and continuously. To achieve our objective, we conducted an experiment using the defects4j dataset [20].

Test suites ordered by SBTP were largely more effective at finding defects than random permutation (VDA: 0.76–0.99 observed using NCD across all subjects). This indicates running the most dissimilar test cases early in the testing process (maximizing the diversity) increases the test suite’s fault-detection capability. This is also verified by our sanity check where the reverse approach was applied (VDA: 0.03–0.34). Of the 5 SBTP implementations investigated, no technique was found to be superior with respect to the effectiveness. LSH was, to a small extent, less effective than other SBTP techniques (VDA: 0.38 observed in comparison to NCD), but its speed largely outperformed the other techniques (i.e., it was approximately 5–111 times faster). From practical perspective, NCD seems to be the best choice because it achieved high effectiveness with

relatively low average method execution time. Yet, LSH is more practical when the prioritization time is critical.

Findings from our study bring to mind the well-known adage “don’t put all your eggs in one basket”. To effectively consume a limited testing budget, one should spread it evenly across different parts of the system by running the most dissimilar test cases early in the testing process. The underlying intuition is that *test cases that capture the same faults tend to be more similar to each other, and test cases that capture different faults tend to be more different* [11, 18, 23]. In comparison to other TCP techniques, SBTP requires minimal information (i.e., only the required information is encoded in the test suite) and has potential applications. SBTP can be applied in different contexts and during initial testing where no information about the system under test is available (e.g., code coverage or historical data). SBTP is an especially interesting approach when code instrumentation is too costly or impossible (e.g., in automotive system testing where source-code is not always available [8, 10]). SBTP can also be applied in a complementary fashion and combined with other TCP techniques (e.g., history-based diversity proposed in our previous work [9]).

To realize SBTP in practice, one must measure the similarities among test cases. This similarity measurement can be performed using string metrics and on different properties (i.e., the source code, documentation, or any other information about the test cases). As acknowledged by Ledru et al. [23], string metrics are based on lexicographic information and do not necessarily capture the semantics behind the test cases. Two test cases might consequently be considered similar, although they are distant and correspond to different execution paths. Future works are required to investigate possible approaches that precisely measure the semantic similarities among test cases. The candidate approach should not come with a high overhead; otherwise, its application remains in theory.

Once similarity measurement has been performed, this information should be leveraged to perform TCP. One can argue that diversification is perhaps the best strategy when no strong clues about fault-revealing test cases are available. Test diversity is a classical heuristic in the literature and has been applied previously [1, 5, 7, 11, 13, 15, 18, 23, 32]. The opposite viewpoint is the intensification strategy, where the testing budget is consumed by and around the most probable fault-revealing test cases. Theoretically, both strategies can be applied simultaneously (i.e., intensify where it is necessary and diversify the remaining budget). However, making decisions about when and how to apply these strategies, either individually or combined, remains a challenge. To the best of our knowledge, the application of these strategies, as well as their relevance and impact, have not been widely investigated in the literature. The only exception we are aware of is the recent study by Patrick and Jia [28] wherein the authors investigated the trade-off between diversification and intensification in adaptive random testing.

Regardless of which strategy is chosen, a TCP algorithm needs to iteratively find the most (dis)-similar item to the set of already prioritized test cases. This can be done using different search techniques. TCP using a pairwise algorithm does not scale, and its performance becomes inefficient as the test suite’s size

increases. In this work, we have investigated LSH as one popular solution to the similarity search problem. There are other solutions proposed in the literature. Future work should also investigate the effectiveness and performance of candidate solutions.

## 5.2 Threats to Validity

In empirical software engineering, validity threats can be grouped into four distinct classes: construct validity, internal validity, external validity, and reliability [33]. In the present context, construct validity relates to the use of right measures. To assess the investigated TCP techniques' effectiveness, we used the APFD metric, which is widely used in the literature (see the latest systematic literature review on TCP by Khatibsyarbini et al. [22]). Internal validity concerns the relationship between the constructs and the proposed explanation. This corresponds to the potential faults in our implementation. Our implementation was piloted on a small sample before running the actual experiment. Furthermore, the implementation and results were discussed and reviewed in regular meetings, which were held among the co-authors of this study.

External validity relates to the generalizability of the study and whether the subjects of our study are real-world projects. Our experiment was conducted on the defects4j dataset [20], which contains 395 real faults from 6 real-world open-source Java programs. Our conclusions are drawn based on ex-post analysis of software artifacts. This motivates our future work to replicate our experiment in industry and to larger systems. Reliability concerns the repeatability and reproducibility of the research procedure and conclusions. This required access to the analyzed subjects and a throughout report of the experiment. The data that we used is publicly available, and detailed information about our experiment and its implementation were presented in this paper.

## 6 Concluding Remarks

The ultimate objective of our study was to detect regression faults early in the testing process, allowing software developers to perform regression testing more frequently and continuously. To achieve this objective, we conducted an experiment using the defects4j dataset, which contains 395 real faults from 6 real-world open-source Java programs. In summary, the results from our experiments suggest the following:

(1) Test suites ordered by SBTP were largely more effective at finding defects than random permutation (VDA: 0.76–0.99 observed using NCD across all subjects), which means running the most dissimilar test cases early in the testing process improves the test suite's fault-detection capability; (2) Of the 5 SBTP implementations investigated, no technique was found to be superior with respect to the effectiveness. LSH was, to a small extent, less effective than other SBTP techniques (VDA: 0.38 observed in comparison to NCD), but its speed was faster than the other techniques studied (approximately 5–111 times faster).

Taken together, these results bring to mind the well-known adage “don’t put all your eggs in one basket”. To effectively consume a limited testing budget, one should spread it evenly across different parts of the system by running the most dissimilar test cases early in the process. Our study contributes to the literature by providing empirical evidence in support of test diversity and its impact on TCP.

## References

1. Arafeen, M.J., Do, H.: Test case prioritization using requirements-based clustering. In: 2013 IEEE Sixth International Conference on Software Testing, Verification and Validation (ICST), pp. 312–321. IEEE (2013)
2. Arcuri, A., Briand, L.: A practical guide for using statistical tests to assess randomized algorithms in software engineering. In: 2011 33rd International Conference on Software Engineering (ICSE), pp. 1–10. IEEE (2011)
3. Cartaxo, E.G., Machado, P.D., Neto, F.G.O.: On the use of a similarity function for test case selection in the context of model-based testing. *Softw. Test. Verif. Reliab.* **21**(2), 75–100 (2011)
4. Engström, E., Runeson, P.: A qualitative survey of regression testing practices. In: Ali Babar, M., Vierimaa, M., Oivo, M. (eds.) PROFES 2010. LNCS, vol. 6156, pp. 3–16. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-13792-1\\_3](https://doi.org/10.1007/978-3-642-13792-1_3)
5. Feldt, R., Poulding, S., Clark, D., Yoo, S.: Test set diameter: quantifying the diversity of sets of test cases. In: 2016 IEEE International Conference on Software Testing, Verification and Validation (ICST), pp. 223–233. IEEE (2016)
6. Feldt, R., Torkar, R., Gorschek, T., Afzal, W.: Searching for cognitively diverse tests: towards universal test diversity metrics. In: IEEE International Conference on Software Testing Verification and Validation Workshop, ICSTW 2008, pp. 178–186. IEEE (2008)
7. Flemström, D., Potena, P., Sundmark, D., Afzal, W., Bohlin, M.: Similarity-based prioritization of test case automation. *Softw. Qual. J.*, 1–29 (2017). <https://doi.org/10.1007/s11219-017-9401-7>
8. Haghighatkah, A., Banijamali, A., Pakanen, O.P., Oivo, M., Kuvaja, P.: Automotive software engineering: a systematic mapping study. *J. Syst. Softw.* **128**, 25–55 (2017)
9. Haghighatkah, A., Mäntylä, M., Oivo, M., Kuvaja, P.: Test prioritization in continuous integration environments. *J. Syst. Softw.* (2018). <https://doi.org/10.1016/j.jss.2018.08.061>, <http://www.sciencedirect.com/science/article/pii/S0164121218301730>
10. Haghighatkah, A., Oivo, M., Banijamali, A., Kuvaja, P.: Improving the state of automotive software engineering. *IEEE Softw.* **34**(5), 82–86 (2017)
11. Hemmati, H., Arcuri, A., Briand, L.: Reducing the cost of model-based testing through test case diversity. In: Petrenko, A., Simão, A., Maldonado, J.C. (eds.) ICTSS 2010. LNCS, vol. 6435, pp. 63–78. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-16573-3\\_6](https://doi.org/10.1007/978-3-642-16573-3_6)
12. Hemmati, H., Arcuri, A., Briand, L.: Empirical investigation of the effects of test suite properties on similarity-based test case selection. In: 2011 IEEE Fourth International Conference on Software Testing, Verification and Validation (ICST), pp. 327–336. IEEE (2011)

13. Hemmati, H., Arcuri, A., Briand, L.: Achieving scalable model-based testing through test case diversity. *ACM Trans. Softw. Eng. Methodol. (TOSEM)* **22**(1), 6 (2013)
14. Hemmati, H., Briand, L.: An industrial investigation of similarity measures for model-based test case selection. In: 2010 IEEE 21st International Symposium on Software Reliability Engineering (ISSRE), pp. 141–150. IEEE (2010)
15. Hemmati, H., Fang, Z., Mäntylä, M.V., Adams, B.: Prioritizing manual test cases in rapid release environments. *Softw. Test. Verif. Reliab.* **27**(6), e1609 (2017)
16. Henard, C., Papadakis, M., Harman, M., Jia, Y., Le Traon, Y.: Comparing white-box and black-box test prioritization. In: 2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE), pp. 523–534. IEEE (2016)
17. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pp. 604–613. ACM (1998)
18. Jiang, B., Zhang, Z., Chan, W.K., Tse, T.: Adaptive random test case prioritization. In: Proceedings of the 2009 IEEE/ACM International Conference on Automated Software Engineering, pp. 233–244. IEEE Computer Society (2009)
19. Jiang, L., Misherghi, G., Su, Z., Glondu, S.: Deckard: scalable and accurate tree-based detection of code clones. In: Proceedings of the 29th International Conference on Software Engineering, pp. 96–105. IEEE Computer Society (2007)
20. Just, R., Jalali, D., Ernst, M.D.: Defects4j: a database of existing faults to enable controlled testing studies for Java programs. In: Proceedings of the 2014 International Symposium on Software Testing and Analysis, pp. 437–440. ACM (2014)
21. Kessiss, M., Ledru, Y., Vandome, G.: Experiences in coverage testing of a Java middleware. In: Proceedings of the 5th International Workshop on Software Engineering and Middleware, pp. 39–45. ACM (2005)
22. Khatibsyarbini, M., Isa, M.A., Jawawi, D.N., Tumeng, R.: Test case prioritization approaches in regression testing: a systematic literature review. *Inf. Softw. Technol.* **93**, 74–93 (2017)
23. Ledru, Y., Petrenko, A., Boroday, S., Mandran, N.: Prioritizing test cases with string distances. *Autom. Softw. Eng.* **19**(1), 65–95 (2012)
24. Leskovec, J., Rajaraman, A., Ullman, J.D.: Mining of Massive Datasets. Cambridge University Press, New York (2014)
25. Mäntylä, M.V., Adams, B., Khomh, F., Engström, E., Petersen, K.: On rapid releases and software testing: a case study and a semi-systematic literature review. *Empir. Softw. Eng.* **20**(5), 1384–1425 (2015)
26. Miranda, B., Verdecchia, R., Cruciani, E., Bertolino, A.: Fast approaches to scalable similarity-based test case prioritization. In: 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE). IEEE (2018)
27. Orso, A., Rothermel, G.: Software testing: a research travelogue (2000–2014). In: Proceedings of the on Future of Software Engineering, pp. 117–132. ACM (2014)
28. Patrick, M., Jia, Y.: KD-ART: should we intensify or diversify tests to kill mutants? *Inf. Softw. Technol.* **81**, 36–51 (2017)
29. Rodríguez, P., et al.: Continuous deployment of software intensive products and services: a systematic mapping study. *J. Syst. Softw.* **123**, 263–291 (2017)
30. Rothermel, G., Untch, R.H., Chu, C., Harrold, M.J.: Prioritizing test cases for regression testing. *IEEE Trans. Softw. Eng.* **27**(10), 929–948 (2001)
31. Shahbazi, A., Miller, J.: Black-box string test case generation through a multi-objective optimization. *IEEE Trans. Softw. Eng.* **42**(4), 361–378 (2016)
32. Thomas, S.W., Hemmati, H., Hassan, A.E., Blostein, D.: Static test case prioritization using topic models. *Empir. Softw. Eng.* **19**(1), 182–212 (2014)



33. Wohlin, C., Runeson, P., Host, M., Ohlsson, C., Regnell, B., Wesslén, A.: *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers, Norwell (2000)
34. Yoo, S., Harman, M.: Regression testing minimization, selection and prioritization: a survey. *Softw. Test. Verif. Reliab.* **22**(2), 67–120 (2012)